SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 1–17

ASYMPTOTIC SHAPE OF THE ERLANG CAPACITY REGION OF A CRITICALLY LOADED MULTISERVICE SHARED RESOURCE*

JOHN A. MORRISON†

Abstract. We consider a loss model of an unbuffered resource having C channels, which are shared by several different types of service connections. Connections of each type arrive in a Poisson stream and request a number of channels, which depends on the type. An arriving connection is blocked and lost if there are not enough free channels. Otherwise, the channels are held for the duration of the connection, and the holding period is generally distributed. It is assumed that C and the traffic intensities are proportionately large, and that the resource is critically loaded. The admission control problem is considered for specified upper bounds on the blocking probabilities, and the boundary of the admissible set is investigated asymptotically. It is shown that the boundary of the admissible set is not convex, although only very slightly so. This completes the investigation of a critically loaded resource, initiated in [J. A. Morrison and D. Mitra, *SIAM J. Appl. Math.*, to appear], which also investigated overloaded and underloaded resources.

Key words. admissible set, asymptotics, Erlang capacity, network design, network economics

AMS subject classifications. 60K30, 90B12

DOI. 10.1137/S0036139902401094

1. Introduction.

We consider an unbuffered resource having C channels, which are shared by J different types of connections. Connections of type j arrive in a Poisson stream with mean rate λ_j, and they require d_j channels. An arriving connection is blocked and lost if there are fewer than d_j free channels. Otherwise, d_j channels are held for the duration of the connection, and the holding period is generally distributed with mean $1/\mu_j$ and is independent of earlier arrival and holding times. The traffic intensity of type j connections is $\rho_j = \lambda_j/\mu_j$, and the product form and the insensitivity property hold [3], [4], [6]; i.e., the joint stationary distribution of the number of active connections of each type depends on the distributions only through ρ_i, $i = 1, \ldots, J$. The blocking probabilities L_j for type j connections satisfy $L_j > 0$ for $\rho_j > 0$ and, assuming that $C \geq \max_i d_i$, $L_j \to 0+$ only if $\rho_i \to 0+$, $i = 1, \ldots, J$. The admissible set in $\mathbb{R}^J$ contains all combinations of ρ_j, $j = 1, \ldots, J$, such that the blocking probability for each connection type satisfies specified bounds, i.e., $L_j \leq \ell_j$, $j = 1, \ldots, J$, where ℓ_j is a prescribed function of C.

Characterization of the admissible set is extremely useful, not only for connection-level admission control, which is the context in which this topic has typically been considered in the past, but also for higher level objectives, such as network economics and network design and operations. The asymptotic view of the admissible set is particularly appropriate for the latter, where the fine details are not as important as the qualitative properties of the shape of the set and tractability of the numerical calculations for large systems.

Special importance is attached to admissible regions with linear boundaries; the solution space is determined by its vertices, which are relatively easy to compute. Optimizations within such spaces are also much easier computationally. Network economics applications are given in [5]. In one such example, the objective function is the

*Received by the editors January 16, 2002; accepted for publication (in revised form) March 21, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/siap/64-1/40109.html
†Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974 (johnmorrison@lucent.com).

profit of a service provider giving several quality of service (QoS) levels at prices that are the solution to the corresponding optimization problem. The admissible region of solutions is defined by a collection of inequalities imposed by available capacity, one for each QoS level. For further details see [5], and for other such applications see [2], [7], [17], and references therein. See [1], [10] for applications to routing and control. For recent work on loss models of optical networking see [15] and [16].

Convexity is also important, since the tangent hyperplane at points on the boundary of the admissible set intersects the positive axes. Hence, if the boundary of the admissible set is convex, then the region in the positive orthant bounded by the tangent hyperplane at a point on the boundary is admissible and may be used in an approximate optimization.

Mitra and Morrison [9] considered our model here (as well as the finite-sources version), and they investigated the case where C and the traffic intensities $\rho_j = \alpha_j C$, $j = 1, \ldots, J$, are proportionately large, so that $\alpha_j = O(1)$ is bounded away from zero. They derived uniform asymptotic approximations to the blocking probabilities L_j, $j = 1, \ldots, J$, for type j connections. The results for Poisson arrivals are obtained from the finite-sources version as a limiting case. They presented numerical results for $J = 2$ and $J = 3$ types for the finite-sources model. These results constitute a numerical procedure but do not provide a characterization of the admissible set, nor do they resolve specific questions on the linearity and convexity of the boundary.

Consequently, Morrison and Mitra [12] investigated the boundary of the admissible set in the case of Poisson arrivals. The uniform asymptotic approximations to the blocking probabilities are specialized [9] to three regimes in which their behavior is markedly different, namely, the overloaded, the critically loaded, and the underloaded regimes, corresponding to $\sum_{j=1}^{J} d_j \alpha_j > 1$, $\sum_{j=1}^{J} d_j \alpha_j - 1 = O(1/\sqrt{C})$, and $\sum_{j=1}^{J} d_j \alpha_j < 1$, respectively. The corresponding blocking probabilities L_j are $O(1)$, $O(1/\sqrt{C})$, and exponentially small in C, respectively, so that the critically loaded regime is of greatest interest. In [12], the shape of the admissible set is investigated separately for each of the three regimes.

In the asymptotic limit $C \to \infty$, with $\rho_j = \alpha_j C$, $j = 1, \ldots, J$, the boundary of the admissible set lies in a hyperplane if the resource is critically loaded or overloaded. If the resource is underloaded, the boundary of the admissible set, in the limit $C \to \infty$, is convex, but not strictly so, except when $J = 2$.

Refined results, which pertain to $C \gg 1$, are derived in [12]. The correction terms are $O(1/C)$ if the resource is overloaded or underloaded and $O(1/\sqrt{C})$ if the resource is critically loaded. In general, the boundary of the admissible set is not convex. If $J = 2$, then the boundary is slightly convex if the resource is critically loaded, but slightly concave if the resource is overloaded. For $J = 2$, the convexity is maintained from the $C \to \infty$ limit for an underloaded resource. Unfortunately, for $J \geq 3$, the boundary of the admissible set is *not* convex, although only slightly so, whether the resource is overloaded or underloaded. The case of a critically loaded resource requires further investigation, which is carried out in this paper.

In section 2 we first state the refined result obtained in [12] for the critically loaded regime, which is based on refined asymptotic approximations [13] to the blocking probabilities. This illustrates why a further refinement is required. We then state the further refined result, which is based on the refined uniform asymptotic approximations [14] to the blocking probabilities. The result establishes that the boundary of the admissible set is *not* convex, although only very slightly so. In spite of this negative result, it is important that the practitioner be aware of it. Moreover, since

the boundary is only slightly nonconvex, a slightly smaller admissible region with a convex boundary may be used by the practitioner.

In section 3, we derive a refined asymptotic approximation to the boundary of the admissible set, based on the refined uniform asymptotic approximations [14] to the blocking probabilities. In section 4, we express the boundary of the admissible set in appropriate coordinates with respect to the tangent hyperplane at a point on the boundary. In section 5, we show that the discriminant of the quadratic form which arises is negative, and hence that the boundary of the admissible set is *not* convex.

2. Refined results. Throughout the paper, we assume that

$$C \gg 1;\ \rho_j = \alpha_j C, \quad j = 1, \ldots, J, \tag{2.1}$$

where C is an integer and $\alpha_j > 0$ is $O(1)$ and bounded away from zero. We also assume that d_j, $j = 1, \ldots, J$, are distinct positive integers, not large relative to C, and that the greatest common divisor of $d_1, \ldots, d_J$ is 1. The admissible set corresponds to

$$L_j\left(\alpha_1, \ldots, \alpha_J;\ C\right) \le \ell_j, \quad j = 1, \ldots, J, \tag{2.2}$$

where the function L_j gives the blocking probability for type j connections. It is shown [12], in all three regimes, that asymptotically

$$\frac{\partial L_j}{\partial \alpha_k} > 0, \quad j, k = 1, \ldots, J, \tag{2.3}$$

and the boundary of the admissible set is expressed in the form

$$\alpha_J = \alpha_J(\boldsymbol{\alpha}; C), \quad \boldsymbol{\alpha} = \left(\alpha_1, \ldots, \alpha_{J-1}\right). \tag{2.4}$$

If

$$\sqrt{C} \min_j \ell_j = O\left(1\right) \tag{2.5}$$

is bounded below by a positive constant, then on the boundary of the admissible set the resource is critically loaded. The boundary satisfies

$$\sum_{j=1}^{J} d_j \alpha_j = 1 + O\left(1/\sqrt{C}\right), \tag{2.6}$$

which in the asymptotic limit $C \to \infty$ lies in a hyperplane.

We have the following refined approximation [12] to the boundary of the admissible set.

PROPOSITION 2.1. *If* (2.5) *holds, then*

(i) *if* $J = 2$, *then* $0 > d\alpha_2/d\alpha_1 = O(1)$ *and* $0 < d^2\alpha_2/d\alpha_1^2 = O(1/\sqrt{C})$, *so that the boundary of the admissible set is convex, although only slightly so;*

(ii) *we consider the case when*

$$\frac{d_J}{\sqrt{C}\,\ell_J} - \frac{d_i}{\sqrt{C}\,\ell_i} \gg \frac{1}{\sqrt{C}}, \quad i = 1, \ldots, J-1. \tag{2.7}$$

With this assumption the boundary of the admissible set is given by $L_J = l_J$. *Let* $(\boldsymbol{\alpha}^{(0)}, \alpha_J^{(0)})$, *where corresponding to* (2.4) $\alpha_J^{(0)} = \alpha_J(\boldsymbol{\alpha}^{(0)}; C)$, *be a point on the boundary of the admissible set. If* $J \ge 3$, *the linear transformation of variables*

$$\zeta_J = d_J \left[\alpha_J - \alpha_J^{(0)} - \sum_{i=1}^{J-1} \frac{\partial \alpha_J}{\partial \alpha_i}\left(\boldsymbol{\alpha}^{(0)}\right)\left(\alpha_i - \alpha_i^{(0)}\right)\right], \tag{2.8}$$

$$\zeta_{J-1} = \sum_{i=1}^{J-1} d_i(d_i - d_J)\left(\alpha_i - \alpha_i^{(0)}\right), \tag{2.9}$$

$$\zeta_{J-2} = \sum_{i=1}^{J-1} d_i\left(d_i^2 - d_J^2\right)\left(\alpha_i - \alpha_i^{(0)}\right), \tag{2.10}$$

and

$$\zeta_i = \alpha_i - \alpha_i^{(0)}, \quad i = 1, \ldots, J-3 \ (J \geq 4), \tag{2.11}$$

is nonsingular, and $\zeta_J = 0$ *corresponds to the tangent hyperplane to the boundary of the admissible set at* $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(0)}$. *From* (2.6) *and* (2.8) *we have*

$$\zeta_J = \sum_{j=1}^{J} d_j\left(\alpha_j - \alpha_j^{(0)}\right) + O\left(\frac{1}{\sqrt{C}}\right). \tag{2.12}$$

Let

$$\sigma_0^2 = 2\sum_{j=1}^{J} d_j^2 \alpha_j^{(0)}, \quad \sigma_0 > 0\,. \tag{2.13}$$

If $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$, *then*

$$\begin{aligned}\zeta_J = {} & \frac{\zeta_{J-1}^2}{2\sqrt{C}}\left[P_2(\sigma_0) + O\left(\frac{1}{\sqrt{C}}\right) + O\left(\zeta_{J-1}\right)\right] \\ & + \frac{\zeta_{J-1}\zeta_{J-2}}{C}\left[R_2(\sigma_0) + O(\zeta_{J-1})\right] + O\left(\frac{\epsilon^2}{C\sqrt{C}}\right)\end{aligned} \tag{2.14}$$

and $P_2(\sigma_0) > 0$. □

If $\zeta_{J-1} = O(\epsilon/\sqrt{C})$, then the leading terms in (2.14) are all $O(\epsilon^2/C\sqrt{C})$. Hence, the next order term in the asymptotic expansion in powers of $1/\sqrt{C}$ is needed to investigate whether or not the boundary of the admissible set is convex. In this paper we establish the following.

PROPOSITION 2.2. *Assume that* (2.5) *and* (2.7) *hold. If in* (2.9)

$$\zeta_{J-1} = Z_{J-1}/\sqrt{C}, \tag{2.15}$$

where $Z_{J-1} = O(1)$, *then*

$$\zeta_J = \frac{1}{2C^{3/2}}\left[P_2(\sigma_0)Z_{J-1}^2 + 2R_2(\sigma_0)Z_{J-1}\zeta_{J-2} + T_2(\sigma_0)\zeta_{J-2}^2\right] + O\left(\frac{1}{C^2}\right). \tag{2.16}$$

Moreover, the discriminant of the quadratic form in (2.16) *is negative. Hence, asymptotically, the boundary of the admissible set is* not *convex, although only very slightly so.* □

We note that in this result we do not need $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$.

3. Boundary of the admissible set. To derive the appropriate refinement of (2.14), we make use of the refined uniform asymptotic approximations [14] to the blocking probabilities, which we now summarize. Let

$$f(z) = \sum_{j=1}^{J} \alpha_j (z^{d_j} - 1) - \log z, \tag{3.1}$$

and let z^* be the unique positive solution of $f'(z) = 0$, so that

$$\sum_{j=1}^{J} \alpha_j d_j (z^*)^{d_j} = 1, \qquad z^* > 0. \tag{3.2}$$

Since $f'(z^*) = 0$,

$$v \triangleq (z^*)^2 f''(z^*) = \sum_{j=1}^{J} \alpha_j d_j^2 (z^*)^{d_j}, \tag{3.3}$$

$$\tau \triangleq (z^*)^3 f^{(3)}(z^*) = \sum_{j=1}^{J} \alpha_j d_j^2 (d_j - 3)(z^*)^{d_j}, \tag{3.4}$$

and

$$y \triangleq (z^*)^4 f^{(4)}(z^*) = \sum_{j=1}^{J} \alpha_j d_j^2 (d_j^2 - 6d_j + 11)(z^*)^{d_j}. \tag{3.5}$$

Next, let

$$K = \frac{1}{(1 - z^*)} - \frac{\sqrt{v}\ \mathrm{sgn}(1 - z^*)}{\sqrt{-2f(z^*)}}, \quad z^* \neq 1, \tag{3.6}$$

$$h_j(z) = \frac{(z^{d_j} - 1)}{z(z - 1)}, \quad z \neq 1, \quad h_j(1) = d_j, \tag{3.7}$$

and

$$N_j = \frac{z^*}{8v} \left(\frac{y}{v} - \frac{5\tau^2}{3v^2} \right) h_j(z^*) + \frac{\tau (z^*)^2}{2v^2} h_j'(z^*) - \frac{(z^*)^3}{2v} h_j''(z^*). \tag{3.8}$$

Since $f(1) = 0$, $f'(z^*) = 0$, and $(z^*)^2 f''(z^*) = v$, the expression for K in (3.6) remains finite as $z^* \to 1$. We define

$$\Omega = \sqrt{\frac{\pi v}{2}} e^{-Cf(z^*)}\ \mathrm{Erfc}\ \left[\mathrm{sgn}(1 - z^*) \sqrt{-Cf(z^*)} \right], \tag{3.9}$$

where the complementary error function is given by

$$\mathrm{Erfc}\,(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-\xi^2} d\xi. \tag{3.10}$$

Then [14], the blocking probabilities are asymptotically given by

$$L_j = \frac{[z^* h_j(z^*) + N_j/C + O(1/C^2)]}{[\Omega\sqrt{C} + K + O(1/C)]}. \tag{3.11}$$

We consider the case where

$$\frac{d_J}{\sqrt{C}\ell_J} - \frac{d_i}{\sqrt{C}\ell_i} \gg \frac{1}{\sqrt{C}}, \quad i = 1, \ldots, J-1, \tag{3.12}$$

and let

$$\kappa = \frac{d_J}{\sqrt{C}\ell_J}. \tag{3.13}$$

We assume that (2.5) holds, so that $\kappa = O(1)$. Since $z^* - 1 = O(1/\sqrt{C})$ in the critically loaded regime, L_j is asymptotically proportional to d_j, to lowest order, and [12] the boundary of the admissible set is given by $L_J = \ell_J$. Hence, from (3.11) and (3.13),

$$\Omega = \frac{\kappa}{d_J}\left[z^* h_J(z^*) + \frac{N_J}{C}\right] - \frac{K}{\sqrt{C}} + O\left(\frac{1}{C^{3/2}}\right). \tag{3.14}$$

We define

$$\phi(w) = \frac{\sqrt{\pi}}{2} e^{w^2}\ \mathrm{Erfc}\,(-w)\ . \tag{3.15}$$

Then, from (3.9) and (3.14),

$$\phi\left[\mathrm{sgn}(z^* - 1)\sqrt{-Cf(z^*)}\right] = \frac{1}{\sqrt{2v}}\left\{\frac{\kappa}{d_J}\left[z^* h_j(z^*) + \frac{N_J}{C}\right] - \frac{K}{\sqrt{C}} + O\left(\frac{1}{C^{3/2}}\right)\right\}. \tag{3.16}$$

But, from (3.10) and (3.15),

$$\phi(w) = e^{w^2}\int_{-w}^{\infty} e^{-\xi^2} d\xi = \int_0^{\infty} e^{2wu} e^{-u^2} du\,. \tag{3.17}$$

Hence $\phi'(w) > 0$, $-\infty < w < \infty$, so that $\phi(w)$ has a unique inverse

$$w(y) = \phi^{-1}(y), \quad y > 0. \tag{3.18}$$

It follows from (3.16) that

$$\mathrm{sgn}(z^* - 1)\sqrt{-Cf(z^*)} = w\left[\frac{1}{\sqrt{2v}}\left\{\frac{\kappa}{d_J}\left[z^* h_J(z^*) + \frac{N_J}{C}\right] - \frac{K}{\sqrt{C}} + O\left(\frac{1}{C^{3/2}}\right)\right\}\right]. \tag{3.19}$$

We let

$$1 - \sum_{j=1}^{J} \alpha_j d_j = \frac{\delta}{\sqrt{C}}, \tag{3.20}$$

where $\delta = O(1)$ from (2.6), and we define

$$\sigma^2 = 2\sum_{j=1}^{J} \alpha_j d_j^2, \qquad \sigma > 0, \tag{3.21}$$

and

$$\eta = \sum_{j=1}^{J} \alpha_j d_j^3, \qquad \rho = \sum_{j=1}^{J} \alpha_j d_j^4. \tag{3.22}$$

It is shown in Appendix A that

$$\operatorname{sgn}(z^*-1)\sqrt{-Cf(z^*)} = \frac{\delta}{\sigma} - \frac{2\delta^2\eta}{3\sigma^5\sqrt{C}} + \frac{\delta^3}{9\sigma^9 C}(16\eta^2 - 3\sigma^2\rho) + O\left(\frac{1}{C^{3/2}}\right), \tag{3.23}$$

and

$$\begin{aligned}
&\frac{1}{\sqrt{2v}}\left\{\frac{\kappa}{d_J}\left[z^* h_J(z^*) + \frac{N_J}{C}\right] - \frac{K}{\sqrt{C}}\right\} \\
&= \frac{\kappa}{\sigma} + \frac{\kappa\delta}{\sigma^5\sqrt{C}}[(d_J-1)\sigma^2 - 2\eta] - \frac{1}{6\sigma^3\sqrt{C}}(2\eta + 3\sigma^2) \\
&\quad + \frac{\kappa\delta^2}{\sigma^9 C}\left[\frac{1}{3}(d_J-1)(2d_J-1)\sigma^4 - 4(d_J-1)\sigma^2\eta + 2(5\eta^2 - \sigma^2\rho)\right] \\
&\quad + \frac{\kappa}{\sigma^7 C}\left(\frac{1}{2}\sigma^2\rho + 2\sigma^2\eta - \sigma^4 - \frac{5}{3}\eta^2\right) \\
&\quad + \frac{\kappa h_J'(1)}{d_J\sigma^5 C}(2\eta - 3\sigma^2) - \frac{\kappa h_J''(1)}{d_J\sigma^3 C} \\
&\quad + \frac{\delta}{6\sigma^7 C}(\sigma^4 - 3\sigma^2\rho + 6\sigma^2\eta + 10\eta^2) + O\left(\frac{1}{C^{3/2}}\right).
\end{aligned} \tag{3.24}$$

The boundary of the admissible set is asymptotically given in terms of $\alpha_1, \ldots, \alpha_J$ by substituting (3.23) and (3.24) into (3.19) and using (3.20)–(3.22).

4. Behavior of the boundary. We now scale by $\zeta_{J-1} = Z_{J-1}/\sqrt{C}$, where $Z_{J-1} = O(1)$. The goal is to express the boundary of the admissible set in the form (2.4). From (2.9) and (2.15) we have

$$\sum_{i=1}^{J-1} d_i(d_i - d_J)(\alpha_i - \alpha_i^{(0)}) = \frac{Z_{J-1}}{\sqrt{C}}\,. \tag{4.1}$$

But, from (2.6),

$$\sum_{j=1}^{J} d_j(\alpha_j - \alpha_j^{(0)}) = O\left(\frac{1}{\sqrt{C}}\right), \tag{4.2}$$

and, from (2.13) and (3.21),

$$\sigma^2 - \sigma_0^2 = 2\sum_{j=1}^{J} d_j^2(\alpha_j - \alpha_j^{(0)}). \tag{4.3}$$

If we use (4.2) to eliminate $\alpha_J - \alpha_J^{(0)}$ to lowest order, it follows from (4.1) that $\sigma^2 - \sigma_0^2 = O(1/\sqrt{C})$, and hence $\sigma = \sigma_0 + O(1/\sqrt{C})$.

From (3.19), (3.23), and (3.24),

$$\delta = \sigma w(\kappa/\sigma) + O\left(\frac{1}{\sqrt{C}}\right). \tag{4.4}$$

Hence,

$$\delta = \delta_0 + O\left(\frac{1}{\sqrt{C}}\right), \quad \delta_0 = \sigma_0 w(\kappa/\sigma_0), \tag{4.5}$$

and, from (3.20),

$$\sum_{j=1}^{J} d_j(\alpha_j - \alpha_j^{(0)}) = O\left(\frac{1}{C}\right). \tag{4.6}$$

If we use (4.6) to eliminate $\alpha_J - \alpha_J^{(0)}$, it follows from (4.1) and (4.3) that

$$\sigma = \sigma_0 + \frac{Z_{J-1}}{\sigma_0\sqrt{C}} + O\left(\frac{1}{C}\right). \tag{4.7}$$

Also, from (2.10) and (3.22), we obtain

$$\eta - \eta_0 = \sum_{j=1}^{J} d_j^3(\alpha_j - \alpha_j^{(0)}) = \zeta_{J-2} + O\left(\frac{1}{C}\right) \tag{4.8}$$

and

$$\rho - \rho_0 = \sum_{j=1}^{J-1} d_j(d_j^3 - d_J^3)(\alpha_j - \alpha_J^{(0)}) + O\left(\frac{1}{\sqrt{C}}\right). \tag{4.9}$$

Now, from (3.19), (3.23), and (3.24),

$$\begin{aligned}\delta = \sigma w\left(\frac{\kappa}{\sigma}\right) + \frac{2\delta^2\eta}{3\sigma^4\sqrt{C}} \\ + \frac{w'(\kappa/\sigma)}{\sqrt{C}}\left[\frac{\kappa\delta}{\sigma^2}\left(d_J - 1 - \frac{2\eta}{\sigma^2}\right) - \left(\frac{1}{2} + \frac{\eta}{3\sigma^2}\right)\right] + O\left(\frac{1}{C}\right).\end{aligned} \tag{4.10}$$

We define

$$\delta_1 = \frac{2\delta_0^2\eta_0}{3\sigma_0^4} + w'\left(\frac{\kappa}{\sigma_0}\right)\left[\frac{\kappa\delta_0}{\sigma_0^2}\left(d_J - 1 - \frac{2\eta_0}{\sigma_0^2}\right) - \left(\frac{1}{2} + \frac{\eta_0}{3\sigma_0^2}\right)\right], \tag{4.11}$$

$$a = \left\{\frac{1}{\sigma}\frac{d}{d\sigma}\left[\sigma w\left(\frac{\kappa}{\sigma}\right)\right]\right\}_{\sigma=\sigma_0} = \frac{1}{\sigma_0}\left[w\left(\frac{\kappa}{\sigma_0}\right) - \frac{\kappa}{\sigma_0}w'\left(\frac{\kappa}{\sigma_0}\right)\right], \tag{4.12}$$

and

$$b = \frac{2\delta_0^2}{3\sigma_0^4} - \frac{1}{\sigma_0^2}w'\left(\frac{\kappa}{\sigma_0}\right)\left(\frac{2\kappa\delta_0}{\sigma_0^2} + \frac{1}{3}\right). \tag{4.13}$$

Then, from (4.5), (4.7), (4.8), and (4.10), we obtain

$$\delta = \delta_0 + \frac{1}{\sqrt{C}}(\delta_1 + aZ_{J-1} + b\zeta_{J-2}) + O\left(\frac{1}{C}\right). \tag{4.14}$$

Next, from (3.20) and (4.14),

$$\sum_{j=1}^{J} d_j(\alpha_j - \alpha_j^{(0)}) = -\frac{1}{C}(aZ_{J-1} + b\zeta_{J-2}) + O\left(\frac{1}{C^{3/2}}\right). \tag{4.15}$$

If we use (4.15) to eliminate $\alpha_J - \alpha_J^{(0)}$, it follows from (4.1) and (4.3) that

$$\frac{1}{2}(\sigma^2 - \sigma_0^2) = \frac{Z_{J-1}}{\sqrt{C}} - \frac{d_J}{C}(aZ_{J-1} + b\zeta_{J-2}) + O\left(\frac{1}{C^{3/2}}\right). \tag{4.16}$$

Hence,

$$\sigma = \sigma_0 + \frac{Z_{J-1}}{\sigma_0\sqrt{C}} - \frac{1}{\sigma_0 C}\left[d_J(aZ_{J-1} + b\zeta_{J-2}) + \frac{Z_{J-1}^2}{2\sigma_0^2}\right] + O\left(\frac{1}{C^{3/2}}\right). \tag{4.17}$$

We are now in a position to obtain a further refinement of the expression for δ, from (3.19), (3.23), and (3.24), with the help of (4.8), (4.9), and (4.17). Some of the details are given in Appendix B. Since we are interested only in the quadratic terms, we do not give the linear terms explicitly. After considerable algebra, it is found that

$$\begin{aligned}
\delta = {} & (\text{linear terms}) - \frac{w(\kappa/\sigma_0)}{2\sigma_0^3 C} Z_{J-1}^2 \\
& + \frac{w''(\kappa/\sigma_0)}{2\sigma_0^5 C}\left[\kappa Z_{J-1} + \left(\frac{2\kappa\delta_0}{\sigma_0^2} + \frac{1}{3}\right)\zeta_{J-2}\right]^2 \\
& + \frac{w'(\kappa/\sigma_0)}{\sigma_0^5 C}\left\{\frac{1}{2}\kappa\sigma_0 Z_{J-1}^2 + 2\sigma_0 Z_{J-1}\zeta_{J-2}\left(\frac{1}{3} - \kappa a + \frac{4\kappa\delta_0}{\sigma_0^2}\right)\right. \\
& \qquad \left. + \zeta_{J-2}^2\left[5\left(\frac{2\kappa\delta_0^2}{\sigma_0^3} - \frac{\kappa}{3\sigma_0} + \frac{\delta}{3\sigma_0}\right) - 2\kappa\sigma_0 b\right]\right\} \\
& + \frac{4\delta_0}{3\sigma_0^4 C}\zeta_{J-2}\left[\left(a - \frac{2\delta_0}{\sigma_0^2}\right)Z_{J-1} + \left(b - \frac{4\delta_0^2}{3\sigma_0^4}\right)\zeta_{J-2}\right] + O\left(\frac{1}{C^{3/2}}\right).
\end{aligned} \tag{4.18}$$

The linear terms include δ_0 in (4.14).

We define

$$P = w\left(\frac{\kappa}{\sigma_0}\right) - \frac{\kappa}{\sigma_0}w'\left(\frac{\kappa}{\sigma_0}\right) - \left(\frac{\kappa}{\sigma_0}\right)^2 w''\left(\frac{\kappa}{\sigma_0}\right), \tag{4.19}$$

$$\begin{aligned}
Q = {} & \frac{4\delta_0}{3\sigma_0}\left(\frac{2\delta_0}{\sigma_0} - \sigma_0 a\right) - \frac{\kappa}{\sigma_0}w''\left(\frac{\kappa}{\sigma_0}\right)\left(\frac{2\kappa\delta_0}{\sigma_0^2} + \frac{1}{3}\right) \\
& - 2w'\left(\frac{\kappa}{\sigma_0}\right)\left(\frac{1}{3} - \kappa a + \frac{4\kappa\delta_0}{\sigma_0^2}\right),
\end{aligned} \tag{4.20}$$

and

$$\begin{aligned}
R = {} & \frac{8\delta_0}{3\sigma_0}\left(\frac{4\delta_0^2}{3\sigma_0^2} - \sigma_0^2 b\right) - w''\left(\frac{\kappa}{\sigma_0}\right)\left(\frac{2\kappa\delta_0}{\sigma_0^2} + \frac{1}{3}\right)^2 \\
& + 2w'\left(\frac{\kappa}{\sigma_0}\right)\left[2\kappa\sigma_0 b - 5\left(\frac{2\kappa\delta_0^2}{\sigma_0^3} - \frac{\kappa}{3\sigma_0} + \frac{\delta_0}{3\sigma_0}\right)\right].
\end{aligned} \tag{4.21}$$

Then, from (2.8) and (3.20), it follows that

$$\zeta_J = \frac{1}{2\sigma_0^5 C^{3/2}}\left(P\sigma_0^2 Z_{J-1}^2 + 2Q\sigma_0 Z_{J-1}\zeta_{J-2} + R\zeta_{J-2}^2\right) + O\left(\frac{1}{C^2}\right). \tag{4.22}$$

If we let

$$P = \sigma_0^3 P_2(\sigma_0), \quad Q = \sigma_0^4 R_2(\sigma_0), \quad R = \sigma_0^5 T_2(\sigma_0), \tag{4.23}$$

then we obtain (2.16). In Appendix C we show that the expressions for $P_2(\sigma_0)$ and $R_2(\sigma_0)$ are consistent with those obtained in [12].

In the next section we consider

$$D = PR - Q^2\,, \tag{4.24}$$

which is proportional to the discriminant of the quadratic form in (4.22).

5. Nonconvexity of the boundary. We first express P, Q, and R as functions of $w(\kappa/\sigma_0)$. From (3.18), we have

$$\phi[w(y)] = y. \tag{5.1}$$

It follows that

$$w' = \frac{1}{\phi'(w)}, \quad w'' = -\frac{\phi''(w)}{[\phi'(w)]^3}\,. \tag{5.2}$$

But, from (4.5), (4.12), and (4.13), with $w = w(\kappa/\sigma_0)$ and the argument w of ϕ suppressed,

$$\frac{\delta_0}{\sigma_0} = w, \quad \sigma_0 a = w - \frac{\phi}{\phi'} \tag{5.3}$$

and

$$\sigma_0^2 b = \frac{2}{3}w^2 - \frac{1}{\phi'}\left(2w\phi + \frac{1}{3}\right). \tag{5.4}$$

Then, from (4.19)–(4.21), we obtain

$$P = w - \frac{\phi}{\phi'} + \frac{\phi^2\phi''}{(\phi')^3}, \tag{5.5}$$

$$Q = \frac{4}{3}w^2 - \frac{2}{3\phi'}(1 + 7w\phi) - 2\left(\frac{\phi}{\phi'}\right)^2 + \frac{\phi\phi''}{(\phi')^3}\left(\frac{1}{3} + 2w\phi\right), \tag{5.6}$$

and

$$\begin{aligned} R = {} & \frac{16}{9}w^3 + \frac{\phi''}{(\phi')^3}\left(\frac{1}{3} + 2w\phi\right)^2 - \frac{4\phi}{(\phi')^2}\left(\frac{1}{3} + 2w\phi\right) \\ & + \frac{2}{\phi'}\left(\frac{5\phi}{3} - \frac{11w}{9} - 6w^2\phi\right). \end{aligned} \tag{5.7}$$

We now consider the behavior of P, Q, and R for $-w \gg 1$. From (3.15) and [8],

$$\phi(w) = -\frac{1}{2w} + \frac{1}{4w^3} - \frac{3}{8w^5} + O\left(\frac{1}{w^7}\right), \quad -w \gg 1. \tag{5.8}$$

Hence,

$$\phi'(w) = \frac{1}{2w^2} - \frac{3}{4w^4} + \frac{15}{8w^6} + O\left(\frac{1}{w^8}\right), \quad -w \gg 1, \tag{5.9}$$

and

$$\phi''(w) = -\frac{1}{w^3} + \frac{3}{w^5} - \frac{45}{4w^7} + O\left(\frac{1}{w^9}\right), \quad -w \gg 1. \tag{5.10}$$

After some straightforward algebra, which we omit, it is found that

$$\begin{aligned} P &= -\frac{2}{w^3} + O\left(\frac{1}{w^5}\right), \quad Q = -2 + O\left(\frac{1}{w^2}\right), \\ R &= -8w + O\left(\frac{1}{w}\right), \quad -w \gg 1. \end{aligned} \tag{5.11}$$

It follows from (4.24) that

$$D = -4 + O\left(\frac{1}{w^2}\right), \quad -w \gg 1. \tag{5.12}$$

Next, we consider the behavior of P, Q, and R for $w \gg 1$. From (3.15) and [8],

$$\phi(w) = \sqrt{\pi} e^{w^2} \left[1 - \frac{1}{2}\,\mathrm{Erfc}\,(w)\right] = \sqrt{\pi} e^{w^2} + O\left(\frac{1}{w}\right), \quad w \gg 1. \tag{5.13}$$

Hence,

$$\phi'(w) = 2\sqrt{\pi} w e^{w^2} + O\left(\frac{1}{w^2}\right), \quad w \gg 1, \tag{5.14}$$

and

$$\phi''(w) = 2\sqrt{\pi}(1 + 2w^2) e^{w^2} + O\left(\frac{1}{w^3}\right), \quad w \gg 1. \tag{5.15}$$

It follows from (5.5)–(5.7) that

$$\begin{aligned} P &= w + O\left(\frac{1}{w^3}\right), \quad Q = \frac{4}{3}(w^2 - 1) + O\left(\frac{1}{w^2}\right), \\ R &= \frac{16}{9} w^3 - 4w + O\left(\frac{1}{w}\right), \quad w \gg 1. \end{aligned} \tag{5.16}$$

Hence, from (4.24), we obtain

$$D = -\frac{4}{9} w^2 + O(1), \quad w \gg 1. \tag{5.17}$$

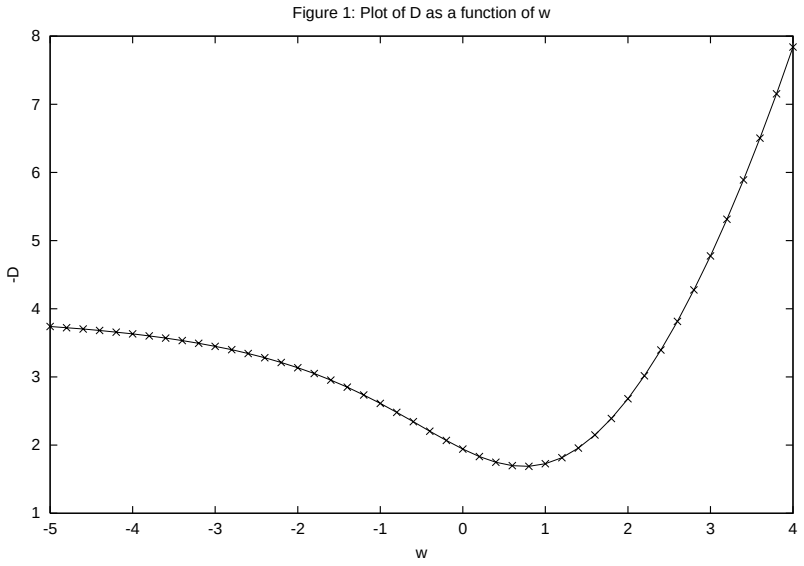


FIG. 1. *Plot of D as a function of w.*

If we differentiate (3.17) with respect to w and integrate by parts, we find that

$$\phi' = 1 + 2w\phi, \quad \phi'' = 2(\phi + w\phi'). \tag{5.18}$$

The quantities ϕ, ϕ', and ϕ'' can be evaluated numerically from (3.15) and (5.18). Then, P, Q, R, and D can be obtained numerically from (5.5)–(5.7) and (4.24). Figure 1 depicts $-D$ as a function of w for $-5 \leq w \leq 4$, and it is seen that $-D > 0$ in that range. Since, from (5.12) and (5.17), $D < 0$ for $-w \gg 1$ and for $w \gg 1$, we conclude that $D < 0$ for $-\infty < w < \infty$. This completes the proof of Proposition 2.2.

We note, as is evident from (5.11) and (5.16), that Q must change sign for some real value of w. If $D < 0$, then (4.24) implies that $PR < 0$ when $Q = 0$. However, we know that $P > 0$ for $-\infty < w < \infty$. Hence, $R < 0$ when $Q = 0$. Figure 2 depicts Q and R for $0 \leq w \leq 2$, and it is seen that $R < 0$ in a range of w which includes the value for which $Q = 0$.

Appendix A. We derive here the expressions in (3.23) and (3.24). If we substitute

$$z^* = 1 + \frac{c_1}{\sqrt{C}} + \frac{c_2}{C} + \frac{c_3}{C\sqrt{C}} + O\left(\frac{1}{C^2}\right) \tag{A.1}$$

in (3.2) and use (3.20)–(3.22), we obtain, after some algebra,

$$c_1 = \frac{2\delta}{\sigma^2}, \quad c_2 = \frac{2\delta^2}{\sigma^6}(\sigma^2 - 2\eta), \quad c_3 = \frac{4\delta^3}{3\sigma^{10}}\left[\sigma^4 - 2\sigma^2(\rho + 3\eta) + 12\eta^2\right]. \tag{A.2}$$

Since $f(1) = 0$ and $f'(z^*) = 0$,

$$\begin{aligned} 0 = f(z^*) + \frac{1}{2}(1 - z^*)^2 f''(z^*) + \frac{1}{6}(1 - z^*)^3 f^{(3)}(z^*) \\ + \frac{1}{24}(1 - z^*)^4 f^{(4)}(z^*) + O[(1 - z^*)^5]. \end{aligned} \tag{A.3}$$

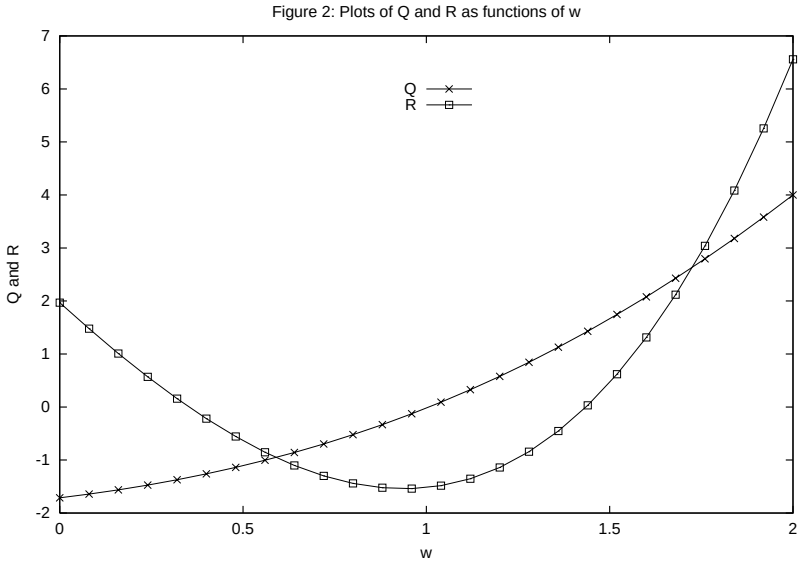


FIG. 2. *Plots of Q and R as functions of w.*

If we expand the expressions for the derivatives of $f(z)$ at $z = z^*$, given by (3.3)–(3.5), in powers of $z^* - 1$ and use (A.1) and (A.2), we obtain, after considerable algebra,

$$-Cf(z^*) = \frac{\delta^2}{\sigma^2}\left[1 - \frac{4\delta\eta}{3\sigma^4\sqrt{C}} + \frac{2\delta^2}{3\sigma^8 C}(6\eta^2 - \sigma^2\rho) + O\left(\frac{1}{C^{3/2}}\right)\right]. \tag{A.4}$$

This result was checked by using the Taylor series expansion of $f(z^*)$ about $z^* = 1$ and calculating the derivatives of $f(z)$ at $z = 1$ from (3.1). It follows from (A.1) and (A.2) that

$$\operatorname{sgn}(z^* - 1) = \operatorname{sgn}\delta. \tag{A.5}$$

Hence, from (A.4), since $\sigma > 0$, we obtain (3.23).

Next, from (3.3), (A.1), and (A.2), we find that

$$2v = \sigma^2\left[1 + \frac{4\delta\eta}{\sigma^4\sqrt{C}} + \frac{4\delta^2}{\sigma^8 C}(\sigma^2\rho - 2\eta^2) + O\left(\frac{1}{C^{3/2}}\right)\right], \tag{A.6}$$

and hence

$$\frac{1}{\sqrt{2v}} = \frac{1}{\sigma}\left[1 - \frac{2\delta\eta}{\sigma^4\sqrt{C}} + \frac{2\delta^2}{\sigma^8 C}(5\eta^2 - \sigma^2\rho) + O\left(\frac{1}{C^{3/2}}\right)\right]. \tag{A.7}$$

Also,

$$\frac{[(z^*)^{d_J} - 1]}{(z^* - 1)} = d_J + \frac{\delta}{\sigma^2\sqrt{C}} d_J(d_J - 1) + \frac{\delta^2}{3\sigma^6 C} d_J(d_J - 1)\left[(2d_J - 1)\sigma^2 - 6\eta\right] + O\left(\frac{1}{C^{3/2}}\right). \tag{A.8}$$

From (3.7), (A.7), and (A.8), we obtain

$$\begin{aligned}(\text{A.9})\ &\frac{z^*h_J(z^*)}{d_J\sqrt{2v}}\\ &=\frac{1}{\sigma}+\frac{\delta}{\sigma^5\sqrt{C}}\left[(d_J-1)\sigma^2-2\eta\right]\\ &+\frac{\delta^2}{\sigma^9C}\left[\frac{1}{3}(d_J-1)(2d_J-1)\sigma^4-4(d_J-1)\sigma^2\eta+2(5\eta^2-\sigma^2\rho)\right]+O\left(\frac{1}{C^{3/2}}\right).\end{aligned}$$

Next, from (3.4), (3.5), (3.21), and (3.22), we have

$$(\text{A.10})\qquad \tau=\eta-\frac{3}{2}\sigma^2+O\left(\frac{1}{\sqrt{C}}\right),\qquad y=\rho-6\eta+\frac{11}{2}\sigma^2+O\left(\frac{1}{\sqrt{C}}\right).$$

Hence, from (3.8), since $2v=\sigma^2+O(1/\sqrt{C})$, we obtain

$$\begin{aligned}(\text{A.11})\qquad \frac{N_J}{\sqrt{2v}}&=\frac{d_J}{\sigma^7}\left(\frac{1}{2}\sigma^2\rho+2\sigma^2\eta-\sigma^4-\frac{5}{3}\eta^2\right)\\ &\quad+\frac{h_J'(1)}{\sigma^5}(2\eta-3\sigma^2)-\frac{h_J''(1)}{\sigma^3}+O\left(\frac{1}{\sqrt{C}}\right).\end{aligned}$$

From (3.23),

$$\begin{aligned}(\text{A.12})\qquad &\frac{\operatorname{sgn}(z^*-1)}{\sqrt{-Cf(z^*)}}\\ &=\frac{\sigma}{\delta}+\frac{2\eta}{3\sigma^3\sqrt{C}}+\frac{\delta}{3\sigma^7C}(\sigma^2\rho-4\eta^2)+O\left(\frac{1}{C^{3/2}}\right),\end{aligned}$$

and, from (A.1) and (A.2),

$$\begin{aligned}(\text{A.13})\qquad \frac{1}{\sqrt{C}(z^*-1)}&=\frac{\sigma^2}{2\delta}-\frac{1}{2\sigma^2\sqrt{C}}(\sigma^2-2\eta)\\ &\quad+\frac{\delta}{6\sigma^6C}(\sigma^4+4\sigma^2\rho-12\eta^2)+O\left(\frac{1}{C^{3/2}}\right).\end{aligned}$$

Hence, from (A.7), we obtain

$$(\text{A.14})\qquad \frac{1}{\sqrt{2Cv}(z^*-1)}=\frac{\sigma}{2\delta}-\frac{1}{2\sigma\sqrt{C}}+\frac{\delta}{6\sigma^7C}(\sigma^4-2\sigma^2\rho+6\sigma^2\eta+6\eta^2)+O\left(\frac{1}{C^{3/2}}\right).$$

Finally, from (3.6), (A.12), and (A.14), we have

$$\begin{aligned}(\text{A.15})\qquad \frac{K}{\sqrt{2Cv}}&=\frac{1}{6\sigma^3\sqrt{C}}(2\eta+3\sigma^2)\\ &\quad-\frac{\delta}{6\sigma^7C}(\sigma^4-3\sigma^2\rho+6\sigma^2\eta+10\eta^2)+O\left(\frac{1}{C^{3/2}}\right).\end{aligned}$$

Now (3.24) follows from (A.9), (A.11), and (A.15).

Appendix B. We here give some of the details of the derivation of the expression for δ given in (4.18). From (4.14) and (4.17), it follows that

$$\frac{1}{\sigma} = \frac{1}{\sigma_0}\left\{1 - \frac{Z_{J-1}}{\sigma_0^2\sqrt{C}} + \frac{1}{\sigma_0^2 C}\left[d_J(aZ_{J-1} + b\zeta_{J-2}) + \frac{3Z_{J-1}^2}{2\sigma_0^2}\right]\right\} + O\left(\frac{1}{C^{3/2}}\right), \tag{B.1}$$

$$\frac{\delta}{\sigma^3\sqrt{C}} = \frac{\delta_0}{\sigma_0^3\sqrt{C}} + \frac{1}{\sigma_0^3 C}\left(\delta_1 + aZ_{J-1} + b\zeta_{J-2} - \frac{3\delta_0}{\sigma_0^2}Z_{J-1}\right) + O\left(\frac{1}{C^{3/2}}\right), \tag{B.2}$$

and, from (4.8),

$$\frac{\delta\eta}{\sigma^5\sqrt{C}} = \left[\frac{\delta_0}{\sigma_0^5\sqrt{C}} + \frac{1}{\sigma_0^5 C}\left(\delta_1 + aZ_{J-1} + b\zeta_{J-2} - \frac{5\delta_0}{\sigma_0^2}Z_{J-1}\right)\right](\eta_0+\zeta_{J-2}) + O\left(\frac{1}{C^{3/2}}\right) \tag{B.3}$$

and

$$\frac{\eta}{\sigma^3\sqrt{C}} = \frac{1}{\sigma_0^3\sqrt{C}}\left(1 - \frac{3Z_{J-1}}{\sigma_0^2\sqrt{C}}\right)(\eta_0 + \zeta_{J-2}) + O\left(\frac{1}{C^{3/2}}\right). \tag{B.4}$$

Hence, from (3.24), (4.9), and (B.1)–(B.4), we obtain

$$\begin{aligned}
&\frac{1}{\sqrt{2v}}\left\{\frac{\kappa}{d_J}\left[z^* h_J(z^*) + \frac{N_J}{C}\right] - \frac{K}{\sqrt{C}}\right\} \\
&= \frac{\kappa}{\sigma_0} - \frac{\kappa Z_{J-1}}{\sigma_0^3\sqrt{C}} + \frac{3\kappa Z_{J-1}^2}{2\sigma_0^5 C} + O\left(\frac{1}{C}\right) \times (\text{linear terms}) \\
&+ \frac{\kappa\delta_0}{\sigma_0^3\sqrt{C}}(d_J - 1) - \frac{2\kappa\delta_0}{\sigma_0^5\sqrt{C}}(\eta_0 + \zeta_{J-2}) \\
&\quad - \frac{2\kappa}{\sigma_0^5 C}\zeta_{J-2}\left(aZ_{J-1} + b\zeta_{J-2} - \frac{5\delta_0}{\sigma_0^2}Z_{J-1}\right) \\
&\quad - \frac{1}{3\sigma_0^3\sqrt{C}}(\eta_0 + \zeta_{J-2}) + \frac{Z_{J-1}\zeta_{J-2}}{\sigma_0^5 C} - \frac{1}{2\sigma_0\sqrt{C}} \\
&\quad + \frac{10\kappa\delta_0^2}{\sigma_0^9 C}\zeta_{J-2}^2 - \frac{5\kappa}{3\sigma_0^7 C}\zeta_{J-2}^2 + \frac{5\delta_0}{3\sigma_0^7 C}\zeta_{J-2}^2 + O\left(\frac{1}{C^{3/2}}\right).
\end{aligned} \tag{B.5}$$

Next, from (4.17) and (B.5), it follows that

$$\begin{aligned}
&\sigma w\left(\frac{1}{\sqrt{2v}}\left\{\frac{\kappa}{d_J}\left[z^* h_J(z^*) + \frac{N_J}{C}\right] - \frac{K}{\sqrt{C}}\right\}\right) \\
&= (\text{linear terms}) - \frac{w(\kappa/\sigma_0)}{2\sigma_0^3 C}Z_{J-1}^2 \\
&\quad + \frac{w'(\kappa/\sigma_0)}{\sigma_0^5 C}\left\{\frac{1}{2}\kappa\sigma_0 Z_{J-1}^2 + 2\sigma_0 Z_{J-1}\zeta_{J-2}\left(\frac{1}{3} - \kappa a + \frac{4\kappa\delta_0}{\sigma_0^2}\right)\right. \\
&\qquad\qquad \left. + \zeta_{J-2}^2\left[5\left(\frac{2\kappa\delta_0^2}{\sigma_0^3} - \frac{\kappa}{3\sigma_0} + \frac{\delta_0}{3\sigma_0}\right) - 2\kappa\sigma_0 b\right]\right\} \\
&\quad + \frac{w''(\kappa/\sigma_0)}{2\sigma_0^5 C}\left[\kappa Z_{J-1} + \left(\frac{2\kappa\delta_0}{\sigma_0^2} + \frac{1}{3}\right)\zeta_{J-2}\right]^2 + O\left(\frac{1}{C^{3/2}}\right).
\end{aligned} \tag{B.6}$$

Also, from (4.8), (4.14), and (4.17), we obtain

$$\frac{\delta^2\eta}{\sigma^4\sqrt{C}} = (\text{linear terms}) + \frac{2\delta_0}{\sigma_0^4 C}\zeta_{J-2}\left(aZ_{J-1} + b\zeta_{J-2} - \frac{2\delta_0}{\sigma_0^2}Z_{J-1}\right) + O\left(\frac{1}{C^{3/2}}\right) \tag{B.7}$$

and, from (4.9),

$$\frac{\delta^3}{\sigma^8 C}(16\eta^2 - 3\sigma^2\rho) = (\text{linear terms}) + \frac{16\delta_0^3}{\sigma_0^8 C}\zeta_{J-2}^2 + O\left(\frac{1}{C^{3/2}}\right). \tag{B.8}$$

The expression for δ in (4.18) now follows from (3.19), (3.23), (3.24), and (B.6)–(B.8).

Appendix C. We here reconcile the expressions for $P_2(\sigma_0)$ and $R_2(\sigma_0)$, given by (4.19), (4.20), and (4.23), with those obtained in [12]. Corresponding to [12], we let

$$w(\kappa/\sigma) = \chi(\sigma), \tag{C.1}$$

where $w(y)$ is given by (3.18). Then,

$$-\frac{\kappa}{\sigma^2}w'\left(\frac{\kappa}{\sigma}\right) = \chi'(\sigma), \qquad \frac{\kappa^2}{\sigma^4}w''\left(\frac{\kappa}{\sigma}\right) = \frac{2}{\sigma}\chi'(\sigma) + \chi''(\sigma). \tag{C.2}$$

Hence, from (4.19) and (4.23), we obtain

$$\sigma_0^3 P_2(\sigma_0) = \chi(\sigma_0) - \sigma_0\chi'(\sigma_0) - \sigma_0^2\chi''(\sigma_0), \tag{C.3}$$

which is consistent with the definition in [12].

Next, from (4.5), (4.12), (4.13), and (4.20), we find that

$$Q = \frac{4}{3}w^2 - \frac{2}{3}\left(1 + \frac{7\kappa}{\sigma_0}w\right)w' - 2\left(\frac{\kappa}{\sigma_0}w'\right)^2 - \frac{\kappa}{\sigma_0}\left(\frac{1}{3} + \frac{2\kappa}{\sigma_0}w\right)w'', \tag{C.4}$$

where we have suppressed the argument κ/σ_0 of w. As in [12], we let

$$G(\sigma) = \chi'(\sigma)\left[\frac{2\chi(\sigma)}{\sigma}\left\{1 - \frac{2}{3}[\chi(\sigma)]^2\right\} + \frac{1}{3\kappa}\{1 - 2[\chi(\sigma)]^2\}\right]. \tag{C.5}$$

Hence, from (C.1) and (C.2),

$$G(\sigma) = \frac{1}{\sigma^2}w'\left(\frac{\kappa}{\sigma}\right)\left[\frac{2\kappa}{\sigma}w\left(\frac{\kappa}{\sigma}\right)\left\{\frac{2}{3}\left[w\left(\frac{\kappa}{\sigma}\right)\right]^2 - 1\right\} + \frac{1}{3}\left\{2\left[w\left(\frac{\kappa}{\sigma}\right)\right]^2 - 1\right\}\right]. \tag{C.6}$$

But, from (5.1), (5.2), and (5.18), we have

$$w'\left(\frac{\kappa}{\sigma}\right)\left[1 + \frac{2\kappa}{\sigma}w\left(\frac{\kappa}{\sigma}\right)\right] = 1. \tag{C.7}$$

It follows from (C.6) that

$$G(\sigma) = \frac{2}{3\sigma^2}\left[w\left(\frac{\kappa}{\sigma}\right)\right]^2 - \frac{1}{\sigma^2}\left[\frac{1}{3} + \frac{2\kappa}{\sigma}w\left(\frac{\kappa}{\sigma}\right)\right]w'\left(\frac{\kappa}{\sigma}\right). \tag{C.8}$$

If we differentiate (C.8) and set $\sigma = \sigma_0$, we find from (4.23) and (C.4) that

$$\sigma_0 R_2(\sigma_0) = -G'(\sigma_0), \tag{C.9}$$

which is consistent with the definition in [12].

Acknowledgments. The author is indebted to Judy Seery for writing the programs for the numerical calculations and for obtaining the figures. He is also grateful to the referees for their helpful suggestions for improving the presentation. The main results of this paper have been stated, without proof, in [11].

REFERENCES

[1] E. Bouillet, D. Mitra, and K. G. Ramakrishnan, *The structure and management of service level agreements in networks*, IEEE J. Selected Areas Commun., 20 (2002), pp. 691–699.

[2] C. A. Courcoubetis, A. Dimakis, and M. I. Reiman, *Providing bandwidth guarantees over a best-effort network: Call admission and pricing*, in Proceedings of the Twentieth Annual Joint Conference of the IEEE Computer and Communication Societies, Vol. 1, IEEE Press, Piscataway, NJ, 2001, pp. 459–467..

[3] J. S. Kaufman, *Blocking in a shared resource environment*, IEEE Trans. Comm., 29 (1981), pp. 1474–1481.

[4] F. P. Kelly, *Reversibility and Stochastic Networks*, John Wiley, New York, 1980.

[5] K. Kumaran, M. Mandjes, D. Mitra, and I. Saniee, *Resource usage and charging in a multi-service multi-QoS packet network*, in Proceedings of the MIT/Tufts Workshop on Internet Service Quality Economics, December, 1999.

[6] S. S. Lam, *Queueing networks with population size constraints*, IBM J. Res. Develop., 21 (1977), pp. 370–378.

[7] S. Lanning, W. A. Massey, B. Rider, and Q. Wang, *Optimal pricing in queueing systems with quality of service constraints*, in Teletraffic Engineering in a Competitive World, Vol. 3B, North–Holland, Amsterdam, 1999, pp. 747–756.

[8] W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[9] D. Mitra and J. A. Morrison, *Erlang capacity and uniform approximations for shared unbuffered resources*, IEEE/ACM Trans. Networking, 2 (1994), pp. 558–570.

[10] M. Montgomery and G. de Veciana, *Hierarchical source routing through clouds*, in Proceedings of the Seventeenth Annual Joint Conference of the IEEE Computer and Communication Societies, Vol. 2, IEEE Press, Piscataway, NJ, 1998, pp. 685–692.

[11] J. A. Morrison and D. Mitra, *Asymptotic shape of the Erlang capacity region of a multiservice shared resource*, Perform. Eval., 49 (2002), pp. 273–281.

[12] J. A. Morrison and D. Mitra, *Asymptotic shape of the Erlang capacity region of a multiservice shared resource*, SIAM J. Appl. Math., to appear.

[13] J. A. Morrison and K. G. Ramakrishnan, *Asymptotic solution to an inverse problem for a shared unbuffered resource*, SIAM J. Appl. Math., 63 (2002), pp. 222–240.

[14] J. A. Morrison, K. G. Ramakrishnan, and D. Mitra, *Refined asymptotic approximations to loss probabilities and their sensitivities in shared unbuffered resources*, SIAM J. Appl. Math., 59 (1998), pp. 494–513.

[15] S. Subramaniam, M. Azizoglu, and A. K. Somani, *All-optical networks with sparse wavelength conversion*, IEEE/ACM Trans. Networking, 4 (1996), pp. 544–557.

[16] T. Tripathi and K. N. Sivarajan, *Computing approximate blocking probabilities in wavelength routed all-optical networks with limited-range wavelength conversion*, in Proceedings of the Eighteenth Annual Joint Conference of the IEEE Computer and Communication Societies, Vol. 1, IEEE Press, Piscataway, NJ, 1999, pp. 329–336.

[17] Q. Wang, J. M. Peha, and M. A. Sirbu, *Optimal pricing for integrated services networks*, in Internet Economics, L. W. McKnight and J. P. Bailey, eds., MIT Press, Cambridge, MA, 1997, pp. 352–376.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 18–40

STEADY MOTION OF A DROP ALONG A LIQUID INTERFACE*

JAMES J. KRIEGSMANN† AND MICHAEL J. MIKSIS†

Abstract. We investigate the motion of a liquid drop as it flows along the interface of a liquid film. Steady, two-dimensional solutions are found in the lubrication limit for a horizontal and inclined plane. The effects of the physical parameters on the interface shapes are studied. When the plane is inclined, solutions are found only for a special set of parameter values.

Key words. thin films, triple junctions, free boundaries

AMS subject classifications. 76Dxx, 76Txx

DOI. 10.1137/S0036139901400215

1. Introduction.

Spreading of one liquid over another occurs in many interesting and important physical processes, with applications in various fields. Specific examples include liquid waste spills on bodies of water (e.g., oil spreading on the sea or chemical waste spills on ponds), spills into partially saturated porous materials, polymer-polymer coextrusion, and aerosol delivery of bronchial medicated mists. In each of these situations, the spreading process is strongly influenced by the surface tension of the liquid interfaces and the physics of the triple junction, where the three phases intersect. In order to obtain a better understanding of the effect of the triple junction on the dynamics of the flow, we consider here a model problem of the spreading of a liquid droplet along a thin liquid film flowing down an inclined plane. We assume that the fluids are immiscible and that the motion is two-dimensional, and we obtain steady solutions in the lubrication limit. These steady solutions are found only for a limited set of parameter values when the plane is inclined. The effects of the physical parameters on the steady solutions are investigated.

If a drop of one liquid is placed on top of a second immiscible liquid, a three-phase point can exist at the gas/liquid/liquid intersection. We refer to this point as the triple junction. Clearly, for a droplet resting on a liquid interface in two dimensions, there are two triple junctions. The boundary conditions imposed at the triple junction have a major influence on how one liquid spreads over another. For example, if the spreading coefficient is positive, an equilibrium solution is impossible without additional assumptions [14]. If a droplet is spreading over a base liquid bounded above by air, a positive spreading coefficient, $S > 0$, indicates that the surface tension of the base liquid with air, Σ^F, is larger than the sum of the surface tension of the droplet with air, Σ^D, and the surface tension between the liquid and the droplet, Σ^{DF}; i.e., $S = \Sigma^F - \Sigma^D - \Sigma^{DF} > 0$. Because of the relative strength of the surface tension Σ^F, the droplet will completely wet the second liquid in such a case. Such a situation has been investigated by DiPietro, Huh, and Cox [3], DiPietro and Cox [4], and Foda and Cox [5], who developed a theory for the spreading of a droplet in the completely wetting case. Their model included the additional effect of a leading precursor (monolayer) film. The addition of the precursor film to the model allowed

*Received by the editors December 26, 2001; accepted for publication (in revised form) February 11, 2003; published electronically October 2, 2003. This work was supported in part by NSF grant DMS-0104935.

http://www.siam.org/journals/siap/64-1/40021.html

†Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, Illinois 60208 (miksis@northwestern.edu).

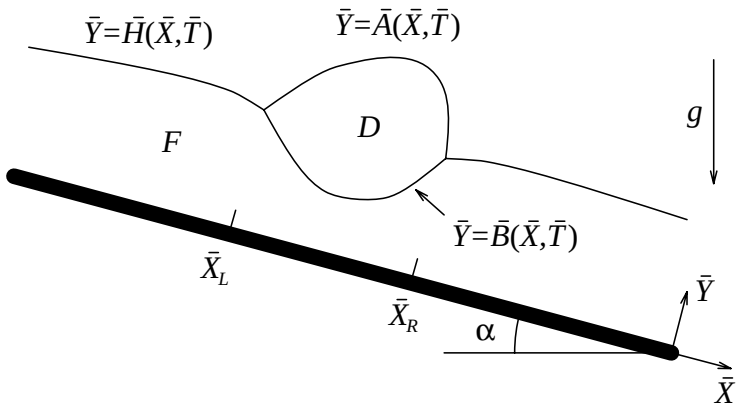


FIG. 1. *Sketch of the physical system.*

them to obtain steady and similarity solutions. The positive spreading coefficient case has also been investigated by Joanny [8] in the lubrication limit. By adding van der Waals forces into his model, he was able to make statements concerning steady state solutions and spreading rates. In particular, he showed that the radius of a wetting droplet will increase like $t^{1/7}$, where t is time. These results were confirmed experimentally by Fraaije and Cazabat [6]. Also recently, Brochard-Wyart, Debregeas, and de Gennes [2] have examined the spreading of a viscous droplet on a nonviscous liquid and determined that the droplet radius should increase like $t^{1/4}$.

Less work has been done for negative spreading coefficients, $S < 0$. In this case an equilibrium situation is possible without the additional assumption of a monolayer precursor film. An example of such a situation would be seen in a water droplet on top of a pool of carbon tetrachloride or nearly any other organic liquid. Equilibrium solutions of droplets resting on a liquid interface were computed by Pujado and Scriven [13]. Recently, a lubrication model was used by Wilson and Williams [17] to study the problem of a dragged film emerging through the free surface of a second liquid. They determined the final thickness of the coating film as a function of the density ratio and the surface tension of the interfaces. Also recently, a similarity solution for the dynamics of a triple junction was investigated by Miksis and Vanden-Broeck [11]. They were able to determine the location of the triple junction and the resulting capillary waves along the interface as a function of the physical parameters. The models used in these investigations forced equilibrium boundary conditions, i.e., zero net force, at the triple junction. The zero-net-force condition will also be assumed here.

2. Physical description. We consider the two-dimensional flow of a two-phase system consisting of a liquid drop floating on a liquid substrate which completely coats a flat solid surface, as illustrated in Figure 1. The fluids are Newtonian, incompressible, and immiscible, forming a well defined interface between the drop and film. We assume that the gas above the drop and film is passive, with sufficiently small viscosity and density so as to impose no effect upon the system.

Our primary concern is to study the behavior of the interfaces: $\bar{H}(\bar{X},\bar{T})$, the surface of the film; $\bar{A}(\bar{X},\bar{T})$, the upper surface of the drop; and $\bar{B}(\bar{X},\bar{T})$, the interface between the drop and film. Here, $\bar{X}$ is the spatial coordinate parallel to the solid surface, and $\bar{T}$ is time. The surface tensions associated with the interfaces are Σ^F

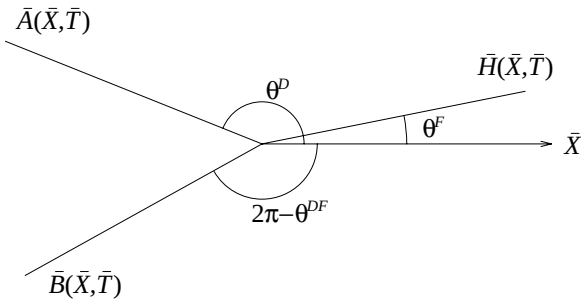


FIG. 2. *Region near the right contact point.*

for the film, Σ^D for the drop, and Σ^{DF} for the interface between the drop and film. These surface tensions are such that

$$\frac{1}{\Sigma^F}\left[\Sigma^F - \Sigma^D - \Sigma^{DF}\right] = S < 0. \tag{1}$$

Here, S is the dimensionless spreading parameter. We assume that Σ^F, Σ^D, Σ^{DF}, and S are well defined for our two liquids, and that they remain constant for all time. The validity of these assumptions is discussed in [7].

Consider the two triple junctions, or contact points, $\bar{X} = \bar{X}_{L,R}(\bar{T})$. We assume that the three interfaces meet here at a well defined massless point. The point bears no mass, and so the surface tension forces must sum to zero, in accordance with Newton's laws. Using the notation shown in Figure 2, this fact is written as

$$\begin{aligned} \Sigma^F \sin\theta^F + \Sigma^D \sin\theta^D + \Sigma^{DF} \sin\theta^{DF} &= 0, \\ \Sigma^F \cos\theta^F + \Sigma^D \cos\theta^D + \Sigma^{DF} \cos\theta^{DF} &= 0. \end{aligned} \tag{2}$$

This force balance is sometimes illustrated through the use of the Neumann triangle, a discussion of which can be found in [14].

Our task is now to determine the dynamics of the droplet illustrated in Figure 1. The liquids are assumed to be viscous, incompressible, and to obey the Navier–Stokes equations. The boundary conditions are no-slip along the solid walls, the continuity of tangential stress at the liquid interfaces, plus the condition that the jump in normal stress is given by the surface tension times the curvature of the interface. Finally, the boundary conditions (2) are assumed to hold at each of the triple junctions. This is a difficult free boundary problem. To simplify matters, we will assume that the liquids are thin, so that the lubrication approximation may be applied. The resulting simplified system of equations for the interface shapes can then be more readily solved.

3. Lubrication approximation. Beyond considering our spreading parameter to be negative, we further assume that it is very small in magnitude,

$$-S = \epsilon^2 \ll 1.$$

If the spreading parameter were identically zero, then (2) would indicate that $\theta^D = \theta^{DF} = \theta^F + \pi$. That is, in the neighborhood of the contact point, the three interfaces would fall upon one shared line. Because S does not vanish, but is very close to zero, we expect that $\left(\theta^D - \theta^F - \pi\right)$ and $\left(\theta^{DF} - \theta^F - \pi\right)$ should not vanish, but they should be very small. If we assume that θ^F is small in magnitude, then θ^D

and θ^{DF} must be very near π, and so all the interface slopes are very small at the contact points.

Since $\bar{H}(\bar{X},\bar{T})$, $\bar{A}(\bar{X},\bar{T})$, and $B(\bar{X},\bar{T})$ all vary slowly in $\bar{X}$ near the triple junctions, we seek solutions for $\bar{H}$, $\bar{A}$, and $\bar{B}$ that vary slowly in $\bar{X}$ everywhere. This suggests that we use a long-wave theory, or, equivalently, that we use the lubrication equations. In order to formally derive the lubrication equations, we introduce two distinct length scales. If the thickness of the film or drop can be characterized by a distance d, we seek solutions that vary over a distance of L, which might characterize the length of the drop, and we seek solutions for which $d/L = \epsilon \ll 1$. Choosing this lubrication ratio equal to $\sqrt{-S}$ leads to a consistent, simplified system. The details of this derivation for our problem can be found in [9]. Here we give only an outline of the derivation and the resulting nonlinear system of evolution equations for the film interfaces.

In order to obtain the leading order lubrication equations in the small parameter ϵ, the Navier–Stokes equation along with the boundary conditions are made dimensionless by the change of variables

$$\begin{aligned} X &= \frac{\bar{X}}{L}, \qquad y = \frac{\bar{Y}}{d}, \qquad T = \frac{\epsilon\rho^F g d\bar{T}}{3\mu^F}, \\ h &= \frac{\bar{H}}{d}, \qquad b = \frac{\bar{B}}{d}, \qquad a = \frac{\bar{A}}{d}, \\ X_L &= \frac{\bar{X}_L}{L}, \quad X_R = \frac{\bar{X}_R}{L}. \end{aligned}$$

Here μ^i and ρ^i are the viscosity and density of the film ($i = F$) and drop ($i = D$), and g is the gravitational acceleration. Following the standard lubrication assumptions [1], [9], the velocity in the direction normal to the plane is assumed to be one order higher in ϵ than the velocity tangent to the plane. In addition, we need to introduce the dimensionless density ratio, viscosity ratio, and Reynolds number,

$$\beta = \frac{\rho^D}{\rho^F}, \quad \lambda = \frac{\mu^D}{\mu^F}, \quad Re = \frac{\epsilon^2(\rho^F)^2 g L^3}{3(\mu^F)^2}.$$

We assume that this Reynolds number is $o(1)$. The dimensionless parameter

$$C = \frac{\epsilon\Sigma^F}{\rho^F g L^2}$$

measures the importance of surface tension. We assume that this parameter, which is the reciprocal of a Bond number, is $O(1)$. Finally, we introduce dimensionless surface tension ratios

$$\begin{aligned} \frac{\Sigma^D}{\Sigma^F} &= \sigma^D, \\ \frac{\Sigma^{DF}}{\Sigma^F} &= \sigma^{DF} + \epsilon^2. \end{aligned}$$

We seek solutions of the dependent variables in the dimensionless equations of motion as regular perturbation expansions in ϵ. To order $O(\epsilon)$, the resulting partial differential equations that describe the evolution of the interfaces are (see [9])

$$
\begin{aligned}
h_T &= \frac{\partial}{\partial X}\left\{h^3\left[\epsilon\cos\alpha\, h_X - C h_{XXX} - \sin\alpha\right]\right\},\\
b_T &= \frac{\partial}{\partial X}\Bigg\{b^3\left[\epsilon(1-\beta)\cos\alpha\, b_X - C\sigma^{DF} b_{XXX}\right]\\
&\qquad + \frac{1}{2}b^2(3a-b)\left[\epsilon\beta\cos\alpha\, a_X - C\sigma^{D} a_{XXX}\right]\\
&\qquad + \sin\alpha\left[\left(\frac{3\beta}{2}-1\right)b^3 - \frac{3\beta}{2}ab^2\right]\Bigg\},\\
a_T &= \frac{\partial}{\partial X}\Bigg\{\frac{1}{2}b^2(3a-b)\left[\epsilon(1-\beta)\cos\alpha\, b_X - C\sigma^{DF} b_{XXX}\right]\\
&\qquad + \left[\left(\frac{1-\lambda}{\lambda}\right)(a-b)^3 + a^3\right]\left[\epsilon\beta\cos\alpha\, a_X - C\sigma^{D} a_{XXX}\right]\\
&\qquad + \sin\alpha\left[-\frac{\beta}{\lambda}a^3 + 3\beta\left(\frac{1-\lambda}{\lambda}\right)a^2 b + \left(\frac{3\beta}{2}-\frac{1}{2}-\frac{\beta}{\lambda}\right)b^2(3a-b)\right]\Bigg\}.
\end{aligned}
\tag{3}
$$

Note that the derivation of these equations parallels the calculation for a single liquid thin film along a substrate; see, e.g., Oron, Davis, and Bankoff [12]. The difference in the calculation occurs in the region where there are two liquid interfaces. The leading order equations from the Navier–Stokes equations are similar to the single-phase case, but when the boundary conditions across the drop/fluid interface are applied, the fluid motion in both liquid regions becomes coupled, resulting in the second two equations in (3).

In order to solve the evolution equations (3), we must supply boundary conditions at the contact points (triple junctions), far-field information, and initial conditions. It is important to keep in mind that the leading order equations we derive have an associated length scale. The equations are therefore really valid only away from the triple junctions, and care also needs to be taken when discussing the far-field. Hence, a proper matched asymptotic analysis needs to be done that accounts for the behavior of the solutions in these different regions. The leading order contact and balance of force conditions at the triple junction follow from a straightforward matched asymptotic analysis, while the derivation of the other conditions local to the triple junction can be found in [9].

The first condition imposed at the triple junctions is continuity of the interfaces. To leading order at $X = X_{R,L}$ these are

$$
\begin{aligned}
h &= b,\\
h &= a.
\end{aligned}
\tag{4}
$$

The balance of surface tension forces (2) must also be imposed at the triple junctions. At $O(1)$, the force balance at the contact line (2) implies

$$
1 - \sigma^{D} - \sigma^{DF} = 0,
\tag{5}
$$

while by pursuing the same equations to $O(\epsilon)$ and $O(\epsilon^2)$, we obtain the boundary conditions at $X = X_{R,L}$,

$$
\begin{aligned}
h_X - \sigma^{D} a_X - \sigma^{DF} b_X &= 0,\\
h_X^2 - \sigma^{D} a_X^2 - \sigma^{DF} b_X^2 &= -2.
\end{aligned}
$$

These conditions affect our leading order system and may be rephrased as

$$
\begin{aligned}
b_X &= h_X - \sqrt{2\frac{\sigma^D}{\sigma^{DF}}} \quad \text{at } X = X_L, \\
a_X &= h_X + \sqrt{2\frac{\sigma^{DF}}{\sigma^D}} \quad \text{at } X = X_L, \\
b_X &= h_X + \sqrt{2\frac{\sigma^D}{\sigma^{DF}}} \quad \text{at } X = X_R, \\
a_X &= h_X - \sqrt{2\frac{\sigma^{DF}}{\sigma^D}} \quad \text{at } X = X_R.
\end{aligned} \tag{6}
$$

The final conditions at the contact points are determined by imposing continuity of pressure and horizontal velocity through the triple junction. To $O(\epsilon^2)$, these conditions imply

$$
\begin{aligned}
h_{XX} - \sigma^D a_{XX} - \sigma^{DF} b_{XX} &= 0, \\
C\left[h_{XXX} - \sigma^D a_{XXX} - \sigma^{DF} b_{XXX}\right] &= \epsilon \cos\alpha \left[h_X - \beta a_X - (1-\beta) b_X\right].
\end{aligned} \tag{7}
$$

We must also supply some information about $h(X,T)$ far upstream and downstream. We will seek solutions for which $h(X,T)$ tends to a constant height, h_{up}, far upstream. Initial conditions for the problem consist of the interface shapes themselves at $T = 0$. A final piece of information that proves to be useful is the speed of the contact points, $\frac{\mathrm{d}}{\mathrm{d}T} X_{L,R}$. It is clear that these speeds must be equivalent to the horizontal component of the velocity of the fluid in the drop at these points. It can be shown that

$$
\frac{\mathrm{d}}{\mathrm{d}T} X_{L,R} = \frac{3}{2}\left[h^2\left(C h_{XXX} + \sin\alpha - \epsilon\cos\alpha\, h_X\right)\right]\Big|_{X=X_{L,R}}. \tag{8}
$$

4. Steady equations and rescaling.

4.1. Equations of motion. The system (3)–(8) has been derived in part by introducing two different length scales, L and d. We have used as our dimensionless horizontal coordinate $X = \bar{X}/L$, and $T = \epsilon\rho^F g d\bar{T}/3\mu^F$ as our dimensionless measure of time. If, instead, we use d as our only length scale, and if we employ

$$
x = \frac{\bar{X}}{d}, \qquad t = \frac{\rho^F g d \bar{T}}{3\mu^F}
$$

as our dimensionless variables, then (3)–(8) are somewhat transformed. Introducing

$$
Bo = \frac{\epsilon^3}{C} = \frac{\rho^F g d^2}{\Sigma^F}
$$

as the Bond number and

$$
x_{r,l}(t) = \frac{\bar{X}_{R,L}}{d}
$$

as the dimensionless locations of the contact points, the transformed system of the partial differential equations can be obtained.

We wish to investigate here only the possibility of steady solutions. Such a solution represents a drop of constant shape moving at a constant speed over the substrate.

The conditions under which such a solution is possible must be determined, as must the speed of the drop and the shape of all interfaces.

The system resulting from the above rescaling may be posed in a moving frame by substituting $x_{\rm new} = x_{\rm old} + Ut$, where U is the speed of either contact point. For steady solutions, the shape of each interface does not vary in time, and so all time derivatives vanish. Upon integrating the resulting equations, we find that

$$h^3\left[Bo^{-1}h''' - \cos\alpha\, h' + \sin\alpha\right] - Uh = -M,$$

$$\begin{aligned} &b^3\left[\frac{\sigma^{DF}}{Bo}b''' - (1-\beta)\cos\alpha\, b'\right] + \frac{1}{2}b^2(3a-b)\left[\frac{\sigma^{D}}{Bo}a''' - \beta\cos\alpha\, a'\right] \\ &\qquad + \sin\alpha\left[\left(1-\frac{3\beta}{2}\right)b^3 + \frac{3\beta}{2}ab^2\right] - Ub = -M, \end{aligned} \tag{9}$$

$$\begin{aligned} &\frac{1}{2}b^2(3a-b)\left[\frac{\sigma^{DF}}{Bo}b''' - (1-\beta)\cos\alpha\, b'\right] \\ &+\left[\left(\frac{1-\lambda}{\lambda}\right)(a-b)^3 + a^3\right]\left[\frac{\sigma^{D}}{Bo}a''' - \beta\cos\alpha\, a'\right] \\ &+\sin\alpha\left[\frac{\beta}{\lambda}a^3 - 3\beta\left(\frac{1-\lambda}{\lambda}\right)a^2b + \left(\frac{3\beta}{2} - \frac{1}{2} - \frac{\beta}{\lambda}\right)b^2(b-3a)\right] - Ua = -M, \end{aligned}$$

where primes denote differentiation with respect to x. The boundary conditions (4) and (7) have been used in the integration of the differential equations (3) to evaluate the constant of integration $M = Uh_{up} - h_{up}^3\sin\alpha$, which is the flux of the lower film. The fact that the film thickness approaches the constant h_{up} far upstream has also been used. The boundary conditions needed to supplement (9) at the contact points are

$$\begin{aligned} h - a &= 0, \\ h - b &= 0, \\ b' &= h' \mp \sqrt{-2S\frac{\sigma^{D}}{\sigma^{DF}}}\,, \\ a' &= h' \pm \sqrt{-2S\frac{\sigma^{DF}}{\sigma^{D}}}\,, \\ h'' - \sigma^{D}a'' - \sigma^{DF}b'' &= 0. \end{aligned} \tag{10}$$

It should be noted that, by evaluating the first of the equations (9) at the contact points and using the expressions (8), the frame speed U can be simply written as

$$U = \frac{3h_{up}^3\sin\alpha}{3h_{up} - h(x_l)} = \frac{3h_{up}^3\sin\alpha}{3h_{up} - h(x_r)}, \tag{11}$$

which implies

$$h(x_l) = h(x_r) \tag{12}$$

so long as $\alpha \neq 0$. We see that, for a solution to be steady, the height of both contact points must be the same.

In seeking steady solutions, we need not concern ourselves with the initial conditions of the system. We must, however, specify the volume of the drop, V:

$$\int_{x_l}^{x_r} (a-b)\mathrm{d}x = V. \tag{13}$$

More properly, of course, this volume is actually an area.

A few points should be kept in mind as we solve the system (9)–(13). First, we have assumed that S is very small in deriving our system, and so a small value should be used if the results are to be meaningful. Second, we should expect that our solutions will vary slowly in x. For example, the slopes of our solutions such as h_x should be $O(\sqrt{|S|})$, since we have assumed that the slopes written as h_X would be $O(1)$. Third, because our solutions should be slowly varying in space, we expect surface tension effects to be important when Bo is small, namely, $O(|S|^{3/2})$. Finally, in deriving our system, we have retained terms to $O(\epsilon)$. The $O(\epsilon)$ quantities have manifested themselves in the $\cos\alpha$ terms. Later, the importance of these presumably small terms is investigated.

4.2. Far-field conditions. In order to solve (9), we must provide some far-field information. Far upstream away from the drop, we have assumed that $h(x)$ approaches the value h_{up}. Far downstream, we expect $h(x)$ to tend to a limiting value h_{dn}. From the first of the equations (9), we see that h_{dn} must satisfy the algebraic equation

$$\sin\alpha\, h_{dn}^3 - U h_{dn} + U h_{up} - \sin\alpha\, h_{up}^3 = 0,$$

which has solutions

$$h_{dn} = h_{up}, \qquad -\frac{h_{up}}{2} \pm \frac{\sqrt{3}}{2} h_{up} \sqrt{\frac{h_{up} + h(x_l)}{3h_{up} - h(x_l)}}\,. \tag{14}$$

We observe that at most two of these solutions are positive and physically significant. If $h(x)$ does tend to a limiting value far downstream, it must asymptote to one of these two allowable values if there is to be a steady solution. A discussion of third order ODEs similar to these can be found in Tuck and Schwartz [15].

In an effort to understand the behavior of $h(x)$ for large values of $|x|$, we seek solutions of the form $h(x) \sim h_\infty + \zeta(x)$, where h_∞ is one of the two values h_{up} or h_{dn} and $\zeta \ll 1$. We substitute this form into the steady equation (9) and retain only linear terms in ζ. The equation admits solutions $\zeta = e^{rx}$, where r satisfies the characteristic equation

$$r^3 - Bo\cos\alpha\, r + \frac{3Bo\sin\alpha[2h_\infty - h(x_l)]}{h_\infty[3h_\infty - h(x_l)]} = 0,$$

which has solutions $r = r_1,\ r_2,\ r_3$. This fact is written as $(r-r_1)(r-r_2)(r-r_3) = 0$, or, by expanding,

$$r^3 - (r_1 + r_2 + r_3)r^2 + (r_1 r_2 + r_1 r_3 + r_2 r_3)r - r_1 r_2 r_3 = 0.$$

By comparing this cubic equation with the characteristic equation, we first can note that $r_1 + r_2 + r_3 = 0$. Also, we see that the sign of $r_1 r_2 r_3$ is determined by the sign of the quotient

$$Q = \frac{\sin\alpha[h(x_l) - 2h_\infty]}{3h_\infty - h(x_l)}. \tag{15}$$

If $Q \neq 0$, then we can be sure that none of the roots are zero. Because we are solving a cubic equation with real coefficients, we are certain that at least one of the roots is real. The fact that the sum of the roots vanishes suggests that one of the roots, r_1, has a negative real part, while another of the roots, r_2, has a positive real part.

If $Q < 0$, then the product of the roots is negative. For the product of the roots to be negative, r_1 must be real, and r_3 must be either the complex conjugate of r_2 or a positive real number, provided that r_2 is real. In summary, r_1 is a negative real number, while the other roots are either positive real numbers or complex conjugates with positive real parts.

If, for the time being, we assume that our steady solution has a contact point height such that $h(x_l) < 2h_{up}$, we see that $Q < 0$. By the preceding analysis, we see in this case that there is one growing mode as $x \to -\infty$. Far downstream, $h(x)$ can approach one of two values, h_{dn}. If, for the time being, we assume that our steady solution has a contact point height such that $h(x_l) < 2h_{dn}$ as well, we see that $Q < 0$. By the preceding analysis, we see in this case that there are two growing modes as $x \to \infty$. Additional discussion of the roots for nonnegative values of Q can be found in Kriegsmann [9].

We will apply the conclusions of the linear analysis to our nonlinear problem and assume that $Q < 0$. To suppress the growth mode far upstream, we require only that

$$\lim_{x \to -\infty} h'(x) = 0, \tag{16}$$

while far downstream we require that

$$\begin{aligned} \lim_{x \to \infty} h'(x) &= 0, \\ \lim_{x \to \infty} h(x) &= h_{dn}, \end{aligned} \tag{17}$$

to eliminate both growth modes. It should be understood that these boundary conditions are to be imposed numerically on a finite computational domain. Note that we have assumed that the upstream height h_{up} and the downstream height h_{dn} both take on values such that $Q < 0$. This fact must be checked in all calculated solutions. Later, the possibility for solutions where this constraint is not met is discussed.

Our steady problem is now fully stated. The differential equations (9) must be solved with boundary conditions (10), extra conditions (11)–(13), and numerical boundary conditions (16)–(17). It should be noted that there are eight free parameters, α, Bo, σ^D, S, β, λ, h_{up}, and V, while $\sigma^{DF} = 1 - \sigma^D$, as given by (5).

Finally, we should remark about the behavior of the interfaces near the contact point. For the steady problem, Kriegsmann [9] has shown that the interfaces have finite first, second, and third derivatives at the triple junction, but in the time dependent case only the second derivative is known to be finite. He also showed that this model implies integrable stresses in the neighborhood of the triple junction.

5. Zero-tilt-angle solutions.

5.1. Equations. We now consider steady solutions when $\alpha = 0$. These are presented to illustrate the effects of the physical parameters in our model. Pujado and Scriven [13] calculated such steady solutions for a more general drop-film system. They permited the interfaces to vary on a length scale similar to the thickness, and they also found both two-dimensional and three-dimensional solutions. We consider only two-dimensional solutions that vary on a length scale much larger than the thickness.

For any such solution, the frame speed U and the flux M must vanish. The steady equations (9) simplify considerably. In order to solve the system, an origin must be selected. This is done by selecting the midpoint of the drop so that the contact points are located at $x = \pm x_c$. In addition to the conditions at the contact points and the volume constraint we must also impose the constraint $\lim_{x\to-\infty} h_0 = h_{up}$. Note that we will denote the zero-tilt-angle solutions with a zero subscript. The resulting system is linear in h_0, b_0, and a_0, but it is nonlinear in x_c. This can be seen by rescaling x so that the contact points would be located at $x = \pm 1$. The differential equations would then have factors of x_c^2 present.

The zero-tilt-angle version of (9) can be integrated, yielding

$$
\begin{aligned}
h_0 &= C_h + C_{hh} e^{-\sqrt{Bo}\,x} \quad \text{for } x > x_c, \\
h_0 &= C_h + C_{hh} e^{\sqrt{Bo}\,x} \quad \text{for } x < -x_c, \\
b_0 &= \begin{cases} C_b + C_{bb} \cosh \sqrt{\frac{(1-\beta)Bo}{\sigma^{DF}}}\, x & \text{if } \beta < 1, \\ C_b + C_{bb} \cos \sqrt{\frac{(\beta-1)Bo}{\sigma^{DF}}}\, x & \text{if } \beta > 1, \\ C_b + C_{bb} x^2 & \text{if } \beta = 1, \end{cases} \\
a_0 &= C_a + C_{aa} \cosh \sqrt{\frac{\beta Bo}{\sigma^D}}\, x,
\end{aligned}
\tag{18}
$$

where we have retained only symmetric terms. Enforcing the boundary conditions at only one contact point and also the condition for $x \to -\infty$, we are left with seven algebraic equations to solve for the seven unknown variables C_h, C_{hh}, C_b, C_{bb}, C_a, C_{aa}, and x_c. This algebraic system is linear in all the variables except x_c and is solved numerically (see Kriegsmann [9] for details).

One should note that, of our eight free parameters for the steady problem, α, Bo, σ^D, S, β, λ, h_{up}, and V, we set α to zero, but the rest are free. For some values of the parameters, the solutions might not be realizable. For example, the drop might dip below the solid surface. In general, however, we may choose the parameters as we please. One should also note that the zero-tilt-angle solutions do not depend upon our choice of λ in any way. This could be anticipated. Since there is no fluid motion, viscosity should not come into play, and λ is a measure of viscosity.

5.2. Solutions. Some zero-tilt-angle solutions are shown in Figure 3. The film height monotonically approaches the value h_{up} both upstream and downstream, though this is not obvious in every graph. The views of the drop and film are chosen as such so that the drop can be seen in greater detail. Also, the labeling conventions shown in the first graph hold for all of the solutions shown in this paper.

In an effort to understand the importance of the density ratio, β is varied while the other dimensionless parameters are held constant. As we consider drops which are progressively more dense, the drops sink progressively lower into the liquid substrate. In the case when $\beta = 0.7$, we see that the film has a constant height $h(x) = h_{up}$ for $|x| > x_c$. This is true in any situation when $\beta = \sigma^D$ (see [9]). If $\beta \neq \sigma^D$, the film height $h(x)$ is not constant but is an exponential function, as stated earlier. If $\beta > \sigma^D$, we see that the contact points dip below h_{up}, and if $\beta < \sigma^D$, the contact points rest above h_{up}.

Next, we investigate the importance of the relative surface tensions by varying σ^D. Figure 4 shows such solutions. Recalling the relationship (5) between σ^D and

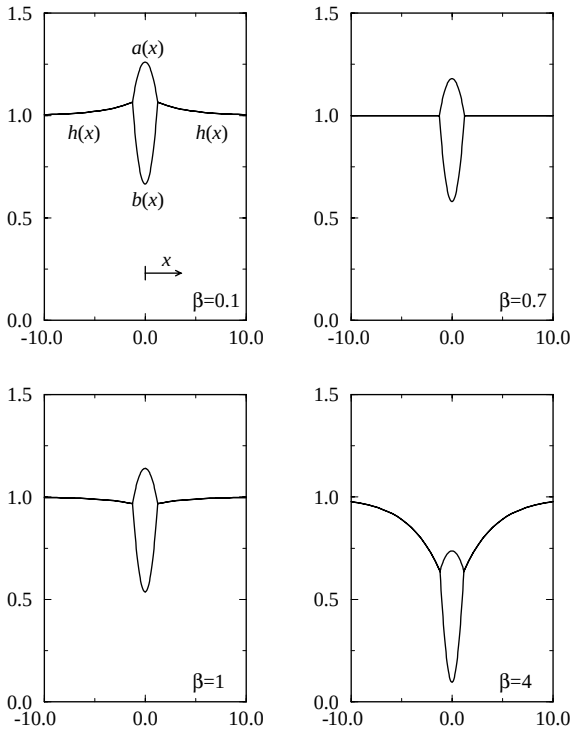


FIG. 3. *Zero-tilt-angle solutions for* $Bo = 0.1$, $\sigma^D = 0.7$, $S = -0.1$, $h_{up} = 1$, *and* $V = 1$. *The density ratio* β *is varied as shown.*

σ^{DF}, $\sigma^D + \sigma^{DF} = 1$, as the tension of the upper surface of the drop is increased, the tension of the lower surface of the drop is necessarily decreased. This is not to say that the surface tensions are dependent on one another; rather a limitation of our model demands that we consider only interactions between liquids for which our assumption (5) is met, i.e., interactions with small spreading coefficients.

That having been said, it can be seen from the solutions that when the tension on the upper surface of the drop is relatively low, the upper surface deforms more dramatically than the lower surface, as expected. On the other hand, if the tension on the upper surface of the drop is relatively high, the upper surface deforms less dramatically than the lower surface.

Figure 5 illustrates the dependence of solutions on the spreading parameter S. It should be remembered that S is the basis of our lubrication approximation. We have agreed to study cases in which S is very small, and we have argued that because of the small magnitude, solutions should vary slowly in x. Because we have derived our equations in this limit, it is pushing the limits of reason to consider a solution to our equations if $S = -1$. Nevertheless, the solutions depicted in Figure 5 show that as the magnitude of S is increased, the width of the drops becomes shorter. This is to be expected, as we have assumed that the ratio of the film thickness to the width of the drop is $O(\sqrt{-S})$.

Finally, we investigate the importance of overall surface tension by varying the Bond number. As the value of Bo is decreased, the surface tension associated with each interface becomes greater and the deformation of each interface is lessened, as shown in Figure 6.

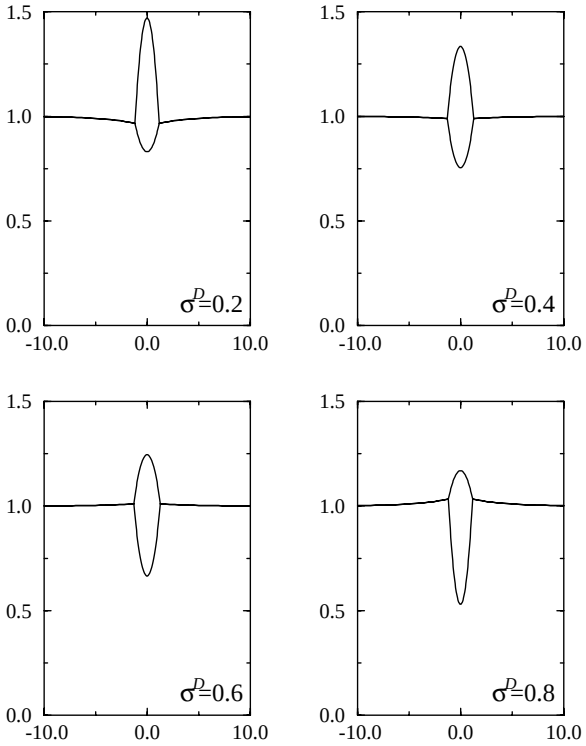


FIG. 4. *Zero-tilt-angle solutions for* $Bo = 0.1$, $S = -0.1$, $\beta = 0.5$, $h_{up} = 1$, *and* $V = 1$. *The surface tension ratio* σ^D *is varied as shown.*

The solutions shown thus far are all accurate solutions of the mathematical system (9)–(13), though they do not all necessarily reflect accurate physical solutions. In the extreme case shown in Figure 3, with $\beta = 4$, we have found a mathematical solution for which the slopes are not small. Nevertheless, by pressing the domains of validity to the extreme, we are easily able to see trends in the solutions.

6. Finite-tilt-angle solutions. We now seek steady solutions when α is greater than zero. The zero-tilt-angle solutions we have found thus far all share the trait that the film height far downstream tends to the upstream height, i.e., $h_{dn} = h_{up}$. Although we will not impose this far-field condition, the solutions we have found for nonzero tilt-angle all share this trait. Hence in this section we will use h_{up} when referring to both the downstream and upstream heights. Here we will assume that the height of the contact point is less than $2h_{up}$, and so it is necessary to apply the boundary conditions (16), (17).

6.1. Numerical method. In order to numerically approximate solutions, we follow a multiple-step procedure. Since we are free to select an origin, we pick it halfway between the contact points so that $x_r = -x_l = x_c$. Next, we guess at the values of $h(x_c)$, $h'(-x_c)$, $b'(x_c)$, $a'(x_c)$, and x_c that are to be held by the steady solution. We further recall that $b(x_c)$, $a(x_c)$, $b(-x_c)$, $a(-x_c)$, $h(-x_c)$, and $h(x_c)$ are all equal by the boundary conditions (10), (12).

We now solve for the interface shapes individually. Here we only briefly outline the numerical method; details are given in Kriegsmann [9]. To begin, we solve the first

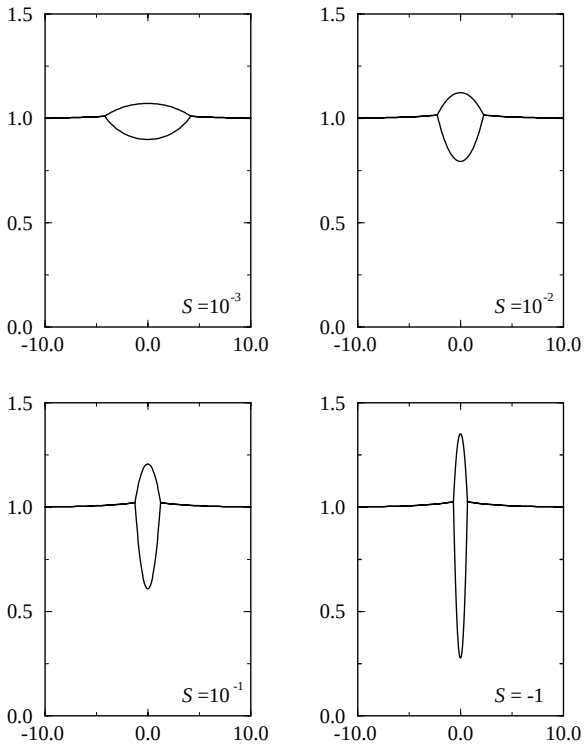


FIG. 5. *Zero-tilt-angle solutions for* $Bo = 0.1$, $\sigma^D = 0.7$, $\beta = 0.5$, $h_{up} = 1$, *and* $V = 1$. *The spreading parameter* S *is varied as shown.*

of the differential equations (9) for $x < -x_c$, given the facts that $h'(x)$ approaches zero for large $|x|$ and that $h(-x_c)$ and $h'(-x_c)$ take on their guessed values. As mentioned earlier, for the film to the left of the drop we impose two boundary conditions at the contact point and one at the end of the computational domain. Next, we solve the same differential equation for $x > x_c$, given the facts that $h(x)$ tends to h_{up} for large x, its slope similarly tends to zero, and that $h(x_c)$ takes on its assumed value. Again as previously mentioned, for the film to the right of the drop, we impose one boundary condition at the contact point and two conditions at the end of the computational domain. Finally, we solve the two coupled differential equations for a and b in (9), with the conditions that $b(-x_c)$, $a(-x_c)$, $b(x_c)$, and $a(x_c)$ take on their guessed height, while $b'(x_c)$ and $a'(x_c)$ are equal to their assumed values.

Given these tentative interface solutions, we check to see whether they satisfy all of the conditions (10), (13). By our construction, the first two conditions of (10) are satisfied at each contact point, leaving us with three conditions at each contact point in addition to the volume constraint. We have made a guess for five different scalars, and we see that seven equations (constraints) must be satisfied. We suspect that there is not always a solution for a given set of physical parameters. In this case, we expect that we may freely select six of the eight physical parameters, for example, α, Bo, σ^D, S, h_{up}, and V, and that there exists a solution only for certain values of λ and β. The viscosity ratio λ and the density ratio β are selected arbitrarily as the two auxiliary dependent variables. The values of the seven scalars, λ, β, $h(x_c)$, $h'(-x_c)$, $b'(x_c)$, $a'(x_c)$, and x_c, are then systematically adjusted by a multidimensional Newton's method to satisfy the seven constraints.

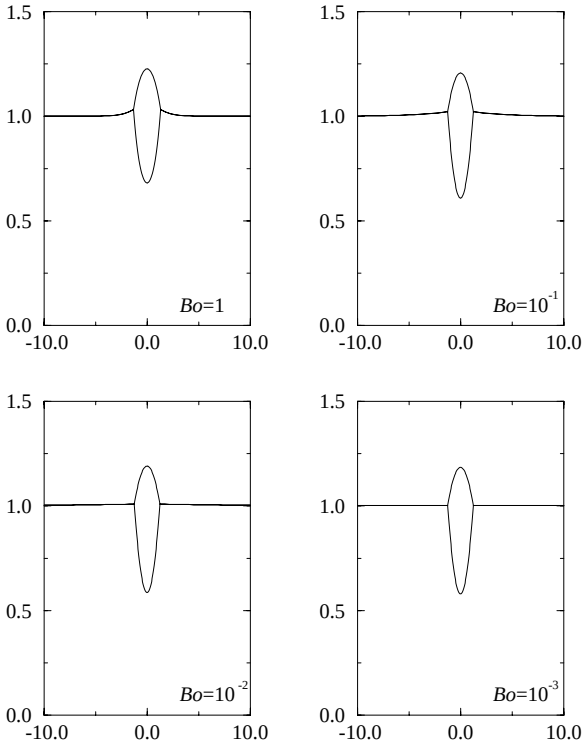


FIG. 6. *Zero-tilt-angle solutions for* $\sigma^D = 0.7$, $S = -0.1$, $\beta = 0.5$, $h_{up} = 1$, *and* $V = 1$. *The Bond number Bo is varied as shown.*

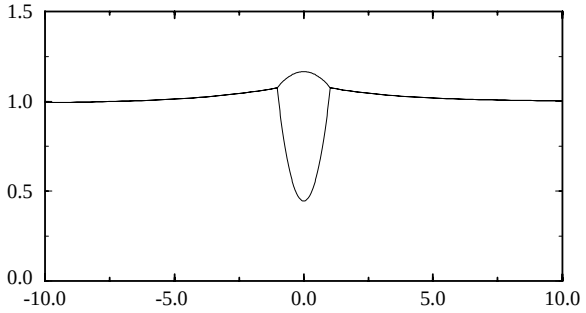


FIG. 7. *Steady solution for* $\alpha = 0.05$, $Bo = 0.1$, $\sigma^D = 0.9$, $S = -0.1$, $h_{up} = 1$, $V = 1$, $\beta = 0.172$, *and* $\lambda = 0.148$.

6.2. Numerical results.

6.2.1. Small-angle solutions. We begin by considering a steady solution for a very small tilt-angle. Figure 7 shows a steady solution for $\alpha = 0.05$, $Bo = 0.1$, $\sigma^D = 0.9$, $S = -0.1$, $h_{up} = 1$, $V = 1$, $\beta = 0.172$, and $\lambda = 0.148$. In this graph, as in most of the graphs to follow in this paper, the computational domain is not shown in its entirety, so that the details near the drop may be observed. Figure 8 illustrates an even closer view of the same solution compared with a solution to the linearized problem, as found by Kriegsmann [9], and a solution to the zero-tilt-angle problem.

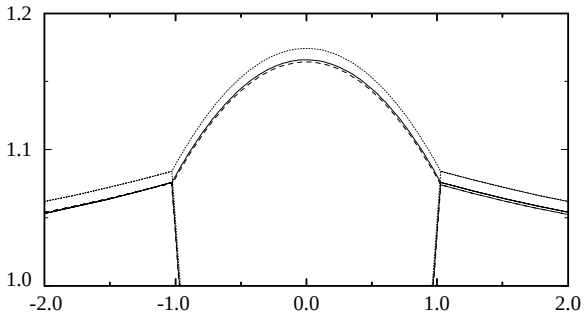


FIG. 8. *Comparison between the zero-tilt-angle solution (dotted), solution to the linearized problem (dashed), and solution to the nonlinear problem (solid).*

By linearized, we imply solutions close to the zero-tilt-angle results, valid in the limit of a small tilt-angle. For the zero-tilt-angle problem, all the same parameters are used, with the exception of $\alpha = 0$. For the linearized problem, the same parameter values are used except for the viscosity ratio, for which the required value is slightly altered, $\lambda = 0.150$. We can see that the solution to the linearized problem is, as expected, a small perturbation away from the zero-tilt-angle solution. The solution to the fully nonlinear problem is reasonably close to the linear solution in this case, for $\alpha = 0.05$. This is to be expected, as solutions to the linearized problem should be accurate in the limit $\alpha \ll 1$.

6.2.2. Large-tilt-angle solutions I. Several nonlinear solutions are shown in Figure 9. The surface tension ratio σ^D is varied, and so the required values of β and λ change as well. In varying σ^D, we see in the steady solutions the same trend present in the zero-tilt-angle solutions. That is, when the surface tension on the upper surface of the drop is relatively low, the upper surface deforms more dramatically than the lower surface. Conversely, when the surface tension on the upper surface of the drop is relatively high, the upper surface deforms less dramatically than the lower surface of the drop.

In fact, the other trends discussed earlier in reference to the zero-tilt-angle solutions are present in the nonlinear finite-tilt-angle solutions as well. As the value of S is raised, the width of the drops becomes shorter. As the Bond number is decreased, the deformation of each interface is lessened. Although these trends are shared by the zero-tilt-angle and finite-tilt-angle solutions, and in truth the solutions even look quite similar, there are quantitative differences. Also, although it is not obvious to the naked eye, the finite-tilt-angle solutions, unlike the zero-tilt-angle solutions, are asymmetric.

As repeatedly mentioned, we have been able to find steady translating solutions only for select sets of physical parameters. In the eight-dimensional parameter space described by α, Bo, σ^D, S, h_{up}, V, β, and λ, we have found steady solutions only on a six-dimensional manifold. To gain an understanding of this manifold, Figures 10–13 reveal numerous cross sections. Each point on the β vs. σ^D and λ vs. σ^D curves represents a steady solution. The final graph in each figure shows the speed of the drop.

Several features can be seen in Figures 10–13. For example, solutions have been found only for $\beta < \sigma^D$. This may be related to the behavior of zero-tilt-angle solutions. The concavity of the zero-tilt-angle steady film interfaces, $h_0(x)$, depends only upon

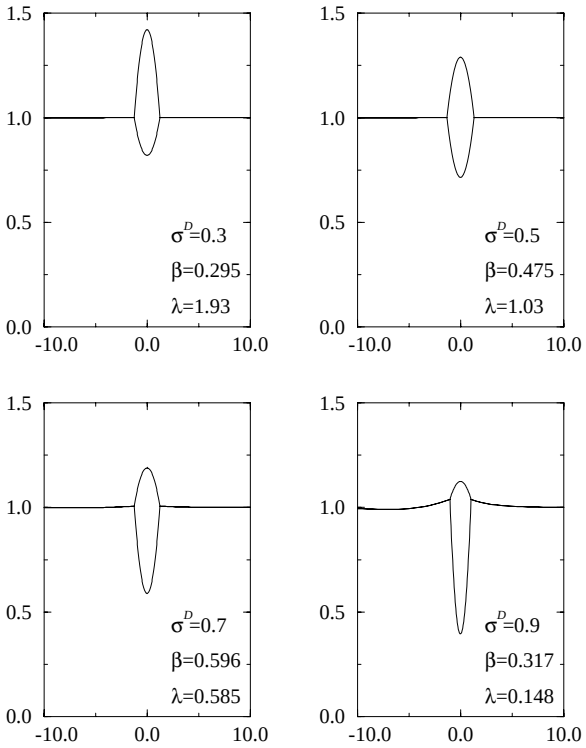


FIG. 9. *Steady solutions for* $\alpha = 0.5$, $Bo = 0.1$, $S = -0.1$, $h_{up} = 1$, *and* $V = 1$. *The values of* σ^D, β, *and* λ *are varied as shown.*

the sign of $(\beta - \sigma^D)$. Furthermore, for reasonably small values of σ^D, we find solutions only if β is very near σ^D and if λ approaches a value which depends almost entirely on σ^D. In this limit, we also observe that the steady solutions found for $\alpha \neq 0$ are almost identical to the zero-tilt-angle solutions found for the same set of parameters, but with $\alpha = 0$. We see from Figure 14 that, for relatively large σ^D, the value of β is far from σ^D, and the steady solution differs somewhat from the zero-tilt-angle solution. Figure 15 shows that, for relatively small σ^D, the value of β is close to σ^D, and the steady solution is almost identical to the zero-tilt-angle solution.

Figures 10–13 further reveal that the necessary value of β depends strongly upon the values of σ^D and h_{up} but only weakly upon the value of the Bond number. Also, we may note that the required value of λ depends weakly upon all parameters except σ^D and β. Finally, we can see that the speed of the drop depends almost exclusively upon α and h_{up}.

Given that the steady solutions never stray too far from the zero-tilt-angle solutions, this final fact is somewhat intuitive. The speed of the drop has been written as

$$U = \frac{3h_{up}^3 \sin\alpha}{3h_{up} - h\left(x_c\right)}.$$

For the zero-tilt-angle solutions, the contact point height $h\left(x_c\right)$ is seen to depend most strongly on the relation between β and σ^D (see Figures 3–6). The steady solutions, which are similar to the corresponding zero-tilt-angle solutions, all include as part

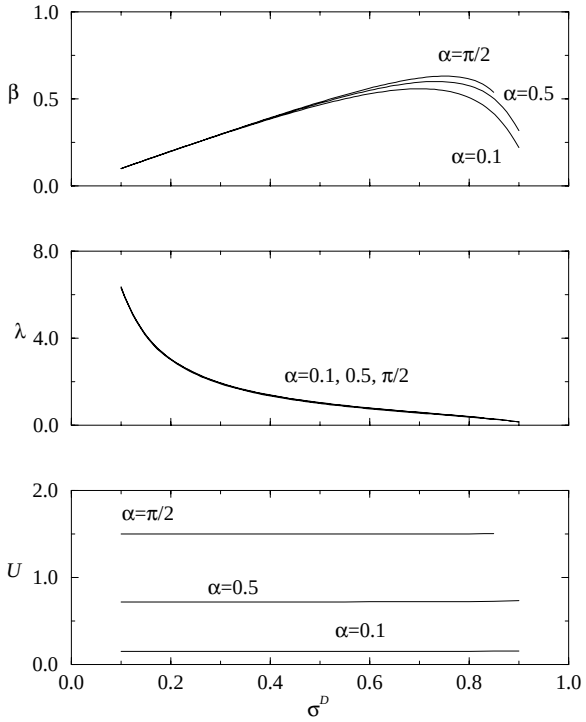


FIG. 10. *Given $Bo = 0.1$, $S = -0.1$, $h_{up} = 1$, $V = 1$, and α, σ^D having values as shown, steady solutions are found only for the special values of β, λ shown. Also, the speed of the drop is displayed.*

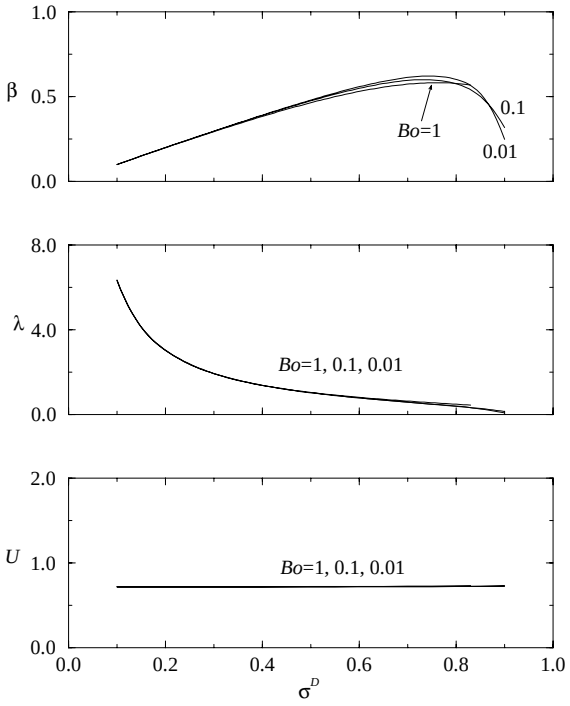


FIG. 11. *Given $\alpha = 0.5$, $S = -0.1$, $h_{up} = 1$, $V = 1$, and Bo, σ^D having values as shown, steady solutions are found only for the special values of β, λ shown. Also, the speed of the drop is displayed.*

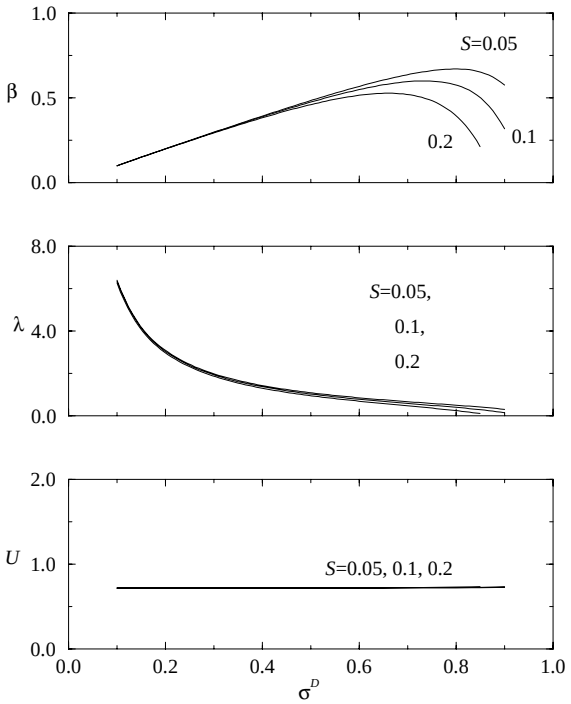


FIG. 12. *Given $\alpha = 0.5$, $Bo = 0.1$, $h_{up} = 1$, $V = 1$, and S, σ^D having values as shown, steady solutions are found only for the special values of β, λ shown. Also, the speed of the drop is displayed.*

of the solution a value of β such that the contact point height $h(x_c)$ is on the same order as the upstream film thickness h_{up}. Since this is true, the drop speed can be approximated:

$$U \approx \frac{3}{2} h_{up}^2 \sin\alpha.$$

This value is the same as the speed of a fluid element on the surface of a perfectly flat film of thickness h_{up}, and it matches the speeds shown in Figures 10–13 almost exactly.

6.2.3. Large-tilt-angle solutions II. The steady finite-tilt-angle solutions found thus far have been to some degree tame. The solutions exist only for a restricted choice of parameters, which conspire to keep the interfaces near those of the corresponding zero-tilt-angle solutions. In some cases, the steady finite-tilt-angle solutions are indistinguishable from the zero-tilt-angle solutions, and in all cases, the drop moves down the plane almost exactly with the surface speed of the undisturbed film.

In order to locate each of these steady solutions, we have considered λ and β as dependent variables. We can also seek solutions by designating different parameters as dependent variables. By choosing h_{up} and λ as such, we are now free to specify the value of β. Many of the solutions described by Figures 10–13 require that β be very near σ^D, and these solutions are somewhat uninteresting, in that they very closely resemble zero-tilt-angle solutions. We may now dictate that β be far from σ^D and hope for more interesting solutions.

Figure 16 shows a family of steady solutions, which are calculated by specifying all the parameters except h_{up} and λ. By comparing the steady solutions for $\alpha = 0.5$

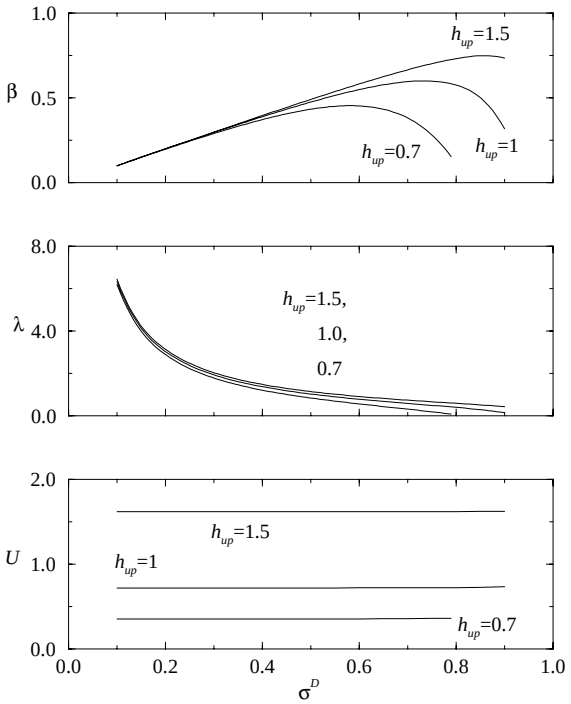


FIG. 13. *Given $\alpha = 0.5$, $Bo = 0.1$, $S = -0.1$, $V = 1$, and h_{up}, σ^D having values as shown, steady solutions are found only for the special values of β, λ shown. Also, the speed of the drop is displayed.*

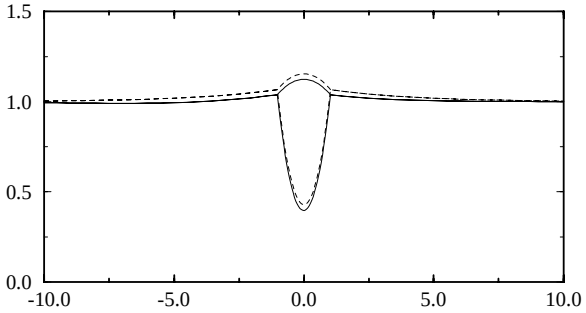


FIG. 14. *Steady solutions for $Bo = 0.1$, $\sigma^D = 0.9$, $S = -0.1$, $h_{up} = 1$, $V = 1$, $\beta = 0.317$, and $\lambda = 0.148$. The tilt-angle is set to $\alpha = 0.5$ (solid) and $\alpha = 0$ (dashed).*

with the zero-tilt-angle solutions for $\alpha = 0$, it is clear once again that, if the value of β is near that of σ^D, the steady solution and zero-tilt-angle solution are nearly identical. By reducing β far below σ^D, we see that the steady solution does differ somewhat from the zero-tilt-angle solution, though the two configurations are quite similar in character.

It is interesting to note that for the solutions shown in Figure 16, the required value of h_{up} increases with β. On the other hand, the solutions themselves are still unexciting, in that they closely resemble zero-tilt-angle solutions. Also, steady solutions still have not been found for $\beta > \sigma^D$.

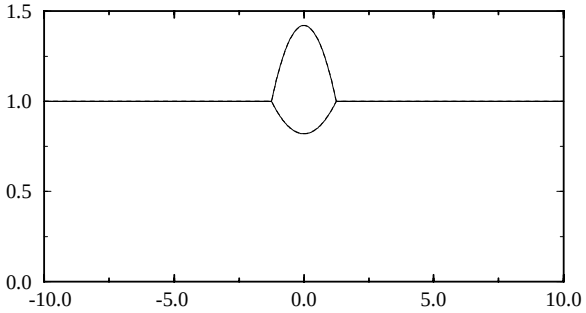


FIG. 15. *Steady solutions for* $Bo = 0.1$, $\sigma^D = 0.3$, $S = -0.1$, $h_{up} = 1$, $V = 1$, $\beta = 0.295$, *and* $\lambda = 1.93$. *The tilt-angle is set to* $\alpha = 0.5$ *(solid) and* $\alpha = 0$ *(dashed). The two solutions are nearly indistinguishable.*

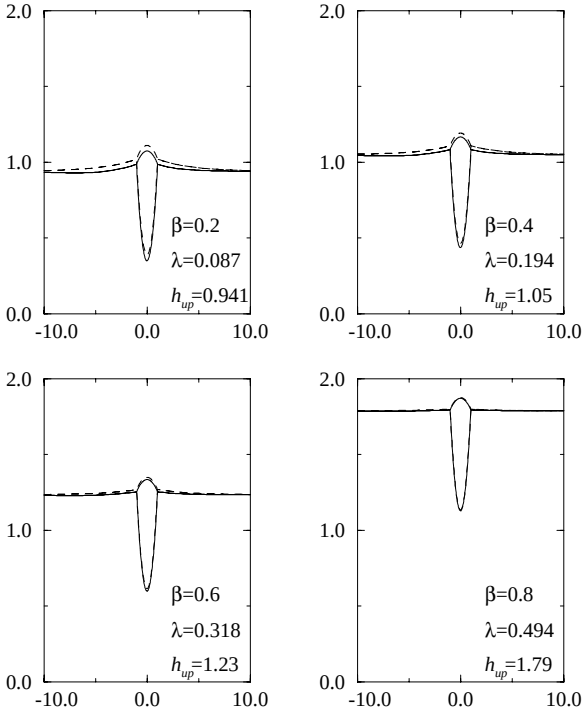


FIG. 16. *Steady solutions for* $Bo = 0.1$, $\sigma^D = 0.9$, $S = -0.1$, *and* $V = 1$. *The values of* β, λ, *and* h_{up} *are varied as shown. Solutions for* $\alpha = 0.5$ *(solid) are compared to the zero-tilt-angle solutions with* $\alpha = 0$ *(dashed).*

6.2.4. Importance of $O(\epsilon)$ terms. In section 4 it is observed that, in deriving our system, we retained terms to $O(\epsilon)$, and that these $O(\epsilon)$ effects are felt in the $\cos\alpha$ terms. Very little difference is noticed in the solutions obtained with and without the $O(\epsilon)$ term, except for α near zero; this is illustrated in Figure 17, where we plot the values that β and λ must assume for the existence of a steady solution if the $O(\epsilon)$ terms are retained or discarded. We see that for $\alpha = \pi/2$ the $\cos\alpha$ terms obviously have no effect. For α very near zero, we see that the removal of the $\cos\alpha$ terms changes the results noticeably. In such a limit, we could not have been justified in neglecting these terms in the formulation of our lubrication equations (seeAcheson

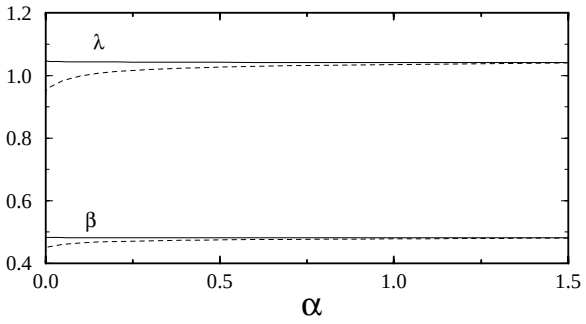


FIG. 17. *Required values of β and λ for existence of solution. The angle α is varied, while $Bo = 0.1$, $\sigma^D = 0.5$, $S = -0.1$, $h_{up} = 1$, and $V = 1$. The $O(\epsilon)$ terms are discarded (solid) and retained (dashed).*

[1]). For intermediate values of α, the inclusion of the $O(\epsilon)$ terms only perturbs our results. It seems proper to have included the $\cos\alpha$ terms in our formulation. The terms are necessary for very small α, and for larger values of α the $O(\epsilon)$ terms simply provide small corrections.

6.2.5. A singular limit. At this point, it seems as though we have reached another unexpected result. We have found that, although it is possible to realize a steady zero-tilt-angle solution for any given values of Bo, σ^D, S, h_{up}, V, β, and λ, if α is increased slightly, then we can find a steady solution only if the parameters take on special values. Figure 17 shows that, although we can find a zero-tilt-angle solution for any values of β and λ, this limit is singular, since β and λ take on single limiting values as $\alpha \to 0$.

To view this result from another angle, consider the case when $Bo = 0.1$, $\sigma^D = 0.5$, $S = -0.1$, $h_{up} = 1$, $V = 1$, $\beta = 0.465$, and $\lambda = 0.999$. If $\alpha = 0$, a steady solution can be found. If α is increased slightly, our claim thus far is that there is no steady solution. If α is further increased to 0.1, then we find that there is a steady solution. Our computations show that the profiles of these two solutions are nearly identical, yet if we perturb α slightly away from 0.1, there is no steady solution. Alternatively, if we perturb any other parameter, there is no steady solution for $\alpha = 0.1$. It would seem reasonable to expect a steady solution if α were between 0 and 0.1, yet we have not found this to be the case.

7. Discussion. In finding all of the preceding finite-tilt-angle solutions, we have found that the downstream film height tends to h_{up}, the upstream film height, and we have assumed that the value of h_{up} is such that the quotient Q is negative (see (15)). These two conditions are not violated by any of the preceding solutions, and so our solution process has been self-consistent.

Previous research (e.g., [10], [16]) has shown that when an external pressure acts on a thin film, in general the downstream height tends to a different value than the upstream height. The case we have considered, where $h_{dn} = h_{up}$, is more the exception than the rule, according to these sources. It is true that our problem is different than one involving a simplified external forcing; in our problem, the drop shape and speed are unknowns coupled with the film dynamics, while in [10], the external forcing is prescribed. Nevertheless, there exists the distinct possibility that we have found solutions that are exceptions to the general rule, and we must consider cases when $h_{dn} \neq h_{up}$.

It seems possible that a steady solution might exist for which $2h_{dn} < h(x_l) < 2h_{up}$. In this case, we see from (15) that Q is negative upstream and positive downstream. The arguments presented in section 4.2 suggest that there is now only one growth mode both upstream and downstream, and that we need only impose reduced far-field conditions, $\lim_{x\to-\infty} h'(x) = 0$ and $\lim_{x\to\infty} h'(x) = 0$, numerically at the ends of the finite computational domain. Hence, it may now be possible to find solutions with $h_{dn} \neq h_{up}$. Although neither h_{up} nor h_{dn}, the value of which is given by (14), is specified directly through boundary conditions, the values are felt through the flux and speed quantities in the differential equations.

In this situation, following the discussion of our numerical method, we now may guess at six scalars to satisfy seven equations, and so solutions might presumably be found as a one-parameter family. For example, β could now be chosen arbitrarily, while λ alone would be a dependent function of the other parameters. Such a result was found in the linearized case by Kriegsmann [9]. If h_{up} is instead chosen as the auxiliary dependent variable, it seems possible that steady solutions could be found for relatively arbitrary choices of the other physical parameters. The time dependent solutions in [10] show, for a simpler problem, the possibility of a film's evolving towards a steady state with adjusted heights upstream and downstream. These heights are determined as part of the steady solution.

In summary, we note that here we have found steady solutions only for certain select parameter sets. The effects of surface tension, spreading coefficient, and viscosity on the solutions have been studied. We speculate that, in the cases for which the parameter sets are chosen differently, there may be steady solutions for which $h_{dn} \neq h_{up}$.

REFERENCES

[1] D. J. Acheson, *Elementary Fluid Dynamics*, Oxford University Press, London, 1990.

[2] F. Brochard-Wyart, G. Debregeas, and P. G. de Gennes, *Spreading of viscous droplets on a nonviscous liquid*, Colloid Polymer Sci., 274 (1996), pp. 70–72.

[3] N. D. DiPietro, C. Huh, and R. G. Cox, *The hydrodynamics of the spreading of one liquid on the surface of another*, J. Fluid Mech., 84 (1978), pp. 529–549.

[4] N. D. DiPietro and R. G. Cox, *The containment of an oil slick by a boom placed across a uniform stream*, J. Fluid Mech., 96 (1980), pp. 613–640.

[5] M. Foda and R. G. Cox, *The spreading of thin liquid films on a water-air interface*, J. Fluid Mech., 101 (1980), pp. 33–51.

[6] J. G. E. M. Fraaije and A. M. Cazabat, *Dynamics of spreading on a liquid substrate*, J. Colloid Interface Sci., 133 (1989), pp. 452–460.

[7] W. D. Harkins, *Physical Chemistry of Surface Films*, Reinhold, New York, 1952.

[8] J.-F. Joanny, *Wetting of a liquid substrate*, PhysicoChemical Hydrodynamics, 9 (1987), pp. 183–196.

[9] J. J. Kriegsmann, *Spreading on a liquid interface*, Ph.D. thesis, Northwestern University, Evanston, IL, 1999.

[10] J. J. Kriegsmann, M. J. Miksis, and J.-M. Vanden-Broeck, *Pressure driven disturbances on a thin viscous film*, Phys. Fluids, 10 (1998), pp. 1249–1255.

[11] M. J. Miksis and J.-M. Vanden-Broeck, *Motion of a triple junction*, J. Fluid Mech., 437 (2001), pp. 385–394.

[12] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Modern Phys., 69 (1997), pp. 931–980.

[13] P. R. Pujado and L. E. Scriven, *Sessile lenticular configurations: Translationally and rotationally symmetric lenses*, J. Colloid Interface Sci., 40 (1972), pp. 82–98.

[14] J. S. Rowlinson and B. Widom, *Molecular Theory of Capillarity*, Clarendon Press, Oxford, UK, 1989.

[15] E. O. Tuck and L. W. Schwartz, *A numerical and asymptotic study of some third-order ordinary differential equations relevant to draining and coating flows*, SIAM Rev., 32 (1990), pp. 453–469.

[16] E. O. Tuck and J.-M. Vanden-Broeck, *Influence of surface tension on jet-stripped continuous coating of sheet materials*, Amer. Inst. Chem. Eng. (AIChE) J., 30 (1984), pp. 808–811.
[17] S. D. R. Wilson and J. Williams, *The flow of a liquid film on the inside of a rotating cylinder, and some related problems*, Phys. Fluids, 9 (1997), pp. 2184–2190.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 41–80

STRONGLY DEGENERATE PARABOLIC-HYPERBOLIC SYSTEMS MODELING POLYDISPERSE SEDIMENTATION WITH COMPRESSION*

STEFAN BERRES†, RAIMUND BÜRGER†, KENNETH H. KARLSEN‡, AND ELMER M. TORY§

Abstract. We show how existing models for the sedimentation of monodisperse flocculated suspensions and of polydisperse suspensions of rigid spheres differing in size can be combined to yield a new theory of the sedimentation processes of polydisperse suspensions forming compressible sediments ("sedimentation with compression" or "sedimentation-consolidation process"). For N solid particle species, this theory reduces in one space dimension to an $N \times N$ coupled system of quasi-linear degenerate convection-diffusion equations. Analyses of the characteristic polynomials of the Jacobian of the convective flux vector and of the diffusion matrix show that this system is of strongly degenerate parabolic-hyperbolic type for arbitrary N and particle size distributions. Bounds for the eigenvalues of both matrices are derived. The mathematical model for $N = 3$ is illustrated by a numerical simulation obtained by the Kurganov–Tadmor central difference scheme for convection-diffusion problems. The numerical scheme exploits the derived bounds on the eigenvalues to keep the numerical diffusion to a minimum.

Key words. polydisperse suspensions, sedimentation, systems of conservation laws, strongly degenerate parabolic-hyperbolic systems, central difference approximation

AMS subject classifications. 35K65, 35L40, 35L65, 65M06, 76T05

DOI. 10.1137/S0036139902408163

1. Introduction.

Mathematical models for the (controlled) sedimentation of polydisperse suspensions of small particles, which belong to a finite number of species differing in size or density and are suspended in a viscous fluid, are important to many applications such as the chemical engineering, ceramic, pulp and paper, and food industries, mineral processing, wastewater treatment, and medicine [3, 50, 88, 89, 101, 122]. The characteristic behavior of such mixtures is differential sedimentation, which leads to areas of different composition if an initially homogeneous suspension is allowed to settle. In this paper, we consider the additional property that the solid particles possibly form a compressible sediment layer. A mathematical model for polydisperse suspensions forming compressible sediments is developed, analyzed, and simulated, focusing on three different aspects.

First, we show how two existing sedimentation models—one for monodisperse flocculated suspensions, which are described by *scalar* strongly degenerate parabolic-hyperbolic equations, and one for polydisperse suspensions of rigid spheres differing

*Received by the editors May 20, 2002; accepted for publication (in revised form) March 24, 2003; published electronically October 14, 2003. This work was supported (in part) by the Sonderforschungsbereich 404 at the University of Stuttgart, by the BeMatA program of the Research Council of Norway, and by the European network HYKE, funded by the EC as contract HPRN-CT-2002-00282.

http://www.siam.org/journals/siap/64-1/40816.html

†Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany (berres@mathematik.uni-stuttgart.de, buerger@mathematik.uni-stuttgart.de).

‡Department of Mathematics, University of Bergen, Johs. Brunsgt. 12, N-5008 Bergen, Norway, and Centre of Mathematics for Applications, Department of Mathematics, University of Oslo, P.O. Box 1053, Blindern, N-0316 Oslo, Norway (kennethk@mi.uib.no).

§Department of Mathematics and Computer Science (Emeritus), Mount Allison University, Sackville, NB, E4L 1E8 Canada (sherpa@nbnet.nb.ca).

in size, which lead to *first-order systems* of conservation laws—can be combined into a model of sedimentation of polydisperse suspensions of particles (or flocs) forming compressible sediments.

Secondly, we prove that this model gives rise to strongly degenerate parabolic-hyperbolic systems of PDEs. (A precise definition of that type property is given below.) This type characterization is valid for arbitrary numbers N of sizes of equal-density particles. The application considered thus provides provably strongly degenerate parabolic-hyperbolic systems of arbitrary size. The well-posedness analysis and design of numerical schemes for such equations has received considerable interest in recent years, but, especially in the system (nonscalar) case, only a few applications are known. The present paper provides such an application.

Finally, for $N = 3$ we illustrate the model by numerical examples using the high-resolution Kurganov–Tadmor central difference scheme [64]. Its exposition in [64] is biased towards systems of conservation laws but also suggests an extension to parabolic-hyperbolic systems. This paper presents the first (to our knowledge) application of that extension to a realistic model.

In what follows, we outline the paper and put it in perspective relative to the existing literature. In section 2, we derive a set of spatially multidimensional model equations for the sedimentation of polydisperse suspensions forming compressible sediments (also called a sedimentation-consolidation process). The modeling starts from the usual mass and linear momentum balance equations for the N solids species (each regarded as one phase) and the fluid. The generic material properties of the suspension are introduced by constitutive assumptions concerning the solid and fluid stress tensors and the solid-fluid interaction forces. In particular, the solid phase pressures and the fluid pressure are replaced by the effective solid stress σ_e and the pore pressure. Here we assume that σ_e is a function of the total solids concentration $\phi := \phi_1 + \cdots + \phi_N$ only, where ϕ_i is the concentration of species i having diameter d_i and density ϱ_i. The way in which σ_e depends on ϕ_1 to ϕ_N determines the resulting diffusion matrix of the above-mentioned degenerate system. Specifying the solid-fluid interaction force for each species and finally performing a dimensional analysis, which permits our neglecting several terms of the linear momentum balance equations, we obtain explicit expressions for the solid-fluid relative velocity (or slip velocity) of each species as a function of $\Phi := (\phi_1, \ldots, \phi_N)^{\mathrm{T}}$ and $\nabla\Phi$, which in turn yield the fluxes of the continuity equations. The final (spatially multidimensional) model equations form a strongly degenerate system of N convection-diffusion equations for $\phi_1, \ldots, \phi_N$ coupled to the divergence-free condition of the volume-average mixture velocity and a three-component equation for the motion of the mixture. These last two equations account for viscous effects and reduce for $\Phi \equiv 0$ to the Stokes system for an incompressible fluid. Finally, we check that for $N = 1$ the strongly degenerate system reduces to the known scalar equation for monodisperse suspensions [26]. An overview of the analysis, numerics, and applications of strongly degenerate parabolic equations is given in section 5. For incompressible sediments, i.e., when $\sigma_e \equiv 0$, the model reduces to the Masliyah–Lockett–Bassoon (MLB) model [71, 73] for polydisperse suspensions of rigid spheres.

The effect of compressible sediment in polydisperse sedimentation has been studied only infrequently [98, 102]. Unfortunately, these treatments are incomplete in that they are not embedded in the appropriate mathematical PDE framework or are limited to $N = 2$. We assume that the mixture forms a compressible sediment layer whenever the cumulative solids concentration ϕ exceeds a critical value (or "gel

point") ϕ_c, and is in hindered settling for $\phi \le \phi_c$. In [102], however, the transition between the hindered settling zone (where $\phi \le \phi_c$) and the compression region (where $\phi > \phi_c$) is introduced by an artificial moving boundary condition, which we avoid by the concept of a degenerate diffusion equation. Our model also describes the relative movement of the solids species against each other within the sediment under the influence of the effective solid stress. This is unlike any previous treatment. Thus, our model is new and therefore derived completely in section 2. On the other hand, we essentially combine arguments that have been discussed extensively in previous works that focus on modeling either flocculated monodisperse [26, 27, 35] or nonflocculated polydisperse suspensions [14, 19, 22]. Thus the presentation in section 2 is fairly concise, and we refer to the cited papers for additional details and justification.

To continue the discussion, we need a precise definition of strongly degenerate parabolic-hyperbolic systems. In fact, in recent years we have seen an increased interest in quasi-linear systems of PDEs that in one space dimension can be written as

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \boldsymbol{\varphi}(\mathbf{u})}{\partial x} = \frac{\partial}{\partial x}\left(\mathbf{D}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x}\right), \quad x \in \mathbb{R},\ t > 0, \tag{1.1}$$

where $\mathbf{u} : \mathbb{R} \times \mathbb{R}^+ \to \mathcal{D} \subset \mathbb{R}^N$ is the sought solution vector, $\boldsymbol{\varphi} : \mathcal{D} \to \mathbb{R}^N$ is a flux vector, and $\mathbf{D} : \mathcal{D} \to \mathbb{R}^{N\times N}$ is a diffusion matrix. We allow the system to be degenerate in the sense that $\mathbf{D}(\mathbf{u}) = 0$ for $\mathbf{u} \in \mathcal{D}' \subset \mathcal{D}$; i.e., the system reduces to first order on $\mathcal{D}'$. The system is called *strongly degenerate* if $\mathcal{D}'$ is of nonzero N-dimensional measure. Moreover, we recall that the system (1.1) is *strictly parabolic* at a point $\mathbf{u}_0 \in \mathcal{D}$ if $\mathbf{D}(\mathbf{u}_0) > 0$; i.e., the matrix $\mathbf{D}(\mathbf{u}_0)$ has only positive eigenvalues. On the other hand, if $\mathbf{u}_0$ is chosen such that $\mathbf{D}(\mathbf{u}_0) = 0$, then, according to the usual terminology for conservation laws, the system (1.1) is called *hyperbolic* if the Jacobian $\boldsymbol{\mathcal{J}}_{\boldsymbol{\varphi}}(\mathbf{u}_0)$ has N real eigenvalues, and *strictly hyperbolic* if these eigenvalues are moreover pairwise distinct. Finally, we shall call (1.1) a *strongly degenerate parabolic-hyperbolic system* if, at any point $\mathbf{u}_0$ belonging to the interior $\mathcal{D}^0$ of $\mathcal{D}$, the system (1.1) is either strictly parabolic or strictly hyperbolic in the sense given above and the set $\mathcal{D}^0 \cap \mathcal{D}'$ on which the system is strictly hyperbolic is of nonzero N-dimensional measure. We emphasize here that points $\mathbf{u} \in \mathcal{D}\backslash\mathcal{D}^0$, which are on the boundary of the physically relevant region $\mathcal{D}$, do not enter the type characterization [59]. Note that a strictly hyperbolic first-order system of conservation laws, for which the right-hand side of (1.1) vanishes identically, is included as a special case. Of course, solutions of (1.1) are in general discontinuous, even for smooth initial data. Further properties are discussed in section 5.

If the multidimensional sedimentation equations developed in section 2 are restricted to one space dimension, the motion of the mixture is determined by the velocity at one end of the computational domain. For a closed vessel, this velocity is zero, and only the degenerate system for $\phi_1, \ldots, \phi_N$ needs to be solved. This system of second-order PDEs can then be written as

$$\frac{\partial \Phi}{\partial t} + \frac{\partial \mathbf{f}(\Phi)}{\partial z} = \frac{\partial}{\partial z}\left(\mathbf{A}(\Phi)\frac{\partial \Phi}{\partial z}\right), \tag{1.2}$$

where t is time and z is height. In section 2, we assume that $\sigma_e = 0$ for $\phi \le \phi_c$. This assumption implies $\mathbf{A}(\Phi) = 0$ for $\phi \le \phi_c$. In this case, the system (1.2) is reduced to the first-order system

$$\frac{\partial \Phi}{\partial t} + \frac{\partial \mathbf{f}(\Phi)}{\partial z} = 0. \tag{1.3}$$

We consider the system (1.2) for vectors $\Phi \in \mathcal{D}_{\phi_{\max}}$, where $0 < \phi_{\max} \leq 1$ denotes the maximum admissible cumulative solids concentration, and we define for $0 < \phi_{\mathrm{M}} \leq 1$

$$\mathcal{D}_{\phi_{\mathrm{M}}} := \big\{\Phi = (\phi_1, \ldots, \phi_N) \in \mathbb{R}^N : \phi_1 \geq 0, \ldots, \phi_N \geq 0;\ \phi_1 + \cdots + \phi_N \leq \phi_{\mathrm{M}}\big\}.$$

Moreover, we denote by $\mathcal{D}^0_{\phi_{\mathrm{M}}}$ the interior of $\mathcal{D}_{\phi_{\mathrm{M}}}$, that is,

$$\mathcal{D}^0_{\phi_{\mathrm{M}}} := \big\{\Phi = (\phi_1, \ldots, \phi_N) \in \mathbb{R}^N : \phi_1 > 0, \ldots, \phi_N > 0;\ \phi_1 + \cdots + \phi_N < \phi_{\mathrm{M}}\big\}.$$

Obviously, the type of (1.2) is determined by the properties of $\mathbf{A}(\Phi)$ for $\phi_{\mathrm{c}} < \phi < \phi_{\max}$ and by those of the vector $\mathbf{f}(\Phi)$ (more precisely, of its Jacobian $\boldsymbol{\mathcal{J}}_{\mathbf{f}}(\Phi)$) for $0 < \phi \leq \phi_{\mathrm{c}}$ and $\phi = \phi_{\max}$. Since every component of $\mathbf{f}(\Phi) = (f_1(\Phi), \ldots, f_N(\Phi))^{\mathrm{T}}$ depends nonlinearly on every component of Φ, and since $\boldsymbol{\mathcal{J}}_{\mathbf{f}}(\Phi)$ is unsymmetric, it is by no means obvious that the system (1.3) is strictly hyperbolic. Since all entries of $\mathbf{A}(\Phi)$ are nonzero on $\mathcal{D}^0_{\phi_{\max}} \backslash \mathcal{D}_{\phi_{\mathrm{c}}}$, it is not apparent either that the system (1.2) is strictly parabolic for $\phi \in \mathcal{D}^0_{\phi_{\max}} \backslash \mathcal{D}_{\phi_{\mathrm{c}}}$. The core of this paper is formed by sections 3 and 4, where these properties are established by analyzing the characteristic polynomials of $\boldsymbol{\mathcal{J}}_{\mathbf{f}}(\Phi)$ and $\mathbf{A}(\Phi)$, respectively, where the vector $\mathbf{f}(\Phi)$ and the matrix $\mathbf{A}(\Phi)$ are chosen according to the model developed in section 2. Moreover, for the analysis of section 3, we assume that the particles all have the same density and that the species differ in size only. Our treatment has in part been inspired by Rosso and Sona's recent analysis of equations modeling the separation of oil-water dispersions [87]. In section 3, we discuss the properties of the system (1.3) with $\mathbf{f}(\Phi) = \mathbf{f}^{\mathrm{M}}(\Phi)$,

$$\frac{\partial \phi_i}{\partial t} + \frac{\partial f_i^{\mathrm{M}}(\Phi)}{\partial z} = 0, \quad i = 1, \ldots, N, \tag{1.4}$$

which arises from the model derived in section 2 by considering one space dimension and a closed settling vessel and assuming that the effective solid stress vanishes ($\sigma_{\mathrm{e}} \equiv 0$). The "M" indicates that the constitutive assumptions in section 2 have been chosen according to the MLB approach [22, 71, 73] (see [14, 22] for alternate equations for $\mathbf{f}(\Phi)$). The analysis of section 3 leads to the type of the system (1.2) for $\phi \leq \phi_{\mathrm{c}}$ and fully determines its type for $\sigma_{\mathrm{e}} \equiv 0$, that is, for a suspension of rigid particles [14, 19, 22]. The main result is that in the equal-density case, the system (1.4) is indeed strictly hyperbolic for all $\Phi \in \mathcal{D}^0_{\phi_{\max}}$ for $0 < \phi_{\max} \leq 1$. Strict hyperbolicity holds for all N and arbitrary particle sizes $d_1 > d_2 > \cdots > d_N > 0$.

To outline the significance of the analysis of section 3 in nontechnical terms, let us first say that hyperbolicity of a first-order system of conservation laws like (1.3) is in general a desirable property. In fact, the existence of a complete set of pairwise-distinct eigenvalues at each relevant point Φ of the state space ensures that the solution of (1.3) involves (simple) waves, i.e., solutions which essentially involve one eigenvalue of the Jacobian $\boldsymbol{\mathcal{J}}_{\mathbf{f}}(\Phi)$ and a corresponding eigenvector; see [51] for details. The important point is that each eigenvalue represents a finite propagation speed of solution information. For a mixture of flowing phases (in our case, the N "particulate" phases and the fluid), we should expect not only that a good model predicts finite speeds of propagation, but that moreover no solution information travels faster than any of the physical phases. We shall show later (Lemma 6.1 in section 6.3) that the MLB model for equal-density spheres and dilute to moderately concentrated suspensions indeed satisfies this requirement.

To put the hyperbolicity result in the proper perspective, let us now look at the opposite situation. Loss of hyperbolicity for a given vector $\Phi \in \mathbb{R}^N$ means that system

(1.4) has at least one pair of complex-conjugate eigenvalues. For $N = 2$, we then say that the system is *elliptic*. In most cases, for vectors Φ chosen from some subregion of the relevant state space, the system (1.4) is nonhyperbolic or elliptic and is hyperbolic elsewhere. Such systems are called *mixed systems*; see [47] for a survey of applications. In some applications, such as multiphase flow in porous media, the significance of mixed systems is essentially unclear, and loss of hyperbolicity is sometimes related to a model error. For polydisperse sedimentation, however, it is shown in [22] that for arbitrary N the degeneracy into nonhyperbolic type is a criterion for the possible occurrence of horizontal structures like fingers, columns, or blobs during sedimentation. This interpretation of nonhyperbolicity generalizes a criterion formulated in [4] for $N = 2$. Such instabilities have been observed in experiments [4, 117] at certain initial concentrations and are particularly likely to occur in suspensions including one species that is heavier and one that is lighter than the fluid. On the other hand, instabilities have never been observed with equal-density particles.

For a given polydisperse sedimentation model, expressed by the specific algebraic form of the flux vector $\mathbf{f}(\Phi)$, the ellipticity region (which usually has to be determined numerically [22]) for given particle densities and sizes should agree with those concentration regions for which instabilities have been observed experimentally. On the other hand, the model equations should be strictly hyperbolic for arbitrary N and equal-density particles. In [22] we show that the MLB model satisfies the first of these properties and is, in particular, not hyperbolic in general for suspensions in which two or more species have different densities. However, in [22] we were able to prove strict hyperbolicity for the system (1.4) with equal-density particles in the case $N = 2$ only. We indicated in [22] that numerical tests with $N = 3$ never produced an instability region, and we conjectured that the MLB equations were hyperbolic for arbitrary N, which is now proved in the present paper.

The properties of the MLB model contrast with those of several other models. For example, the model proposed by Davis and Gecol [38] again leads to a system of the form (1.4), but which for equal-density particles is hyperbolic only for small values of d_1/d_2 (for example, for $N = 2$ the restriction is $d_1/d_2 < 5$; see [22]), and for which the size of the ellipticity region drastically increases when d_1/d_2 is increased. Since no instabilities have been observed experimentally with equal-density suspensions, these ellipticity regions are unphysical and limit the use of the Davis and Gecol model to small values of d_1/d_2. We refer to [22] for a thorough discussion of mixed systems modeling polydisperse sedimentation and the consequences for the mathematical analysis.

In section 4, we consider the right-hand side of (1.2) using the diffusion matrix $\mathbf{A}(\Phi)$ derived in section 2. While it is obvious that $\mathbf{A}(\Phi) = 0$ on $\mathcal{D}_{\phi_c}$, it is not apparent that $\mathbf{A}(\Phi)$ is positive definite on $\mathcal{D}^0_{\phi_{\max}} \backslash \mathcal{D}_{\phi_c}$. The hyperbolicity and parabolicity properties of (1.2) associated with the matrices $\boldsymbol{\mathcal{J}}_{\mathbf{f}^M}(\Phi)$ and $\mathbf{A}(\Phi)$ are controlled by the independent model functions $V(\phi)$ and $\sigma_e(\phi)$, but their entries are analogous. Thus, the formula for the characteristic polynomial of $\boldsymbol{\mathcal{J}}_{\mathbf{f}^M}(\Phi)$ derived in section 3 also provides (after substitutions) a formula for that of $\mathbf{A}(\Phi)$. It is then straightforward to prove that $\mathbf{A}(\Phi)$ has N distinct nonnegative eigenvalues, which are positive if $\phi_c < \phi < \phi_{\max}$. Thus, (1.2) is strictly parabolic for $\phi_c < \phi < \phi_{\max}$, which is the main result of section 4.

In contrast to the hyperbolicity of the first-order system (1.4), the parabolicity property established in section 4 does not admit a direct physical interpretation. Rather, parabolicity is a condition ensuring the well-posedness (existence, uniqueness,

and stability) of systems of PDEs of the form (1.2). It should, however, be pointed out that mathematically rigorous well-posedness results are available for certain special cases of (1.2) only. These include, on one hand, *scalar* strongly degenerate parabolic-hyperbolic equations [6, 7, 8, 17, 25, 31, 61, 62, 72, 118], and, on the other hand, certain *uniformly parabolic* systems, that is, systems that do not degenerate into first-order type [42, 49, 63, 66]. A closed mathematical theory for the strongly degenerate systems considered in this paper is not available despite the increased interest this kind of equation has attained in recent years. Section 5 provides a short overview of the existing literature on mathematical and numerical theory for strongly degenerate parabolic problems.

In section 6, we first describe the central difference scheme due to Kurganov and Tadmor [64], which is used in this paper. The system (1.2) is discretized by a high-resolution central difference (Riemann solver free) scheme for the convection part (corresponding to the first-order equation (1.4)) combined with a central difference discretization of the parabolic parts (the right-hand side of (1.2)). Then we illustrate the (new) model of polydisperse sedimentation with compression by numerical examples with $N = 3$ and compare the results with simulations of the two (conventional) models of settling of monodisperse flocculated and polydisperse rigid-sphere suspensions (where $\sigma_{\mathrm{e}} \equiv 0$). We refer to [14, 19] for the application of similar numerical schemes to first-order systems like (1.4) describing sedimentation of polydisperse suspensions without compression effects.

The closing section 7 discusses various aspects of the paper. We first show that the type analysis of sections 3 and 4 is also valid in several space dimensions. Next, we briefly comment on the possible extension of the model to polydisperse suspensions with particles of different densities, and we furthermore provide a physical interpretation of one of the eigenvalue bounds derived in section 3. One frequent topic in the sedimentation literature is hydrodynamic diffusion, which is associated with particle velocity fluctuations. We give a brief survey of the literature on hydrodynamic diffusion and provide justification for not including this effect in our model. An important new property of the model is the prediction of diffusive relative movement of the different solids species within the sediment. This effect is clearly visible in the numerical simulations, which correspond to a hypothetical material, and may be less pronounced for real materials. We therefore discuss several alternative gradual and structural modifications of the present model that could reduce sediment diffusivity. Finally, some applications in which the sediment compressibility is important are discussed.

2. Derivation of the model of polydisperse sedimentation with compression.

2.1. Mass and linear momentum balance equations. A suspension may be represented as a superposition of continuous media, each following its own movement with the only restrictions imposed by the interaction between components. Each component obeys the laws of conservation of mass and momentum, incorporating terms to account for the interchange between components [27]. We assume that there is no mass transfer between species.

The local mass balance equations of the solid species and of the fluid can be written as

$$\frac{\partial \phi_i}{\partial t} + \nabla \cdot \big(\phi_i \mathbf{v}_i\big) = 0, \quad i = 1, \ldots, N, \qquad -\frac{\partial \phi}{\partial t} + \nabla \cdot \big((1-\phi)\mathbf{v}_{\mathrm{f}}\big) = 0, \tag{2.1}$$

where $\mathbf{v}_i$ is the phase velocity of solids species i, $i = 1, \dots, N$, and $\mathbf{v}_\mathrm{f}$ is the fluid phase velocity. Defining the volume-average velocity of the mixture $\mathbf{q} := (1-\phi)\mathbf{v}_\mathrm{f} + \phi_1\mathbf{v}_1 + \cdots + \phi_N\mathbf{v}_N$ and the relative velocities or slip velocities $\mathbf{u}_i := \mathbf{v}_i - \mathbf{v}_\mathrm{f}$ for $i = 1, \dots, N$, we derive easily that

$$\phi_i\mathbf{v}_i = \phi_i\big(\mathbf{u}_i + \mathbf{q} - (\phi_1\mathbf{u}_1 + \cdots + \phi_N\mathbf{u}_N)\big), \quad i = 1, \dots, N; \tag{2.2}$$

hence the solids mass balance equations can be rewritten in terms of $\mathbf{q}$ and $\mathbf{u}_1, \dots, \mathbf{u}_N$ as

$$\frac{\partial\phi_i}{\partial t} + \nabla\cdot\big(\phi_i\mathbf{u}_i + \phi_i\mathbf{q} - \phi_i(\phi_1\mathbf{u}_1 + \cdots + \phi_N\mathbf{u}_N)\big) = 0, \quad i = 1, \dots, N. \tag{2.3}$$

The sum of all equations in (2.1) produces the simple mass balance of the mixture, $\nabla\cdot\mathbf{q} = 0$. The momentum balance equations for the N solid species and the fluid are

$$\varrho_i\phi_i\frac{\mathrm{D}\mathbf{v}_i}{\mathrm{D}t} = \nabla\cdot\mathbf{T}_i + \varrho_i\phi_i\mathbf{b} + \mathbf{m}_i^\mathrm{f} + \mathbf{m}_i^\mathrm{s}, \quad i = 1, \dots, N, \tag{2.4}$$

$$\varrho_\mathrm{f}(1-\phi)\frac{\mathrm{D}\mathbf{v}_\mathrm{f}}{\mathrm{D}t} = \nabla\cdot\mathbf{T}_\mathrm{f} + \varrho_\mathrm{f}(1-\phi)\mathbf{b} - (\mathbf{m}_1^\mathrm{f} + \cdots + \mathbf{m}_N^\mathrm{f}). \tag{2.5}$$

Here ϱ_f is the mass density of the fluid, $\mathbf{T}_i$ denotes the stress tensor of particle species i, $i = 1, \dots, N$, $\mathbf{T}_\mathrm{f}$ that of the fluid, $\mathbf{b}$ is the body force, $\mathbf{m}_i^\mathrm{f}$ and $\mathbf{m}_{ij}^\mathrm{s}$ are the interaction forces per unit volume between solid species i and the fluid and between the solid species i and j, respectively, $\mathbf{m}_i^\mathrm{s} := \mathbf{m}_{i1}^\mathrm{s} + \cdots + \mathbf{m}_{iN}^\mathrm{s}$ is the particle-particle interaction term of species i, and we use the standard notation $\mathrm{D}\mathbf{v}/\mathrm{D}t := \partial\mathbf{v}/\partial t + (\mathbf{v}\cdot\nabla)\mathbf{v}$.

2.2. Solid and fluid stress tensors. We assume that the stress tensors of the solid and fluid phases can be written as $\mathbf{T}_i = -p_i\mathbf{I} + \mathbf{T}_i^\mathrm{E}$ for $i = 1, \dots, N$ and $\mathbf{T}_\mathrm{f} = -p_\mathrm{f}\mathbf{I} + \mathbf{T}_\mathrm{f}^\mathrm{E}$, respectively, where p_i denotes the phase pressure of particle species i, p_f that of the fluid, $\mathbf{I}$ denotes the identity tensor, and $\mathbf{T}_i^\mathrm{E}$ and $\mathbf{T}_\mathrm{f}^\mathrm{E}$ are the corresponding extra (or viscous) stress tensors, all of which could be given by expressions that correspond, for example, to a viscous-linear fluid. Since the focus here is on the continuity equations for the solids and we assume that viscous effects due to the motion of the mixture are not dominant, all viscous effects are assigned to the fluid extra-stress tensor. To make this simplification visible in the dimensional analysis, we assume that ν_0^f and $\nu_0^\mathrm{s} < \nu_0^\mathrm{f}$ are characteristic viscosities associated with the fluid and the solid species, respectively.

2.3. Partial pressures, pore pressure, and effective solid stress. The phase pressures $p_1, \dots, p_N$ and p_f are theoretical variables (arising from the averaging procedure [27]), which cannot be measured experimentally. As in [26], they are replaced by the pore pressure p and the effective solid stress σ_e, which are measurable. We assume that σ_e is given by a constitutive equation $\sigma_\mathrm{e} = \sigma_\mathrm{e}(\Phi)$, that is, as a function of the local composition of the sediment. To our knowledge (see also [102]), no suitable function $\sigma_\mathrm{e} = \sigma_\mathrm{e}(\Phi)$ for the polydisperse case has been derived either theoretically or empirically so far. However, most researchers utilize formulas that relate σ_e to the sediment porosity or, equivalently, to the total volumetric solids concentration ϕ [74, 98].

In stating the generic assumptions on σ_e, we follow [81, 92] and consider that during sedimentation, when $\phi \le \phi_\mathrm{c}$, there is no permanent contact between the particles (or aggregates of them), and the momentum transfer between the particles occurs entirely through the fluid or through collisions (although in a moment we shall

show that the latter effect is negligible here). This means that the total stress of the mixture, p_t, which can be decomposed in two different ways as

$$p_\mathrm{t} = p_\mathrm{f} + p_1 + \cdots + p_N = p + \sigma_\mathrm{e}(\phi), \tag{2.6}$$

equals the pore pressure, and therefore $\sigma_\mathrm{e}(\phi) = 0$ for $\phi \le \phi_\mathrm{c}$. (The second equality in (2.6) reflects the well-known effective-stress principle [39].) During consolidation, when $\phi > \phi_\mathrm{c}$, permanent contact is established between the solid particles, and the contact forces are transmitted through solid-solid contacts. Moreover, it can be assumed that the part of the total stress supported by the skeleton of networked solid particles is an increasing function of their concentration ϕ, i.e., $\sigma_\mathrm{e}'(\phi) := d\sigma_\mathrm{e}(\phi)/d\phi > 0$ for $\phi > \phi_\mathrm{c}$. These generic assumptions on $\sigma_\mathrm{e}(\phi)$ can be summarized as

$$\sigma_\mathrm{e}(\phi) \begin{cases} = 0 & \text{for } \phi \le \phi_\mathrm{c}, \\ > 0 & \text{for } \phi > \phi_\mathrm{c}, \end{cases} \qquad \sigma_\mathrm{e}'(\phi) \begin{cases} = 0 & \text{for } \phi \le \phi_\mathrm{c}, \\ > 0 & \text{for } \phi > \phi_\mathrm{c}; \end{cases} \tag{2.7}$$

a specific example is given in section 6. Our concept of effective solid stress has been adopted from soil consolidation theory [81, 92] but is consistent with and in some cases mathematically equivalent to the concepts of compressive yield stress [52, 67], effective pressure [40], or yield pressure [54] utilized by research workers with a focus on solid-liquid separation. All these papers have in common that it is assumed that the effective stress takes positive values if and only if the particles are networked, and that this occurs when $\phi > \phi_\mathrm{c}$, where ϕ_c is a distinct critical concentration, also called the "threshold value" or "gel point."

We now relate the fluid and solid phase pressures p_f and $p_1, \ldots, p_N$ to the effective solid stress σ_e and the pore pressure p. While p is defined within the fluid filling the interstices between the solids, the partial fluid pressure p_f is defined in the fluid component occupying the whole volume of the mixture. Let S be the cross-section of a settling column and $S_\mathrm{f} \subset S$ be its part that is filled out by the fluid in the porous medium, and let ϵ denote the surface porosity $\epsilon := |S_\mathrm{f}|/|S|$, i.e., $dS_\mathrm{f} = \epsilon dS$. Then the surface forces exerted on the fluid in a cross section of the sediment are

$$\int_S p_\mathrm{f}\, dS = \int_{S_\mathrm{f}} p\, dS_\mathrm{f} = \int_S p(\epsilon dS). \tag{2.8}$$

Since we may assume that the surface porosity equals the volume porosity [22], we may replace ϵ by $1 - \phi$, and as a consequence of the localization theorem [53], we obtain $p_\mathrm{f} = (1 - \phi)p$ from (2.8).

The effective solid stress σ_e is that part of the total stress p_t which acts on the porous network formed by the solid particles. Assuming that the cross-sectional surface area fraction of each solids species equals its volume fraction [22], we may conclude that $(\phi_i/\phi)\sigma_\mathrm{e}(\phi)$ is that part of σ_e which acts on species i. In view of $p_\mathrm{f} = (1 - \phi)p$, (2.6) may be rewritten as

$$p_1 + \cdots + p_N = \phi p + \frac{\phi_1 + \cdots + \phi_N}{\phi} \sigma_\mathrm{e}(\phi).$$

Thus, the phase pressure p_i is related to p and σ_e by $p_i = (\phi_i/\phi)(\phi p + \sigma_\mathrm{e}(\phi))$ for $i = 1, \ldots, N$.

2.4. Body force, solid-fluid, and particle-particle interaction forces. We assume that the only body force is gravity, $\mathbf{b} = -g\mathbf{k}$, where g is the acceleration of

gravity and $\mathbf{k}$ is the upwards-pointing unit vector. Furthermore, for a monodisperse suspension [26, 27, 34, 35], the interaction force $\mathbf{m}$ between the fluid and the unique solid phase can be modeled by

$$\mathbf{m} = \alpha(\phi)\mathbf{u} + \beta(\phi)\nabla\phi, \tag{2.9}$$

where α is the resistance coefficient and $\mathbf{u} := \mathbf{v}_\mathrm{s} - \mathbf{v}_\mathrm{f}$ is the solid-fluid relative or slip velocity. Equation (2.9) follows from the theorem of representation of isotropic functions [70, 99, 115, 115A, 115B] if we require that $\mathbf{m}$ be given as the most general linear function of $\mathbf{u}$, ϕ, and $\nabla\phi$. A similar result is obtained in [41], and (2.9) is also presented in [82] within a discussion of general principles for the formulation of constitutive equations. The function $\beta(\phi)$ can be shown to coincide with the pore pressure p (see [26]). In the present case, we analogously assume that the solid-fluid interaction term related to species i is given by $\mathbf{m}_i^\mathrm{f} = \alpha_i(\Phi)\mathbf{u}_i + \beta_i(\Phi)\nabla\phi_i$, where α_i is the resistance coefficient for the transfer of momentum between the fluid and solid phase species i, $i = 1, \dots, N$.

The interaction force between the different solid particle species could be specified by the Nakamura and Capes formula [1, 76, 98]:

$$\mathbf{m}_{ij}^\mathrm{s} = \frac{3}{2}\varphi_\mathrm{e}\frac{\varrho_i\varrho_j\phi_i\phi_j(d_i+d_j)^2}{\varrho_i d_i^3 + \varrho_j d_j^3}\,\|\mathbf{v}_i - \mathbf{v}_j\|\,(\mathbf{v}_i - \mathbf{v}_j), \quad i,j = 1,\dots,N, \quad i \neq j,$$

where the parameter φ_e accounts for non–head-on collisions [98] and its value depends on whether these are plastic or elastic. Typical values of φ_e vary between 0 and 5 [1, 76], and numerical simulations have not turned out to be sensitive to φ_e (see [1]). Nevertheless, the elimination of the term $\mathbf{m}_i^\mathrm{s} = \mathbf{m}_{i1}^\mathrm{s} + \cdots + \mathbf{m}_{iN}^\mathrm{s}$ due to the dimensional analysis (see section 2.5) is not dependent on any particular formula, since there is considerable experimental and theoretical evidence (summarized in [22]) that $\mathbf{m}_{ij}^\mathrm{s}$ can be neglected at the very low Reynolds numbers considered here.

To determine $\beta_1(\Phi), \dots, \beta_N(\Phi)$, we insert the constitutive assumptions into (2.4) and (2.5) and consider the mixture at equilibrium ($t \to \infty$) in a settling column. This state is characterized by $\mathbf{v}_\mathrm{f} = 0$, $\mathbf{u}_1 = \cdots = \mathbf{u}_N = 0$, and $\nabla p = -\varrho_\mathrm{f} g\mathbf{k}$, and we obtain $\beta_1(\Phi) = \cdots = \beta_N(\Phi) = p$; i.e., the functions β_i are all constant with respect to Φ [22, 26]. The linear momentum balances now read

$$\varrho_i\phi_i\frac{\mathrm{D}\mathbf{v}_i}{\mathrm{D}t} = -\varrho_i\phi_i g\mathbf{k} + \nabla\cdot\mathbf{T}_i^\mathrm{E} - \phi_i\nabla p + \alpha_i(\Phi)\mathbf{u}_i + \mathbf{m}_i^\mathrm{s} - \nabla\left(\frac{\phi_i}{\phi}\sigma_\mathrm{e}(\phi)\right), \quad i = 1,\dots,N, \tag{2.10}$$

$$\nabla p = -\varrho_\mathrm{f} g\mathbf{k} - \frac{1}{1-\phi}\big(\alpha_1(\Phi)\mathbf{u}_1 + \cdots + \alpha_N(\Phi)\mathbf{u}_N\big) - \varrho_\mathrm{f}\frac{\mathrm{D}\mathbf{v}_\mathrm{f}}{\mathrm{D}t} + \frac{1}{1-\phi}\nabla\cdot\mathbf{T}_\mathrm{f}^\mathrm{E}. \tag{2.11}$$

2.5. Dimensional analysis. We introduce dimensionless (starred) variables by referring all densities to ϱ_f, all velocities to the velocity U, all lengths to a typical length L, all solid and fluid viscosities to ν_0^s and ν_0^f, respectively, and all pressures to the hydrostatic pressure $\varrho_\mathrm{f} gL$. Here, we assume that U is the settling velocity of a single particle of the fastest settling species in an unbounded medium, and L is the depth of the settling vessel. A characteristic time is then given by $T = L/U$. A dimensionless gradient of a variable u is defined by $\nabla^* u = L\nabla u$, and a dimensionless time derivative by $\partial u/\partial t^* = T\partial u/\partial t = (L/U)\partial u/\partial t$. Using the Froude number of the flow $\mathrm{Fr} := U^2/(gL)$ and the sedimentation Reynolds number $\mathrm{Re} := dU/\nu_0^\mathrm{f}$, where d

is the size of the largest particles, we obtain from (2.4) and (2.11) the dimensionless equations

$$\begin{aligned}\varrho_i^*\phi_i\mathrm{Fr}\frac{\mathrm{D}\mathbf{v}_i^*}{\mathrm{D}t^*} =& -\varrho_i^*\phi_i\mathbf{k}+\frac{d}{L}\frac{\nu_0^{\mathrm{s}}}{\nu_0^{\mathrm{f}}}\frac{\mathrm{Fr}}{\mathrm{Re}}\nabla^*\cdot\left(\mathbf{T}_i^{\mathrm{E}}\right)^*-\phi_i\nabla^*p^*+\alpha_i^*(\Phi)\mathbf{u}_i^*\\ &+\frac{L}{d}\mathrm{Fr}(\mathbf{m}_i^{\mathrm{s}})^*-\nabla^*\left(\frac{\phi_i}{\phi}\sigma_{\mathrm{e}}^*(\phi)\right),\quad i=1,\ldots,N,\end{aligned}\tag{2.12}$$

$$\nabla^*p^*=-\mathbf{k}-\frac{1}{1-\phi}\left(\alpha_1^*(\Phi)\mathbf{u}_1^*+\cdots+\alpha_N^*(\Phi)\mathbf{u}_N^*\right)-\mathrm{Fr}\frac{\mathrm{D}\mathbf{v}_{\mathrm{f}}^*}{\mathrm{D}t^*}+\frac{1}{1-\phi}\frac{d}{L}\frac{\mathrm{Fr}}{\mathrm{Re}}\nabla^*\cdot\left(\mathbf{T}_{\mathrm{f}}^{\mathrm{E}}\right)^*.\tag{2.13}$$

The values $d=10^{-4}\,\mathrm{m}$, $g=10\,\mathrm{m/s^2}$, $L=1\,\mathrm{m}$ (height of a settling vessel), $U=10^{-4}\,\mathrm{m/s}$ (settling velocity of a particle of the fastest species in an unbounded fluid), and $\nu_0^{\mathrm{f}}=10^{-6}\,\mathrm{m^2/s}$ (kinematic viscosity of water) are typical for the particulate systems considered here and imply $\mathrm{Fr}=10^{-9}$, $\mathrm{Re}=10^{-2}$, and $d/L=10^{-4}$. Since all viscous effects have been moved onto the fluid extra-stress tensor, we can assume $\nu_0^{\mathrm{s}}/\nu_0^{\mathrm{f}}\ll 1$. We assume that all dimensionless variables are of the order of magnitude $\mathcal{O}(1)$. Then we obtain, by discarding from (2.12) all terms that have a coefficient that is 10^{-5} or smaller, and discarding the advective acceleration term from (2.13) but retaining the viscous term, the following simplified linear momentum balances:

$$\alpha_i(\Phi)\mathbf{u}_i=\varrho_i\phi_i g\mathbf{k}+\phi_i\nabla p+\nabla\left(\frac{\phi_i}{\phi}\sigma_{\mathrm{e}}(\phi)\right),\quad i=1,\ldots,N,\tag{2.14}$$

$$\nabla p=-\varrho_{\mathrm{f}}g\mathbf{k}-\frac{1}{1-\phi}\left(\alpha_1(\Phi)\mathbf{u}_1+\cdots+\alpha_N(\Phi)\mathbf{u}_N\right)+\frac{1}{1-\phi}\nabla\cdot\mathbf{T}_{\mathrm{f}}^{\mathrm{E}},\tag{2.15}$$

which are written again in their dimensional forms. The small viscous term $\nabla\cdot\mathbf{T}_{\mathrm{f}}^{\mathrm{E}}$ is retained in (2.15) when this equation acts as an equation for the motion of the mixture. We shall comment on the necessity of viscous terms in the multidimensional case in section 2.7.

The term $\nabla\cdot\mathbf{T}_{\mathrm{f}}^{\mathrm{E}}$ is, however, deleted when (2.15) is inserted into (2.14), in order to produce a solvable linear system for the slip velocities $\mathbf{u}_1,\ldots,\mathbf{u}_N$. Thus, this system can be written as

$$\begin{aligned}&\frac{\alpha_i(\Phi)(1-\phi)}{\phi_i}\mathbf{u}_i+\sum_{j=1}^{N}\alpha_j(\Phi)\mathbf{u}_j\\ &=(1-\phi)\left[\left(\varrho_i-\varrho_{\mathrm{f}}\right)g\mathbf{k}+\frac{1}{\phi_i}\nabla\left(\frac{\phi_i}{\phi}\sigma_{\mathrm{e}}(\phi)\right)\right],\quad i=1,\ldots,N.\end{aligned}\tag{2.16}$$

2.6. Explicit formula for the slip velocities $\mathbf{u}_i$. Let $\varrho(\Phi):=(1-\phi)\varrho_{\mathrm{f}}+\phi_1\varrho_1+\cdots+\phi_N\varrho_N$ denote the local density of the mixture, and note that $\phi_1(\varrho_1-\varrho_{\mathrm{f}})+\cdots+\phi_N(\varrho_N-\varrho_{\mathrm{f}})=\varrho(\Phi)-\varrho_{\mathrm{f}}$. Then the following explicit equation for the slip velocities $\mathbf{u}_i$ as functions of Φ is obtained as the solution of the system (2.16), which follows from the Sherman–Morrison formula [22]:

$$\mathbf{u}_i=\frac{\phi_i}{\alpha_i(\Phi)}\left[\left(\varrho_i-\varrho(\Phi)\right)g\mathbf{k}+\frac{\sigma_{\mathrm{e}}(\phi)}{\phi_i}\nabla\left(\frac{\phi_i}{\phi}\right)+\frac{1-\phi}{\phi}\nabla\sigma_{\mathrm{e}}(\phi)\right],\quad i=1,\ldots,N.\tag{2.17}$$

Following [22] and being consistent with Masliyah [73] and Lockett and Bassoon [71], we choose $\phi_i/\alpha_i(\Phi)=-d_i^2V(\Phi)/(18\mu_{\mathrm{f}})$, where μ_{f} is the viscosity of the pure fluid, and

the hindered settling factor $V(\Phi)$ can, for example, be chosen as $V(\Phi) = (1-\phi)^{n(\Phi)-2}$ [86]. Since the dependence of n on Φ is through wall effects, which are small when d is very small compared to the diameter of the settling vessel, we may limit the analysis to formulas of the type $V(\Phi) = V(\phi)$ and obtain

$$\mathbf{u}_i = -\frac{d_i^2}{18\mu_{\mathrm{f}}} V(\phi)\left[\big(\varrho_i - \varrho(\Phi)\big) g\mathbf{k} + \frac{\sigma_{\mathrm{e}}(\phi)}{\phi_i}\nabla\Big(\frac{\phi_i}{\phi}\Big) + \frac{1-\phi}{\phi}\nabla\sigma_{\mathrm{e}}(\phi)\right], \quad i = 1, \dots, N. \tag{2.18}$$

The generic assumption to ensure hyperbolicity, which is satisfied by $V(\phi) = (1-\phi)^{n-2}$, $n > 2$, is

$$V(\phi) > 0, \quad V'(\phi) < 0 \quad \text{for } 0 < \phi < \phi_{\max}. \tag{2.19}$$

2.7. Final form of the model equations. The final model equations are the continuity equations of the solids species and of the mixture ($\nabla \cdot \mathbf{q} = 0$), the linear momentum balance of the fluid (2.15), and the equations (2.18) for the slip velocities $\mathbf{u}_i$ derived from the linear momentum balances of the solid species. To derive explicit expressions for the fluxes $\phi_1\mathbf{v}_1, \dots, \phi_N\mathbf{v}_N$ appearing in these equations, we introduce the reduced densities $\bar{\varrho}_{\mathrm{s}} := \varrho_{\mathrm{s}} - \varrho_{\mathrm{f}}$, where ϱ_{s} is the density of the solid particles if they differ only in size, $\bar{\varrho}_i := \varrho_i - \varrho_{\mathrm{f}}$, $i = 1, \dots, N$, the vector $\bar{\boldsymbol{\varrho}} := (\bar{\varrho}_1, \dots, \bar{\varrho}_N)^{\mathrm{T}}$, and the parameters $\mu := -gd_1^2/(18\mu_{\mathrm{f}})$ and $\delta_i := d_i^2/d_1^2$, $i = 1, \dots, N$, such that (2.18) reads

$$\mathbf{u}_i = \mu\delta_i V(\phi)\left[(\bar{\varrho}_i - \bar{\boldsymbol{\varrho}}^{\mathrm{T}}\Phi)\mathbf{k} + \frac{\sigma_{\mathrm{e}}(\phi)}{g\phi_i}\nabla\Big(\frac{\phi_i}{\phi}\Big) + \frac{1-\phi}{g\phi}\nabla\sigma_{\mathrm{e}}(\phi)\right], \quad i = 1, \dots, N. \tag{2.20}$$

From (2.2), we get $\phi_i\mathbf{v}_i = f_i^{\mathrm{M}}(\Phi)\mathbf{k} + \phi_i\mathbf{q} - \mathbf{a}_i(\Phi, \nabla\Phi)$ for $i = 1, \dots, N$, where the components of $\mathbf{f}^{\mathrm{M}}(\Phi)$ (corresponding to the MLB model for suspensions of rigid spheres) are given by

$$f_i(\Phi) = f_i^{\mathrm{M}}(\Phi) = \mu V(\phi)\phi_i\left[\delta_i(\bar{\varrho}_i - \bar{\boldsymbol{\varrho}}^{\mathrm{T}}\Phi) - \sum_{k=1}^{N}\delta_k\phi_k(\bar{\varrho}_k - \bar{\boldsymbol{\varrho}}^{\mathrm{T}}\Phi)\right], \quad i = 1, \dots, N. \tag{2.21}$$

If we let $\boldsymbol{\delta} := (\delta_1, \dots, \delta_N)^{\mathrm{T}}$, then the vectors $\mathbf{a}_i(\Phi, \nabla\Phi)$ are given by

$$\begin{aligned}\mathbf{a}_i(\Phi, \nabla\Phi) = -\frac{\mu V(\phi)}{g}\Bigg\{&\frac{(1-\phi)\phi_i}{\phi}(\delta_i - \boldsymbol{\delta}^{\mathrm{T}}\Phi)\nabla\sigma_{\mathrm{e}}(\phi) \\ &+ \sigma_{\mathrm{e}}(\phi)\left[\delta_i\nabla\Big(\frac{\phi_i}{\phi}\Big) - \phi_i\left(\delta_1\nabla\Big(\frac{\phi_1}{\phi}\Big) + \cdots + \delta_N\nabla\Big(\frac{\phi_N}{\phi}\Big)\right)\right]\Bigg\}, \quad i = 1, \dots, N.\end{aligned} \tag{2.22}$$

The continuity equations for the solids, i.e., for the N unknowns ϕ_1 to ϕ_N, can then be written as

$$\frac{\partial\phi_i}{\partial t} + \nabla\cdot\big(\phi_i\mathbf{q} + f_i^{\mathrm{M}}(\Phi)\mathbf{k}\big) = \nabla\cdot\mathbf{a}_i(\Phi, \nabla\Phi), \quad i = 1, \dots, N. \tag{2.23}$$

Due to the property (2.7), $\mathbf{a}_i(\Phi, \nabla\Phi) = 0$ wherever $\phi \le \phi_{\mathrm{c}}$. At these concentrations, the system (2.23) turns into the first-order system of N scalar equations analyzed in

[22]. The final coupled set of model equations, valid in several space dimensions, is given by (2.23) and the equations

$$\nabla \cdot \mathbf{q} = 0, \tag{2.24}$$

$$\begin{aligned} \nabla p =& -\nabla \sigma_{\mathrm{e}}(\phi) - (\varrho_{\mathrm{f}} + \bar{\boldsymbol{\varrho}} \cdot \Phi) g \mathbf{k} + \frac{1}{1-\phi} \nabla \cdot \mathbf{T}_{\mathrm{f}}^{\mathrm{E}} \\ \equiv& -\nabla \sigma_{\mathrm{e}}(\phi) - \varrho(\Phi) g \mathbf{k} + \frac{1}{1-\phi} \nabla \cdot \mathbf{T}_{\mathrm{f}}^{\mathrm{E}}. \end{aligned} \tag{2.25}$$

Before discussing the role of (2.25), we set $N = 1$ to check consistency with the model of sedimentation of monodisperse flocculated suspensions [26]. With

$$f^{\mathrm{M}}(\phi) = -\frac{g d^2 \bar{\varrho}_{\mathrm{s}}}{18 \mu_{\mathrm{f}}} V(\phi) \phi (1-\phi)^2, \quad \mathbf{a}(\phi, \nabla\phi) = a(\phi) \nabla \phi = -\frac{f^{\mathrm{M}}(\phi)}{\bar{\varrho}_{\mathrm{s}} g \phi} \nabla \sigma_{\mathrm{e}}(\phi),$$

$$a(\phi) = -\frac{f^{\mathrm{M}}(\phi) \sigma_{\mathrm{e}}'(\phi)}{\bar{\varrho}_{\mathrm{s}} g \phi},$$

we see that (2.23) indeed reduces to the scalar equation

$$\frac{\partial \phi}{\partial t} + \nabla \cdot \big(\phi \mathbf{q} + f^{\mathrm{M}}(\phi) \mathbf{k} \big) = \nabla \cdot \big(a(\phi) \nabla \phi \big) \tag{2.26}$$

derived in [26]. It is easy to see that (2.26) is first-order hyperbolic for $\phi \le \phi_{\mathrm{c}}$ and $\phi = 1$, and second-order parabolic for $\phi_{\mathrm{c}} < \phi < 1$, and therefore a strongly degenerate parabolic equation.

Noting that $\mathbf{v}_{\mathrm{f}} = \mathbf{q} - (\phi_1 \mathbf{u}_1 + \cdots + \phi_N \mathbf{u}_N)$, we can rewrite $\mathbf{T}_{\mathrm{f}}^{\mathrm{E}}$ in terms of the mixture velocity $\mathbf{q}$ and the slip velocities $\mathbf{u}_i$, which are now given functions of Φ. For example, if we use the expression $\mathbf{T}_{\mathrm{f}}^{\mathrm{E}} = \mu(\phi)[\nabla \mathbf{v}_{\mathrm{f}} + (\nabla \mathbf{v}_{\mathrm{f}})^{\mathrm{T}} - (2/3)(\nabla \cdot \mathbf{v}_{\mathrm{f}})\mathbf{I}]$ as for a standard viscous-linear fluid but with a concentration-dependent viscosity function, then (2.25) can be rewritten in the form

$$\nabla p = -\varrho(\Phi) g \mathbf{k} + \frac{1}{1-\phi} \Big[\big(\nabla \mu(\phi) \big)^{\mathrm{T}} \big(\nabla \mathbf{q} + (\nabla \mathbf{q})^{\mathrm{T}} \big) + \mu(\phi) \Delta \mathbf{q} \Big] + \mathbf{g}(\Phi, \nabla\Phi, \nabla^2\Phi), \tag{2.27}$$

where $\mathbf{g}$ is a function depending on Φ and the derivatives of its components of up to second order. For pure fluid, i.e., when $\Phi \equiv 0$ (and thus $\mathbf{q} \equiv \mathbf{v}_{\mathrm{f}}$), (2.24) and (2.25) form the Stokes system for an incompressible fluid for the velocity $\mathbf{q}$ and the pressure p.

We now comment on the necessity of retaining a viscosity term, such as $\mu(\phi)\Delta\mathbf{q}$ in (2.25) or (2.27). In fact, deleting *all* terms which are expected to be small according to the dimensional analysis would require that we consider the following equation instead of (2.25):

$$\nabla p = -\nabla \sigma_{\mathrm{e}}(\phi) - \varrho(\Phi) g \mathbf{k}. \tag{2.28}$$

To elucidate the consequences of (2.28), we take the curl of (2.28), which leads to $\partial \varrho(\Phi)/\partial x = \partial \varrho(\Phi)/\partial y = 0$, such that the local density of the mixture depends on height only [93]. For $N = 1$, the implications of this observation are well known [23, 93]. Although the concentration waves (kinematic waves) are one-dimensional, they are embedded in the three-dimensional mixture flow field $\mathbf{q}$. Since $\mathbf{q}$ does not appear in the field equation (2.28), the coupling between the flow field and the kinematic waves

has to be modeled by boundary conditions, which requires introducing boundary layers of sediment or streaming liquid. The resulting kinematic-wave theory has been useful in explaining the behavior of relatively dilute suspensions in vessels with inclined walls [93] or in centrifuges [90, 91]. In [23], this approach is extended to monodisperse suspensions with compressible sediments, for which numerical solutions can be readily obtained. However, it is also shown in [23] that the kinematic-wave theory does not lead to a mathematically well-posed problem, and that this shortcoming is due to the absence of the aforementioned coupling between kinematic waves and the flow field in (2.28). On the other hand, in [24], energy estimates for slight variants of the coupled system (2.23)–(2.25) with $N = 1$ are obtained. These estimates lead to existence and stability results, and follow from the viscosity term in (2.25).

We now consider one space dimension, for which we get $\partial q/\partial z = 0$, and only (2.23) needs to be solved, since q is given by boundary conditions and (2.25) turns into an equation for the pore pressure p, which permits us to calculate this quantity a posteriori from $\phi_1, \dots, \phi_N$.

2.8. Initial and boundary conditions in one space dimension. In a closed one-dimensional vessel, the mixture velocity at the bottom vanishes; hence $q \equiv 0$, and the remaining equations that actually have to be solved are the system of convection-diffusion equations

$$\frac{\partial \phi_i}{\partial t} + \frac{\partial f_i^{\mathrm{M}}(\Phi)}{\partial z} = \frac{\partial}{\partial z}\left[a_i\left(\Phi, \frac{\partial \Phi}{\partial z}\right)\right], \quad i = 1, \dots, N, \tag{2.29}$$

together with an initial concentration distribution and zero flux boundary conditions, i.e.,

$$\Phi(z, 0) = \Phi^0(z) \in \mathcal{D}_{\phi_{\max}}, \ \ 0 \le z \le L, \tag{2.30}$$

$$\phi_i v_i = f_i^{\mathrm{M}}(\Phi) - a_i\left(\Phi, \frac{\partial \Phi}{\partial z}\right) = 0 \quad \text{for } z = 0 \text{ and } z = L,\ i = 1, \dots, N. \tag{2.31}$$

3. Hyperbolicity of the first-order system. We now assume $\varrho_1 = \cdots = \varrho_N = \varrho_{\mathrm{s}}$, so that the components of $\mathbf{f}^{\mathrm{M}}(\Phi)$ are

$$f_i^{\mathrm{M}}(\Phi) = \mu \bar{\varrho}_{\mathrm{s}} V(\phi)(1 - \phi)(\delta_i - \boldsymbol{\delta}^{\mathrm{T}}\Phi)\phi_i, \quad i = 1, \dots, N, \tag{3.1}$$

and we denote by $P(\lambda)$ the characteristic polynomial of $\mathbf{J} := (\mu\bar{\varrho}_{\mathrm{s}})^{-1}\boldsymbol{\mathcal{J}}_{\mathbf{f}^{\mathrm{M}}}(\Phi)$, where $\boldsymbol{\mathcal{J}}_{\mathbf{f}^{\mathrm{M}}}(\Phi) = (\partial f_i^{\mathrm{M}}(\Phi)/\partial\phi_j)_{i,j=1,\dots,N}$ is the Jacobian of $\mathbf{f}^{\mathrm{M}}(\Phi)$. We now derive a closed algebraic expression for $P(\lambda)$. We can write $(\mu\bar{\varrho}_{\mathrm{s}})^{-1}\partial f_i^{\mathrm{M}}(\Phi)/\partial\phi_j = \gamma_j^i(\Phi)\phi_i + \gamma^i(\Phi)\delta_{ij}$ for $i, j = 1, \dots, N$, where

$$\gamma^i(\Phi) := V(\phi)(1 - \phi)(\delta_i - \boldsymbol{\delta}^{\mathrm{T}}\Phi), \quad i = 1, \dots, N, \tag{3.2}$$

$$\gamma_j^i(\Phi) := \big(V(\phi)(1 - \phi)\big)'(\delta_i - \boldsymbol{\delta}^{\mathrm{T}}\Phi) - V(\phi)(1 - \phi)\delta_j, \quad i, j = 1, \dots, N. \tag{3.3}$$

The characteristic polynomial can be written as

$$P(\lambda) := \det(\mathbf{J} - \lambda\mathbf{I}) = \begin{vmatrix} \gamma_1^1\phi_1 + \gamma^1 - \lambda & \gamma_2^1\phi_1 & \cdots & \gamma_N^1\phi_1 \\ \gamma_1^2\phi_2 & \gamma_2^2\phi_2 + \gamma^2 - \lambda & \cdots & \gamma_N^2\phi_2 \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_1^N\phi_N & \gamma_2^N\phi_N & \cdots & \gamma_N^N\phi_N + \gamma^N - \lambda \end{vmatrix}. \tag{3.4}$$

In what follows, we omit the argument Φ and, for later use, note that $\gamma^i - \gamma^l = V(\phi)(1-\phi)(\delta_i - \delta_l)$, which due to $\delta_1 > \delta_2 > \cdots > \delta_N$ implies $\gamma^N < \gamma^{N-1} < \cdots < \gamma^1$. Moreover, we observe that

$$\gamma_j^1 - \gamma_k^1 = \cdots = \gamma_j^N - \gamma_k^N = -V(\phi)(1-\phi)(\delta_j - \delta_k), \quad j,k = 1,\ldots,N,$$
$$\gamma_1^j - \gamma_1^k = \cdots = \gamma_N^j - \gamma_N^k = \big(V(\phi)(1-\phi)\big)'(\delta_j - \delta_k), \quad j,k = 1,\ldots,N.$$

The common values of $\gamma_j^i - \gamma_k^i$ and $\gamma_i^j - \gamma_i^k$ for all $i = 1,\ldots,N$, $j,k = 1,\ldots,N$, will be denoted by $\gamma_{j,k}$ and $\gamma^{j,k}$, respectively. Since $\boldsymbol{\mathcal{J}}_{\mathbf{f}^{\mathrm{M}}}(\Phi)$ and $\mathbf{A}(\Phi)$ have similar structure and therefore similar characteristic polynomials, it is convenient for later use to prove the following lemma separately.

LEMMA 3.1. *The polynomial $P(\lambda)$ defined in* (3.4) *satisfies*

$$P(\lambda) = \left\{1 + \sum_{m=1}^{N} \frac{\phi_m \gamma_m^m}{\gamma^m - \lambda} - \sum_{m=1}^{N} \frac{\phi_m}{\gamma^m - \lambda} \sum_{l=1}^{N} \frac{\phi_l \gamma_{l,N} \gamma^{l,m}}{\gamma^l - \lambda}\right\} \prod_{k=1}^{N} (\gamma^k - \lambda). \tag{3.5}$$

Proof. In this proof we merely use the definitions of $\gamma_{j,k}$ and $\gamma^{j,k}$ in terms of the γ_j^i's and γ_j's. Subtracting column N from columns 1 to $N-1$ in (3.4) yields

$$P(\lambda) = \begin{vmatrix} \gamma_{1,N}\phi_1 + \gamma^1 - \lambda & \cdots & \gamma_{N-1,N}\phi_1 & \gamma_N^1 \phi_1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma_{1,N}\phi_{N-1} & \cdots & \gamma_{N-1,N}\phi_{N-1} + \gamma^{N-1} - \lambda & \gamma_N^{N-1}\phi_{N-1} \\ \gamma_{1,N}\phi_N - \gamma^N + \lambda & \cdots & \gamma_{N-1,N}\phi_N - \gamma^N + \lambda & \gamma_N^N \phi_N + \gamma^N - \lambda \end{vmatrix}. \tag{3.6}$$

Expanding this determinant on the last row, we get

$$P(\lambda) = X + (\gamma^N - \lambda)(-1)^N \big(Y_1 - Y_2 + Y_3 - \cdots + (-1)^N Y_{N-1}\big), \tag{3.7}$$

where X and Y_m are the determinants obtained from the determinant in (3.6) by replacing the last row by $(\gamma_{1,N}\phi_N, \ldots, \gamma_{N-1,N}\phi_N, \gamma_N^N\phi_N + \gamma^N - \lambda)$ and by deleting the last row and the mth column, $m = 1,\ldots,N-1$, respectively. Multiplying the last row in X with $(-\phi_i/\phi_N)$ and adding the result to the ith row, $i = 1,\ldots,N-1$, leads to

$$X = \begin{vmatrix} \gamma^1 - \lambda & \cdots & 0 & \phi_1\big(\gamma^{1,N} - (\gamma^N - \lambda)/\phi_N\big) \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \gamma^{N-1} - \lambda & \phi_{N-1}\big(\gamma^{N-1,N} - (\gamma^N - \lambda)/\phi_N\big) \\ \gamma_{1,N}\phi_N & \cdots & \gamma_{N-1,N}\phi_N & \gamma_N^N\phi_N + \gamma^N - \lambda \end{vmatrix}.$$

Expanding X on the last row yields

$$X = \left(\gamma_N^N\phi_N + \gamma^N - \lambda - \phi_N \sum_{m=1}^{N-1} \frac{\phi_m \gamma_{m,N}}{\gamma^m - \lambda}\left(\gamma^{m,N} - \frac{\gamma^N - \lambda}{\phi_N}\right)\right) \prod_{k=1}^{N-1} (\gamma^k - \lambda). \tag{3.8}$$

Furthermore, we have $Y_m = (-1)^{N-1-m}\tilde{Y}_m$, where

$$\tilde{Y}_m = \begin{vmatrix} \gamma_{1,N}\phi_1 + \gamma^1 - \lambda & \cdots & \gamma_{m-1,N}\phi_1 & \gamma_{m+1,N}\phi_1 & \cdots & \gamma_{N-1,N}\phi_1 & \gamma_N^1\phi_1 \\ \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ \gamma_{1,N}\phi_{m-1} & \cdots & \gamma_{m-1,N}\phi_{m-1} + \gamma^{m-1} - \lambda & \gamma_{m+1,N}\phi_{m-1} & \cdots & \gamma_{N-1,N}\phi_{m-1} & \gamma_N^{m-1}\phi_{m-1} \\ \gamma_{1,N}\phi_{m+1} & \cdots & \gamma_{m-1,N}\phi_{m+1} & \gamma_{m+1,N}\phi_{m+1} + \gamma^{m+1} - \lambda & \cdots & \gamma_{N-1,N}\phi_{m+1} & \gamma_N^{m+1}\phi_{m+1} \\ \vdots & & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{1,N}\phi_{N-1} & \cdots & \gamma_{m-1,N}\phi_{N-1} & \gamma_{m+1,N}\phi_{N-1} & \cdots & \gamma_{N-1,N}\phi_{N-1} + \gamma^{N-1} - \lambda & \gamma_N^{N-1}\phi_{N-1} \\ \gamma_{1,N}\phi_m & \cdots & \gamma_{m-1,N}\phi_m & \gamma_{m+1,N}\phi_m & \cdots & \gamma_{N-1,N}\phi_m & \gamma_N^m\phi_m \end{vmatrix},$$

which implies

$$Y_m = (-1)^{N-1-m}\frac{\phi_m}{\gamma^m - \lambda}\left(\gamma_N^m - \sum_{l=1}^{N-1}\frac{\gamma_{l,N}\gamma^{l,m}\phi_l}{\gamma^l - \lambda}\right)\prod_{k=1}^{N-1}(\gamma^k - \lambda),$$

$$m = 1, \dots, N-1. \tag{3.9}$$

Inserting (3.8) and (3.9) into (3.7), we get

$$\begin{aligned} P(\lambda) =& \left\{\gamma_N^N\phi_N + \gamma^N - \lambda - \phi_N\sum_{m=1}^{N-1}\frac{\phi_m}{\gamma^m - \lambda}\gamma_{m,N}\left(\gamma^{m,N} - \frac{\gamma^N - \lambda}{\phi_N}\right)\right. \\ &\left. + (\gamma^N - \lambda)\sum_{m=1}^{N-1}\frac{\phi_m}{\gamma^m - \lambda}\left(\gamma_N^m - \sum_{l=1}^{N-1}\phi_l\frac{\gamma_{l,N}\gamma^{l,m}}{\gamma^l - \lambda}\right)\right\}\prod_{k=1}^{N-1}(\gamma^k - \lambda) \\ =& \left\{1 - \frac{\phi_N}{\gamma^N - \lambda}\sum_{m=1}^{N-1}\frac{\phi_m\gamma_{m,N}\gamma^{m,N}}{\gamma^m - \lambda} + \sum_{m=1}^{N-1}\frac{\phi_m\gamma_{m,N}}{\gamma^m - \lambda}\right. \\ &\left. + \sum_{m=1}^{N}\frac{\phi_m\gamma_N^m}{\gamma^m - \lambda} - \sum_{m=1}^{N-1}\frac{\phi_m}{\gamma^m - \lambda}\sum_{l=1}^{N-1}\frac{\phi_l\gamma_{l,N}\gamma^{l,m}}{\gamma^l - \lambda}\right\}\prod_{l=1}^{N}(\gamma^l - \lambda). \end{aligned} \tag{3.10}$$

The upper index of summation in the second sum in the second equation of (3.10) can be changed to N since $\gamma_{N,N} = 0$, and the second and third sum can be combined into one using $\gamma_{m,N} + \gamma_N^m = \gamma_m^m$. Furthermore, the first and the fourth sums can be combined into one by changing the upper index of summation for m in the fourth sum from $N-1$ to N, from which we obtain (3.5). ∎

We can now prove the following lemma.

LEMMA 3.2. *Let $\lambda \in \mathbb{R}$ and $\delta(\lambda) := (V(\phi)(1-\phi))^{-1}\lambda + \boldsymbol{\delta}^{\mathrm{T}}\Phi$. Then $P(\lambda)$ is given by*

$$\begin{aligned} P(\lambda) =& \left\{V(\phi)(1-\phi) + \sum_{m=1}^{N}\frac{\phi_m}{\delta_m - \delta(\lambda)}\left[-\delta_m V(\phi)(1-\phi) + \big(V(\phi)(1-\phi)\big)'\right.\right. \\ &\left.\left. \times\left(\delta_m - \boldsymbol{\delta}^{\mathrm{T}}\Phi + \sum_{l=1}^{N}\frac{\delta_l\phi_l(\delta_l - \delta_m)}{\delta_l - \delta(\lambda)}\right)\right]\right\}\big(V(\phi)(1-\phi)\big)^{N-1}\prod_{k=1}^{N}\big(\delta_k - \delta(\lambda)\big). \end{aligned} \tag{3.11}$$

This expression is also well defined for $\lambda \in \{\gamma^1, \dots, \gamma^N\}$ and reads for $k = 1, \dots, N$ as

$$P(\gamma^k) = \phi_k\delta_k\Big\{\big(V(\phi)(1-\phi)\big)' - V(\phi)\Big\}\big(V(\phi)\big)^{N-1}(1-\phi)^N\prod_{\substack{l=1 \\ l\neq k}}^{N}(\delta_l - \delta_k). \tag{3.12}$$

Proof. Using $\gamma^m - \lambda = (\delta_m - \delta(\lambda))V(\phi)(1-\phi)$ and the definitions of $\gamma_{l,N}$ and $\gamma^{l,m}$, we get

(3.13)

$$\begin{aligned}
P(\lambda) =& \Bigg\{ 1 + \sum_{m=1}^{N} \frac{\phi_m[(V(\phi)(1-\phi))'(\delta_m - \boldsymbol{\delta}^{\mathrm{T}}\Phi) - V(\phi)(1-\phi)\delta_m]}{V(\phi)(1-\phi)(\delta_m - \delta(\lambda))} + \frac{(V(\phi)(1-\phi))'}{V(\phi)(1-\phi)} \\
&\times \sum_{m=1}^{N}\sum_{l=1}^{N} \frac{\phi_m\phi_l(\delta_l^2 - \delta_l\delta_m - \delta_l\delta_N + \delta_m\delta_N)}{(\delta_m - \delta(\lambda))(\delta_l - \delta(\lambda))} \Bigg\} \big(V(\phi)(1-\phi)\big)^N \prod_{k=1}^{N} \big(\delta_k - \delta(\lambda)\big) \\
=& \Bigg\{ V(\phi)(1-\phi) + \sum_{m=1}^{N} \frac{\phi_m}{\delta_m - \delta(\lambda)} \Bigg[-V(\phi)(1-\phi)\delta_m + \big(V(\phi)(1-\phi)\big)' \\
&\times \left(\delta_m - \boldsymbol{\delta}^{\mathrm{T}}\Phi + \sum_{l=1}^{N} \frac{\delta_l\phi_l(\delta_l - \delta_m)}{\delta_l - \delta(\lambda)} \right) \Bigg] \Bigg\} \big(V(\phi)(1-\phi)\big)^{N-1} \prod_{k=1}^{N} \big(\delta_k - \delta(\lambda)\big),
\end{aligned}$$

which is (3.11). This expression can be rewritten as

$$\begin{aligned}
P(\lambda) =& \prod_{l=1}^{N} (\gamma^l - \lambda) \\
&+ \Bigg\{ \sum_{m=1}^{N} \phi_m \Big[-\delta_m V(\phi)(1-\phi) + \big(V(\phi)(1-\phi)\big)'(\delta_m - \boldsymbol{\delta}^{\mathrm{T}}\Phi) \Big] \prod_{\substack{l=1 \\ l\neq m}}^{N} \big(\delta_l - \delta(\lambda)\big) \\
&+ \big(V(\phi)(1-\phi)\big)' \sum_{m=1}^{N}\sum_{l=1}^{N} \phi_m\phi_l\delta_l(\delta_l - \delta_m) \prod_{\substack{n=1 \\ n\neq m,l}}^{N} \big(\delta_n - \delta(\lambda)\big) \Bigg\} V(\phi)(1-\phi)^{N-1}.
\end{aligned}$$

For $\lambda = \gamma^k$ the first product vanishes, and in the first sum only the summand with $m = k$ and in the second sum only the summands with $m = k$ or $l = k$ do not vanish. This implies

$$\begin{aligned}
P(\gamma^k) =& \Bigg\{ \phi_k \Big[-\delta_k V(\phi)(1-\phi) + \big(V(\phi)(1-\phi)\big)'(\delta_k - \boldsymbol{\delta}^{\mathrm{T}}\Phi) \Big] \prod_{\substack{l=1 \\ l\neq k}}^{N} (\delta_l - \delta_k) \\
&+ \big(V(\phi)(1-\phi)\big)' \Bigg[\phi_k \sum_{l=1}^{N} \Bigg(\delta_l\phi_l(\delta_l - \delta_k) \prod_{\substack{n=1 \\ n\neq k,l}}^{N} (\delta_n - \delta_k) \Bigg) \\
&+ \sum_{m=1}^{N} \Bigg(\phi_m\delta_k\phi_k(\delta_k - \delta_m) \prod_{\substack{n=1 \\ n\neq k,m}}^{N} (\delta_n - \delta_k) \Bigg) \Bigg] \Bigg\} \big(V(\phi)(1-\phi)\big)^{N-1} \\
=& \Big\{ -\delta_k\phi_k V(\phi)(1-\phi) + \delta_k\phi_k\big(V(\phi)(1-\phi)\big)' - \phi_k\boldsymbol{\delta}^{\mathrm{T}}\Phi\big(V(\phi)(1-\phi)\big)' \\
&+ \phi_k\boldsymbol{\delta}^{\mathrm{T}}\Phi\big(V(\phi)(1-\phi)\big)' - \delta_k\phi_k\phi\big(V(\phi)(1-\phi)\big)' \Big\} \\
&\times \big(V(\phi)(1-\phi)\big)^{N-1} \prod_{\substack{l=1 \\ l\neq k}}^{N} (\delta_l - \delta_k),
\end{aligned}$$

from which (3.12) can be read off immediately. ∎

THEOREM 3.3. *If* $\varrho_1 = \cdots = \varrho_N = \varrho_s$, $\delta_1 > \delta_2 > \cdots > \delta_N$, *and* $\Phi \in \mathcal{D}^0_{\phi_{\max}}$, *then the system* (1.4) *is strictly hyperbolic; i.e., the Jacobian* $\boldsymbol{\mathcal{J}}_{\mathbf{f}^M}(\Phi)$ *has* N *distinct real eigenvalues.*

Proof. From (3.12) we see that

$$P(\gamma^k) = C_k\big(V(\phi)\big)^{N-1}(1-\phi)^N \prod_{\substack{m=1\\ m\neq k}}^{N} (\delta_m - \delta_k), \quad k = 1, \dots, N,$$

with $C_k := \delta_k\phi_k(V'(\phi)(1-\phi) - 2V(\phi))$. Since $C_1, \dots, C_N < 0$ on $\mathcal{D}^0_{\phi_{\max}}$ due to (2.19), we have

$$\operatorname{sgn}\big(P(\gamma^k)\big) = -\operatorname{sgn}\left(\prod_{\substack{m=1\\ m\neq k}}^{N} (\delta_m - \delta_k)\right) = -\operatorname{sgn}\left(\prod_{m=k+1}^{N} (\delta_m - \delta_k)\right) = (-1)^{N-k+1}.$$

Consequently, we have shown that $\operatorname{sgn}(P(\gamma^i)) = (-1)^{N+1-i}$ for $i = 1, \dots, N$. Whether N is even or odd, we have $P(\lambda) \to \infty$ as $\lambda \to -\infty$ and $P(\gamma^N) < 0$. In view of $\gamma^N < \gamma^{N-1} < \cdots < \gamma^1$ and since $P(\gamma^N) < 0$, there exists a number $\lambda_N < \gamma^N$ with $P(\lambda_N) = 0$. Furthermore, $\operatorname{sgn}(P(\gamma^i)) = (-1)^{N+1-i}$ implies that there exist $N-1$ numbers $\lambda_i \in (\gamma^{i+1}, \gamma^i)$, $i = 1, \dots, N-1$, with $P(\lambda_i) = 0$. This shows that $P(\lambda) = \det(\mathbf{J} - \lambda\mathbf{I})$ has N roots $\lambda_1, \dots, \lambda_N$ satisfying

$$\lambda_N < \gamma^N < \lambda_{N-1} < \gamma^{N-1} < \cdots < \lambda_2 < \gamma^2 < \lambda_1 < \gamma^1. \tag{3.14}$$

Thus the system (1.4) is strictly hyperbolic for all $\Phi \in \mathcal{D}^0_{\phi_{\max}}$, and Theorem 3.3 is proved. ∎

The statement of Theorem 3.3 can still be improved. In fact, it is desirable to have lower and upper bounds for all eigenvalues of $\boldsymbol{\mathcal{J}}_{\mathbf{f}^M}(\Phi)$. However, in (3.14) a lower bound for the eigenvalue λ_N of $\mathbf{J}$ is still lacking. The following theorem shows that by evaluating $P(\lambda)$ at a suitable number $\gamma^\infty < \gamma^1$ it is indeed possible to provide that bound.

THEOREM 3.4. *Define* $\gamma^\infty := -2\boldsymbol{\delta}^{\mathrm{T}}\Phi V(\phi)(1-\phi) + (V(\phi)(1-\phi))'(\boldsymbol{\delta}^{\mathrm{T}}\Phi + \phi)$. *Then, under the conditions of Theorem* 3.3, *the eigenvalues* $\nu_1(\Phi), \dots, \nu_N(\Phi)$ *of* $\boldsymbol{\mathcal{J}}_{\mathbf{f}^M}(\Phi)$ *satisfy*

$$\nu_i(\Phi) \in \big(\mu\bar{\varrho}_s V(\phi)(1-\phi)(\delta_i - \boldsymbol{\delta}^{\mathrm{T}}\Phi), \mu\bar{\varrho}_s V(\phi)(1-\phi)(\delta_{i+1} - \boldsymbol{\delta}^{\mathrm{T}}\Phi)\big), \quad i = 1, \dots, N-1, \tag{3.15}$$

$$\nu_N(\Phi) \in \big(\mu\bar{\varrho}_s V(\phi)(1-\phi)(\delta_N - \boldsymbol{\delta}^{\mathrm{T}}\Phi), \mu\bar{\varrho}_s\gamma^\infty\big). \tag{3.16}$$

Proof. We first evaluate $P(\lambda)$ assuming that $\delta(\lambda) < 0$. Moreover, to estimate the factor in curled brackets in the second equation of (3.13), we use $(V(\phi)(1-\phi))'/V(\phi)(1-\phi) < 0$ to justify deleting $-\delta_m$ and replacing δ_l^2 by δ_l in the last sum. Furthermore, we use $1/(\delta_m - \delta(\gamma^\infty)) < -1/\delta(\gamma^\infty)$ in several instances, which leads to

$$\begin{aligned}
&1 + \sum_{m=1}^{N} \frac{\phi_m}{\delta_m - \delta(\lambda)}\left[-\delta_m + \frac{(V(\phi)(1-\phi))'}{V(\phi)(1-\phi)}\left(\delta_m - \boldsymbol{\delta}^{\mathrm{T}}\Phi + \sum_{l=1}^{N} \frac{\delta_l\phi_l(\delta_l - \delta_m)}{\delta_l - \delta(\lambda)}\right)\right] \\
&\geq 1 - \frac{\boldsymbol{\delta}^{\mathrm{T}}\Phi}{\delta(\lambda)}\left(\frac{(V(\phi)(1-\phi))'}{V(\phi)(1-\phi)} - 1\right) \\
&\quad + \frac{(V(\phi)(1-\phi))'}{V(\phi)(1-\phi)}\left(\frac{\phi}{\delta(\lambda)^2}\sum_{l=1}^{N}\delta_l\phi_l - \boldsymbol{\delta}^{\mathrm{T}}\Phi\sum_{m=1}^{N}\frac{\phi_m}{\delta_m - \delta(\lambda)}\right).
\end{aligned}$$

If we delete the last sum, the left-hand part of this inequality will remain positive whenever

$$\delta^2(\lambda) + \delta(\lambda)\boldsymbol{\delta}^{\mathrm{T}}\Phi\left[1 - \frac{(V(\phi)(1-\phi))'}{V(\phi)(1-\phi)}\right] + \boldsymbol{\delta}^{\mathrm{T}}\Phi\phi\frac{(V(\phi)(1-\phi))'}{V(\phi)(1-\phi)} > 0.$$

This can be achieved by letting $\lambda = \gamma^\infty$ such that

$$\delta(\gamma^\infty) = \boldsymbol{\delta}^{\mathrm{T}}\Phi\left[\frac{(V(\phi)(1-\phi))'}{V(\phi)(1-\phi)} - 1\right] + \frac{(V(\phi)(1-\phi))'}{V(\phi)(1-\phi)}\phi.$$

In fact, making obvious simplifications, we then obtain $\phi(\phi + \boldsymbol{\delta}^{\mathrm{T}}\Phi) > 0$, and the inequality is proved. Since $\gamma^\infty < \gamma^N$, $P(\gamma^N) < 0$, and $P(\gamma^\infty) > 0$, the smallest eigenvalue λ_N of $\mathbf{J}$ satisfies $\gamma^\infty < \lambda_N < \gamma^N$. Combining this with (3.14) and recalling that the eigenvalues ν_i of $\boldsymbol{\mathcal{J}}_{\mathbf{f}^{\mathrm{M}}}(\Phi)$ are given by $\nu_i = \mu\bar{\varrho}_{\mathrm{s}}\lambda_i$, $i = 1, \ldots, N$, we obtain the statement of Theorem 3.4. ∎

4. Properties of the diffusion matrix. Using

$$\frac{\partial}{\partial z}\left(\frac{\phi_i}{\phi}\right) = \frac{1}{\phi}\frac{\partial\phi_i}{\partial z} - \frac{\phi_i}{\phi^2}\left(\frac{\partial\phi_1}{\partial z} + \cdots + \frac{\partial\phi_N}{\partial z}\right) = \frac{1}{\phi}\left\{\frac{\partial\phi_i}{\partial z} - \frac{\phi_i}{\phi}\left(\frac{\partial\phi_1}{\partial z} + \cdots + \frac{\partial\phi_N}{\partial z}\right)\right\}$$

for $i = 1, \ldots, N$ and defining $W(\phi) := -\mu V(\phi)/(g\phi)$ and

(4.1)
$$\eta_{ij}(\Phi) := W(\phi)\left\{(1-\phi)\phi_i(\delta_i - \boldsymbol{\delta}^{\mathrm{T}}\Phi)\sigma_{\mathrm{e}}'(\phi) + \left[\delta_i\delta_{ij} - \delta_j\phi_i - \frac{\phi_i}{\phi}(\delta_i - \boldsymbol{\delta}^{\mathrm{T}}\Phi)\right]\sigma_{\mathrm{e}}(\phi)\right\}$$

for $i, j = 1, \ldots, N$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise, we get from (2.22)

$$a_i\left(\Phi, \frac{\partial\Phi}{\partial z}\right) = \eta_{i1}(\Phi)\frac{\partial\phi_1}{\partial z} + \cdots + \eta_{iN}(\Phi)\frac{\partial\phi_N}{\partial z}, \quad i = 1, \ldots, N. \tag{4.2}$$

Defining the matrix $\mathbf{A}(\Phi) := (\eta_{ij}(\Phi))_{1\le i,j\le N}$ and taking $\mathbf{f} = \mathbf{f}^{\mathrm{M}}$, we can rewrite (2.29) in the form (1.2). We show that the eigenvalues of $\mathbf{A}(\Phi)$ are positive and pairwise distinct on $\mathcal{D}^0_{\phi_{\max}}\backslash\mathcal{D}_{\phi_{\mathrm{c}}}$ by evaluating the characteristic polynomial in a fashion similar to section 3. To this end, we first provide an explicit expression for $S(\lambda) := \det(W(\phi)^{-1}\mathbf{A}(\Phi) - \lambda\mathbf{I})$.

LEMMA 4.1. *Let $\delta^*(\lambda) := \lambda/\sigma_{\mathrm{e}}(\phi)$. Then the polynomial $S(\lambda)$ is given by*

$$\begin{aligned} S(\lambda) = &\left\{\sigma_{\mathrm{e}}(\phi) + \sum_{m=1}^{N}\frac{\phi_m}{\delta_m - \delta^*(\lambda)}\left[-\delta_m\sigma_{\mathrm{e}}(\phi) + \left((1-\phi)\sigma_{\mathrm{e}}'(\phi) - \frac{\sigma_{\mathrm{e}}(\phi)}{\phi}\right)\right.\right. \\ &\left.\left.\times\left(\delta_m - \boldsymbol{\delta}^{\mathrm{T}}\Phi + \sum_{l=1}^{N}\frac{\phi_l\delta_l(\delta_l - \delta_m)}{\delta_l - \delta^*(\lambda)}\right)\right]\right\}\big(\sigma_{\mathrm{e}}(\phi)\big)^{N-1}\prod_{k=1}^{N}\big(\delta_k - \delta^*(\lambda)\big). \end{aligned} \tag{4.3}$$

Proof. We write $\eta_{ij}(\Phi)/W(\phi) = s^i_j\phi_i + s^i\delta_{ij}$ for $1 \le i, j \le N$, where we define

$$s^i := \sigma_{\mathrm{e}}(\phi)\delta_i, \quad s^i_j := (1-\phi)(\delta_i - \boldsymbol{\delta}^{\mathrm{T}}\Phi)\sigma_{\mathrm{e}}'(\phi) - \left(\delta_j + \frac{1}{\phi}(\delta_i - \boldsymbol{\delta}^{\mathrm{T}}\Phi)\right)\sigma_{\mathrm{e}}(\phi)$$

for $i,j=1,\dots,N$. Consequently, $S(\lambda)$ can be written as

$$S(\lambda)=\begin{vmatrix} s_1^1\phi_1+s^1-\lambda & s_2^1\phi_1 & \dots & s_N^1\phi_1 \\ s_1^2\phi_2 & s_2^2\phi_2+s^2-\lambda & \dots & s_N^2\phi_2 \\ \vdots & \vdots & \ddots & \vdots \\ s_1^N\phi_N & s_2^N\phi_N & \dots & s_N^N\phi_N+s^N-\lambda \end{vmatrix}.$$

Observe that the numbers s_j^i satisfy

$$s_j^i-s_k^i=-\sigma_{\rm e}(\phi)(\delta_j-\delta_k),\qquad s_i^j-s_i^k=(\delta_j-\delta_k)\left[(1-\phi)\sigma_{\rm e}'(\phi)-\frac{\sigma_{\rm e}(\phi)}{\phi}\right],\qquad i=1,\dots,N; \tag{4.4}$$

i.e., the right-hand parts of (4.4) do not depend on i. Therefore, we may introduce

$$s_{j,k}:=s_j^1-s_k^1=\cdots=s_j^N-s_k^N,\quad s^{j,k}:=s_1^j-s_1^k=\cdots=s_N^j-s_N^k,\quad j,k=1,\dots,N.$$

We can now easily provide an explicit expression for $S(\lambda)$ in terms of the s's, since the rules for the s's correspond to those for the γ's. Thus, replacing $V(\phi)(1-\phi)$ by $\sigma_{\rm e}(\phi)$, $(V(\phi)(1-\phi))'$ by $(1-\phi)\sigma_{\rm e}'(\phi)-\sigma_{\rm e}(\phi)/\phi$, $\delta(\lambda)$ by $\delta^*(\lambda)=\lambda/\sigma_{\rm e}(\phi)$, we obtain (4.3) by closely following the proofs of Lemma 3.1 and of (3.11) in Lemma 3.2. ∎

To localize the eigenvalues of $\mathbf{A}(\Phi)$, we need to evaluate $S(\lambda)$ at $\lambda=0$ and $\lambda=s^1,\dots,s^N$. Using the analogy between $P(\lambda)$ and $S(\lambda)$, we can easily prove the following lemma.

LEMMA 4.2. *The determinant of* $\mathbf{A}(\Phi)$ *is given by* $\det(\mathbf{A}(\Phi))=(W(\phi))^N S(0)$, *where*

$$S(0)=\delta_1\cdots\delta_N\big(\sigma_{\rm e}(\phi)\big)^{N-1}\sigma_{\rm e}'(\phi)\phi(1-\phi)^2\quad \text{for } 0\le\phi\le\phi_{\max} \text{ and } N\ge 2. \tag{4.5}$$

Moreover, for $k=1,\dots,N$ *we have*

$$S(s^k)=\phi_k\delta_k\left\{(1-\phi)\left[(1-\phi)\sigma_{\rm e}'(\phi)-\frac{\sigma_{\rm e}(\phi)}{\phi}\right]-\sigma_{\rm e}(\phi)\right\}\big(\sigma_{\rm e}(\phi)\big)^{N-1}\prod_{\substack{m=1\\ m\ne k}}^{N}(\delta_m-\delta_k). \tag{4.6}$$

Proof. We set $\lambda=\delta^*(\lambda)=0$ in (4.3). Then (4.5) follows from

$$\begin{aligned} S(0)=&\left\{\sigma_{\rm e}(\phi)+\sum_{m=1}^{N}\frac{\phi_m}{\delta_m}\left[-\delta_m\sigma_{\rm e}(\phi)+\left((1-\phi)\sigma_{\rm e}'(\phi)-\frac{\sigma_{\rm e}(\phi)}{\phi}\right)\right.\right.\\ &\left.\left.\times(\delta_m-\boldsymbol{\delta}^{\rm T}\Phi+\boldsymbol{\delta}^{\rm T}\Phi-\delta_m\phi)\right]\right\}\big(\sigma_{\rm e}(\phi)\big)^{N-1}\delta_1\cdots\delta_N\\ =&\left\{\sigma_{\rm e}(\phi)-\phi\sigma_{\rm e}(\phi)+(1-\phi)\phi\left((1-\phi)\sigma_{\rm e}'(\phi)-\frac{\sigma_{\rm e}(\phi)}{\phi}\right)\right\}\big(\sigma_{\rm e}(\phi)\big)^{N-1}\delta_1\cdots\delta_N. \end{aligned} \tag{4.7}$$

Equation (4.6) can then be derived by closely following the proof of (3.12) in Lemma 3.2. ∎

THEOREM 4.3. *Let* $G(\phi):=\phi(1-\phi)^2\sigma_{\rm e}'(\phi)-\sigma_{\rm e}(\phi)$, *and assume that* $V(\phi)\ne 0$ *for* $\phi<\phi_{\max}$ *and* $V(\phi)=0$ *otherwise. Then, for all* $\Phi\in\mathcal{D}^0_{\phi_{\max}}\backslash\mathcal{D}_{\phi_{\rm c}}$, *the matrix* $\mathbf{A}(\Phi)$ *has* N *distinct positive eigenvalues* $\Lambda_1,\dots,\Lambda_N$; *i.e., the system* (1.2) *is strictly parabolic on* $\mathcal{D}^0_{\phi_{\max}}\backslash\mathcal{D}_{\phi_{\rm c}}$. *Moreover, we have the following:*

(a) *If Φ is chosen such that $G(\phi) > 0$, then these eigenvalues satisfy*

$$\begin{aligned} 0 < W(\phi)\sigma_{\mathrm{e}}(\phi)\delta_N < \Lambda_N < W(\phi)\sigma_{\mathrm{e}}(\phi)\delta_{N-1} < \Lambda_{N-1} \\ < \cdots < W(\phi)\sigma_{\mathrm{e}}(\phi)\delta_1 < \Lambda_1 < W(\phi)\delta_1\phi(1-\phi)^2\sigma_{\mathrm{e}}'(\phi). \end{aligned} \tag{4.8}$$

(b) *At those points Φ where $G(\phi) < 0$, we have*

$$\begin{aligned} 0 < W(\phi)\delta_N\phi(1-\phi)^2\sigma_{\mathrm{e}}'(\phi) < \Lambda_N < W(\phi)\sigma_{\mathrm{e}}(\phi)\delta_N < \Lambda_{N-1} \\ < W(\phi)\sigma_{\mathrm{e}}(\phi)\delta_{N-1} < \cdots < \Lambda_1 < W(\phi)\sigma_{\mathrm{e}}(\phi)\delta_1. \end{aligned} \tag{4.9}$$

(c) *If $G(\phi) = 0$, then the eigenvalues are given by $\Lambda_i = W(\phi)\sigma_{\mathrm{e}}(\phi)\delta_i$ for $i = 1, \ldots, N$.*

Proof. Using the function $G(\phi)$, we can rewrite (4.6) as

$$S(s^k) = \frac{\phi_k}{\phi}\delta_k G(\phi)\big(\sigma_{\mathrm{e}}(\phi)\big)^{N-1} \prod_{\substack{m=1 \\ m\neq k}}^{N} (\delta_m - \delta_k), \quad k = 1, \ldots, N. \tag{4.10}$$

This implies

$$\begin{aligned} \operatorname{sgn}\big(S(s^k)\big) &= \operatorname{sgn}\big(G(\phi)\big) \cdot \operatorname{sgn}\left(\prod_{\substack{m=1 \\ m\neq k}}^{N} (\delta_m - \delta_k) \right) \\ &= (-1)^{N-k} \operatorname{sgn}\big(G(\phi)\big), \quad k = 1, \ldots, N. \end{aligned} \tag{4.11}$$

Recall first that, for $\phi > \phi_{\mathrm{c}}$ and due to $\delta_1 > \delta_2 > \cdots > \delta_N$, we have $0 < s^N < s^{N-1} < \cdots < s^1$. If $\operatorname{sgn}(G(\phi)) = 1$, then $S(s^N) > 0$, $S(s^{N-1}) < 0$, and so on, until we obtain $S(s^2) > 0$ and $S(s^1) < 0$ if N is even and $S(s^2) < 0$ and $S(s^1) > 0$ if N is odd. Thus there exist $N - 1$ values

$$0 < s^N < \lambda_N < s^{N-1} < \lambda_{N-1} < s^{N-2} < \cdots < \lambda_3 < s^2 < \lambda_2 < s^1 < \lambda_1 \tag{4.12}$$

with $S(\lambda_2) = \cdots = S(\lambda_N) = 0$. Moreover, $S(\lambda) \to \infty$ for $\lambda \to \infty$ if N is even and $S(\lambda) \to -\infty$ if N is odd. Thus there exists an Nth number $\lambda_1 > s^1$ with $S(\lambda_1) = 0$. Since the determinant of a matrix is the product of its eigenvalues, which are all positive here, (4.7) implies

$$\begin{aligned} \lambda_1 = \frac{S(0)}{\lambda_2 \cdots \lambda_N} < \frac{S(0)}{s^2 \cdots s^N} &= \frac{\phi(1-\phi)^2\sigma_{\mathrm{e}}'(\phi)(\sigma_{\mathrm{e}}(\phi))^{N-1}\delta_1 \cdots \delta_N}{\delta_2 \cdots \delta_N (\sigma_{\mathrm{e}}(\phi))^{N-1}} \\ &= \delta_1\phi(1-\phi)^2\sigma_{\mathrm{e}}'(\phi). \end{aligned} \tag{4.13}$$

If $\operatorname{sgn}(G(\phi)) = -1$, then $S(s^N) < 0$, $S(s^{N-1}) > 0$, and so on, and $S(s^2) < 0$, $S(s^1) > 0$ if N is even, and $S(s^2) > 0$, $S(s^1) < 0$ if N is odd. This means that we have $N - 1$ values

$$s^N < \lambda_{N-1} < s^{N-1} < \lambda_{N-2} < s^{N-2} < \cdots < \lambda_1 < s^1 \tag{4.14}$$

with $S(\lambda_1) = \cdots = S(\lambda_{N-1}) = 0$. Since $S(s^N) < 0$ but $S(0) > 0$ due to Lemma 4.2, there exists an Nth value $\lambda_N \in (0, s^N)$ satisfying $S(\lambda_N) = 0$, and we have

$$\begin{aligned} \lambda_N = \frac{S(0)}{\lambda_1 \cdots \lambda_{N-1}} > \frac{S(0)}{s^1 \cdots s^{N-1}} &= \frac{\phi(1-\phi)^2\sigma_{\mathrm{e}}'(\phi)(\sigma_{\mathrm{e}}(\phi))^{N-1}\delta_1 \cdots \delta_N}{\delta_1 \cdots \delta_{N-1} (\sigma_{\mathrm{e}}(\phi))^{N-1}} \\ &= \delta_N\phi(1-\phi)^2\sigma_{\mathrm{e}}'(\phi). \end{aligned} \tag{4.15}$$

The eigenvalues of $\mathbf{A}(\Phi)$ are given by $\Lambda_i = W(\phi)\lambda_i$, $i = 1, \ldots, N$. Thus, parts (a) and (b) of Theorem 4.3 follow from (4.12)–(4.15). Part (c) is the common limit for $G(\phi) \uparrow 0$ and $G(\phi) \downarrow 0$. □

Since the eigenvalues $\Lambda_1, \ldots, \Lambda_N$ are positive independent of the sign of $G(\phi)$, we see that the system (1.2) is strictly parabolic for all Φ satisfying $\phi_c < \phi < 1$, although, due to the properties of σ_e and W, at least $N - 1$ of these eigenvalues approach zero as $\phi \downarrow \phi_c$ or $\phi \uparrow \phi_{max}$.

5. Strongly degenerate parabolic problems. We have demonstrated that polydisperse sedimentation models taking into account compression effects give rise to strongly degenerate parabolic (also known as mixed hyperbolic-parabolic) systems of PDEs. The general theory of *uniformly parabolic* systems is an old subject and is by now well developed; see [42, 63, 104]. One can consult [100] for some special uniformly parabolic systems, as well as [36, 58, 85] for some results on parabolic systems with weaker parabolicity conditions. The general mathematical theory of *hyperbolic* systems is also fairly well developed (at least in one spatial dimension); see, for example, [32] and the references therein. On the other hand, to date there exists no general theory for *strongly degenerate parabolic* systems. However, the mathematical theory for *scalar* strongly degenerate parabolic equations has advanced significantly in the last few years. It is well known that nonlinear degenerate parabolic equations exhibit "hyperbolic phenomena" like finite speed of propagation or the appearance of interfaces. These effects are consequences of the partial loss of parabolicity. *Strongly* degenerate parabolic equations (e.g., those arising in the theory of sedimentation-consolidation processes) exhibit even more novel hyperbolic features such as the appearance of shock waves, loss of uniqueness, and the need for entropy conditions. Recall that a simple example of a strongly degenerate equation is a hyperbolic equation. Hence, strongly degenerate parabolic equations will in general possess discontinuous (weak) solutions. Moreover, discontinuous solutions are not uniquely determined by their initial (and boundary) data. In fact, an additional condition—the entropy condition—is needed to single out the physically relevant weak solution of the problem.

An entropy condition for strongly degenerate parabolic equations was first proposed in [118], which also established existence of an entropy solution by passing to the limit in a parabolic regularization. In the one-dimensional case, uniqueness of the entropy solution was proved in [119, 120]; see also [6, 7, 8]. Uniqueness of entropy solutions for multidimensional equations was obtained in the recent work [28] for a particular homogeneous boundary value problem. Extensions of this uniqueness result to the initial value problem can be found in [61, 62] for bounded entropy solutions (of more general equations). Uniqueness for unbounded entropy solutions and kinetic solutions is studied in [30] and [31], respectively. The inhomogeneous Dirichlet boundary value problem is treated in [72]. Some other boundary value problems arising in the theory of sedimentation-consolidation processes are studied in [17, 21, 25]. Weakly coupled systems of (strongly) degenerate parabolic equations are treated in [57].

Following up the recent development of a well-posedness theory for scalar strongly degenerate parabolic equations, there has also been a lot of activity on the design and analysis of numerical methods for such equations. Most of this activity can be seen as natural extensions of ideas and techniques from the hyperbolic numerical literature. Let us here mention the studies on monotone finite difference schemes [45], operator splitting methods (see [44] for an overview), finite volume schemes [46, 80], central finite difference schemes [64], the local discontinuous Galerkin method [33], and BGK

schemes [2, 12]. Numerical methods for parabolic systems (with weak degeneracy) are studied and analyzed in [60, 68]. Applications of operator splitting methods and finite difference schemes to scalar sedimentation-consolidation models can be found in [18] and [20], respectively.

In the next section, we will present and apply certain numerical schemes for systems of strongly degenerate parabolic equations. Except for [2], the available numerical literature has so far dealt with *scalar* strongly degenerate parabolic equations. Let us add that the generality in [2] is such that it does not include systems of the form considered in the present paper.

6. Numerical results. The Kurganov–Tadmor (KT) scheme [64] can be regarded as a refinement of the essentially nonoscillatory Nessyahu–Tadmor scheme [77], where the improvement is based on local estimates of the propagation velocities of the Riemann fan emerging from the cell boundaries during each time step. Thus, the accuracy of the resulting scheme depends on how accurately the eigenvalues of the Jacobian of the flux vector are determined. Since only for small systems can these eigenvalues be determined exactly, it is important for large N that sharp estimates can be obtained with low computational effort. The analysis of section 3 indeed provides sharp estimates for the first-order system of equations. Given the importance of these analytical results for the KT scheme, we give in what follows a rather compressed but complete description of this scheme. A general introduction to central schemes for systems of conservation laws is given in [103].

6.1. General difference scheme. Consider the computational domain $Q_T := [0,1] \times [0,T]$ and a rectangular grid defined by $z_j := j\Delta z$, $j = 0, \dots, J$, where J is an even integer and $\Delta z := 1/J$ is the width of a half-cell, and $t_n := n\Delta t$, $n = 0, \dots, \mathcal{N}$, where $\Delta t := T/\mathcal{N}$, $\mathcal{N} \in \mathbb{N}$, and $\lambda := \Delta t/(2\Delta z)$ is the fixed mesh-size ratio. (Thus, all grid-point indices are integers.) The (approximate) cell average of ϕ_i, $i = 1, \dots, N$, with respect to the cell $[z_j, z_{j+2}]$ at time t_n is denoted by $\bar\phi^n_{i,j}$, and we define $\bar\Phi^n_j := (\bar\phi^n_{1,j}, \dots, \bar\phi^n_{N,j})^{\mathrm{T}}$, $j = 1, 3, \dots, J-1$, $n = 0, 1, \dots, \mathcal{N}$. We assume that at time $t = t_n$, $n = 0, 1, \dots, \mathcal{N}-1$, the vector $\bar\Phi^n_j$ either has been calculated from the previous time step (for $n \ge 1$) or is given by the discretization of the initial condition,

$$\bar\phi^0_{i,j} := \frac{1}{2\Delta z} \int_{z_{j-1}}^{z_{j+1}} \phi^0_i(\zeta)\, d\zeta, \quad j = 1, 3, \dots, J-1, \quad i = 1, \dots, N.$$

For the interior cells, the general scheme ("interior scheme") is of the type

$$\bar\Phi^{n+1}_j = \bar\Phi^n_j - \lambda(\mathbf{h}^n_{j+1} - \mathbf{h}^n_{j-1}) + \lambda(\mathbf{p}^n_{j+1} - \mathbf{p}^n_{j-1}), \qquad \begin{matrix} j = 3, 5, \dots, J-3, \\ n = 0, \dots, \mathcal{N}-1, \end{matrix} \tag{6.1}$$

where $\mathbf{h}^n_{j\pm1}$ and $\mathbf{p}^n_{j\pm1}$ are approximations of the "hyperbolic" and "parabolic" fluxes $\mathbf{f}^{\mathrm{M}}$ and $\mathbf{a}$, respectively, through the boundaries of cell $I_j := [z_{j-1}, z_{j+1}]$ at time t_n. The detailed computation of these fluxes from the solution values at time t_n is described in section 6.2.

While the interior scheme (6.1) approximates the field equation (1.2), the boundary conditions (2.31) are discretized by setting $\mathbf{h}^n_0 - \mathbf{p}^n_0 = 0$ and $\mathbf{h}^n_J - \mathbf{p}^n_J = 0$ for $n = 0, \dots, \mathcal{N}-1$. Inserting this into (6.1), where we set $j = 1$ and $j = J-1$, we obtain the following "boundary scheme":

$$\begin{gathered} \bar\Phi^{n+1}_1 = \bar\Phi^n_1 - \lambda\mathbf{h}^n_2 + \lambda\mathbf{p}^n_2, \qquad \bar\Phi^{n+1}_{J-1} = \bar\Phi^n_{J-1} + \lambda\mathbf{h}^n_{J-2} - \lambda\mathbf{p}^n_{J-2}, \\ n = 0, \dots, \mathcal{N}-1. \end{gathered} \tag{6.2}$$

The extension of the CFL stability condition for the explicit KT scheme stated in [64] for scalar equations to the present case of a strongly degenerate parabolic-hyperbolic problem reads

$$\frac{\Delta t}{\Delta z}\max_{\mathcal{D}_{\phi_{\max}}}\rho\big(\boldsymbol{\mathcal{J}}_{\mathbf{f}}(\Phi)\big)+\frac{\Delta t}{2\Delta z^2}\max_{\mathcal{D}_{\phi_{\max}}}\rho\big(\mathbf{A}(\Phi)\big)\le\frac{1}{4}, \tag{6.3}$$

where $\rho(\cdot)$ denotes the spectral radius. We view (6.3) as a *necessary* condition for the present explicit KT scheme to produce a physically relevant numerical result, and we emphasize that no rigorous convergence result is associated with (6.3). For that matter, an existence and uniqueness theory for the system (1.1) is still lacking.

6.2. Computation of the numerical fluxes. Given the vectors $\bar\Phi_j^n$, $j=1,3,\dots,J-1$, we calculate a piecewise linear reconstruction of the solution values at time t_n by determining the slope vector $\Phi_j'=(\phi_{1,j}',\dots,\phi_{N,j}')^{\mathrm{T}}$, $j=1,3,\dots,J-1$, whose components are defined by

$$\phi_{i,j}':=\begin{cases}0 & \text{for } j=1 \text{ and } j=J-1,\\ \mathrm{MM}\big(\theta(\bar\phi_{i,j}^n-\bar\phi_{i,j-2}^n),(\bar\phi_{i,j+2}^n-\bar\phi_{i,j-2}^n)/2,\\ \qquad\qquad \theta(\bar\phi_{i,j+2}^n-\bar\phi_{i,j}^n)\big) & \text{for } j=3,5,\dots,J-3\end{cases}$$

for $i=1,\dots,N$. Here $\mathrm{MM}(\cdot,\cdot,\cdot)$ is the minmod function given by $\mathrm{MM}(a,b,c)=\min(a,b,c)$ if $a,b,c>0$, $\mathrm{MM}(a,b,c)=\max(a,b,c)$ if $a,b,c<0$, and $\mathrm{MM}(a,b,c)=0$ otherwise. The extrapolated values of Φ at the cell boundaries z_j, $j=2,4,\dots,J-2$, are then given by

$$\Phi_j^{\mp}:=\bar\Phi_{j\mp1}^n\pm\frac{1}{2}\Phi_{j\mp1}',\qquad j=2,4,\dots,J-2,$$

and are used to calculate the local speeds of propagation

$$a_j^n:=\max\left\{\rho\left(\boldsymbol{\mathcal{J}}_{\mathbf{f}}(\Phi_j^-)\right),\rho\left(\boldsymbol{\mathcal{J}}_{\mathbf{f}}(\Phi_j^+)\right)\right\},\quad j=2,4,\dots,J-2. \tag{6.4}$$

Of course, it is feasible only for small N to use exact eigenvalues here. However, the analysis of section 3 provides estimates of the eigenvalues that can be used here (see section 6.3). Observe that, for each cell I_j, the solution of (1.4) with the piecewise linear initial data defined by $\bar\Phi_j^n$ and the slope vectors Φ_j' remains smooth for $t_n\le t\le t_{n+1}$ in the subinterval $[z_{j-1}+a_{j-1}^n\Delta t,z_{j+1}-a_{j+1}^n\Delta t]$ for $j=1,\dots,J$. Equipped with the numbers a_j^n and the vectors Φ_j', we next calculate the following vectors, which represent the parts of the cell averages pertaining to the left and right half-cells adjacent to $z=z_j$ that are mapped onto a smooth solution:

$$\Phi_{j,\mathrm{L}}^n:=\bar\Phi_{j-1}^n+\left(\frac{1}{2}-\lambda a_j^n\right)\Phi_{j-1}',\qquad \Phi_{j,\mathrm{R}}^n:=\bar\Phi_{j+1}^n-\left(\frac{1}{2}-\lambda a_j^n\right)\Phi_{j+1}'$$

for $j=2,4,\dots,J-2$. The vectors $\Phi_{j,\mathrm{L}}^n$ and $\Phi_{j,\mathrm{R}}^n$ are used to calculate the flux slope vectors

$$\mathbf{f}'(\Phi_{j,c}^n)=\big(f_1'(\Phi_{j,c}^n),\dots,f_N'(\Phi_{j,c}^n)\big)^{\mathrm{T}},\quad c=\mathrm{L},\mathrm{R},\quad j=2,4,\dots,J-2,$$

whose components are defined by

$$f_i'(\Phi_{2,\mathrm{L}}^n)=f_i'(\Phi_{J-2,\mathrm{L}}^n)=0,\qquad f_i'(\Phi_{2,\mathrm{R}}^n)=f_i'(\Phi_{J-2,\mathrm{R}}^n)=0,$$

and

$$f_i'(\Phi_{j,c}^n) := \mathrm{MM}\Big(\theta\big(f_i(\Phi_{j,c}^n) - f_i(\Phi_{j-2,c}^n)\big), \big(f_i(\Phi_{j+2,c}^n) - f_i(\Phi_{j-2,c}^n)\big)/2, \theta\big(f_i(\Phi_{j+2,c}^n) - f_i(\Phi_{j,c}^n)\big)\Big)$$

for $c = \mathrm{L}, \mathrm{R}$, $i = 1, \ldots, N$, and $j = 4, 6, \ldots, J-4$. We then calculate the predictor solution values

$$\Phi_{j,c}^{n+1/2} := \Phi_{j,c}^n - \frac{\lambda}{2}\mathbf{f}'(\Phi_{j,c}^n), \quad c = \mathrm{L}, \mathrm{R}, \quad j = 2, 4, \ldots, J-2,$$

at which the flux vector $\mathbf{f}$ is evaluated in order to calculate the new approximate values $\bar{\Psi}_j^{n+1}$, $j = 2, 3, \ldots, J-1, J$, of the solution at time t_{n+1}, which are referred to a nonuniform grid as follows. For $j = 2, 4, \ldots, J-2$, approximate cell averages $\bar{\Psi}_j^{n+1}$ referring to the intervals $[z_j - a_j^n \Delta t, z_j + a_j^n \Delta t]$, $j = 2, 4, \ldots, J-2$, are calculated by

$$\bar{\Psi}_j^{n+1} = \frac{1}{2}\big(\bar{\Phi}_{j-1}^n + \bar{\Phi}_{j+1}^n\big) + \frac{1 - \lambda a_j^n}{4}\big(\Phi_{j-1}' - \Phi_{j+1}'\big) - \frac{1}{2a_j^n}\Big[\mathbf{f}\big(\Phi_{j,\mathrm{R}}^{n+1/2}\big) - \mathbf{f}\big(\Phi_{j,\mathrm{L}}^{n+1/2}\big)\Big],$$

while the second family of approximate cell averages $\bar{\Psi}_j^{n+1}$ refers to the nonuniform cells $[z_{j-1} + a_{j-1}^n \Delta, z_{j+1} - a_{j+1}^n \Delta t] \subset I_j$, $j = 3, 5, \ldots, J-3$, and is calculated by

$$\bar{\Psi}_j^{n+1} = \bar{\Phi}_j^n - \frac{\lambda}{2}\big(a_{j+1}^n - a_{j-1}^n\big)\Phi_j' - \frac{\lambda}{1 - \lambda(a_{j-1}^n + a_{j+1}^n)}\Big[\mathbf{f}\big(\Phi_{j+1,\mathrm{L}}^{n+1/2}\big) - \mathbf{f}\big(\Phi_{j-1,\mathrm{R}}^{n+1/2}\big)\Big].$$

Using both families of nonuniform approximate cell averages, we determine the vector of discrete derivatives $\Psi_j' = (\Psi_{1,j}', \ldots, \Psi_{N,j}')^{\mathrm{T}}$ for $j = 2, 4, \ldots, J-2$, setting $\Psi_2' = \Psi_{J-2}' = 0$ and

$$\Psi_{i,j}' = \frac{1}{\Delta z}\mathrm{MM}\left(\theta\frac{\bar{\Psi}_{i,j}^{n+1} - \bar{\Psi}_{i,j-1}^{n+1}}{1 + \lambda(a_j^n - a_{j-2}^n)}, \frac{\bar{\Psi}_{i,j+1}^{n+1} - \bar{\Psi}_{i,j-1}^{n+1}}{2 + \lambda(2a_j^n - a_{j-2}^n - a_{j+2}^n)}, \theta\frac{\bar{\Psi}_{i,j+1}^{n+1} - \bar{\Psi}_{i,j}^{n+1}}{1 + \lambda(a_j^n - a_{j+2}^n)}\right)$$

for $i = 1, \ldots, N$ and $j = 4, 6, \ldots, J-4$, where $\theta \in [0, 2]$ is a parameter. Finally, we can calculate the desired numerical flux vectors

$$\mathbf{h}_j^n = \frac{1}{2}\Big[\mathbf{f}\big(\Phi_{j,\mathrm{R}}^{n+1/2}\big) + \mathbf{f}\big(\Phi_{j,\mathrm{L}}^{n+1/2}\big)\Big] - \frac{a_j^n}{2}\big(\bar{\Phi}_{j+1}^n - \bar{\Phi}_{j-1}^n\big) + \frac{a_j^n(1 - \lambda a_j^n)}{4}\big(\Phi_{j-1}' + \Phi_{j+1}'\big) + \lambda\Delta z\big(a_j^n\big)^2\Psi_j', \quad j = 2, 4, \ldots, J-2.$$

For the diffusion part, we approximate $\partial\Phi/\partial z(z_j, t_n)$ by the slope vector

$$\widetilde{\Phi}_j' = (\tilde{\phi}_{1,j}', \ldots, \tilde{\phi}_{N,j}')^{\mathrm{T}}$$

defined by $(\bar{\phi}_{i,j+1}^n - \bar{\phi}_{i,j-1}^n)/(2\Delta z)$ for $j = 2, 4, \ldots, J-2$ and $i = 1, \ldots, N$. Using the diffusion vector $\mathbf{a}(\Phi, \partial\Phi/\partial z)$ given by (4.2), we can calculate the numerical diffusion vectors by

$$\mathbf{p}_j^n = \frac{1}{2}\Big[\mathbf{a}\big(\bar{\Phi}_{j-1}, \widetilde{\Phi}_j'\big) + \mathbf{a}\big(\bar{\Phi}_{j+1}, \widetilde{\Phi}_j'\big)\Big], \quad j = 2, 4, \ldots, J-2. \tag{6.5}$$

6.3. Application to the model of polydisperse sedimentation with compression. In the numerical examples, we consider the standard Richardson and Zaki [86] hindered settling factor

$$V(\phi) = \begin{cases} 0 & \text{for } \phi \le 0 \text{ and } \phi \ge \phi_{\max}, \\ (1-\phi)^{n-2},\ n > 2, & \text{otherwise} \end{cases} \tag{6.6}$$

and the widely used power-law effective solid-stress formula

$$\sigma_{\mathrm{e}}(\phi) = \begin{cases} 0 & \text{for } \phi \le \phi_{\mathrm{c}}, \\ \sigma_0\big((\phi/\phi_{\mathrm{c}})^k - 1\big) & \text{for } \phi > \phi_{\mathrm{c}}, \end{cases} \quad \text{i.e.,} \quad \sigma_{\mathrm{e}}'(\phi) = \begin{cases} 0 & \text{for } \phi < \phi_{\mathrm{c}}, \\ (\sigma_0/\phi_{\mathrm{c}}^k) k \phi^{k-1} & \text{for } \phi > \phi_{\mathrm{c}}, \end{cases} \tag{6.7}$$

with parameters $\sigma_0 > 0$ and $k \ge 1$. Values of σ_0, ϕ_{c}, and k for real materials are given in [16, 105].

For our choice of $V(\phi)$ and under the mild assumption $\phi_{\max} > 1/n$, we may significantly sharpen the upper bound for the eigenvalues of $\boldsymbol{\mathcal{J}}_{\mathbf{f}^{\mathrm{M}}}(\Phi)$ compared with the bound given by Theorem 3.4.

LEMMA 6.1. *For the hindered settling function* (6.6) *and* $\Phi \in \mathcal{D}^0_{1/n}$, *i.e.,* $\phi < 1/n < \phi_{\max}$, *the eigenvalues* $\nu_1(\Phi)$ *to* $\nu_N(\Phi)$ *of* $\boldsymbol{\mathcal{J}}_{\mathbf{f}^{\mathrm{M}}}(\Phi)$ *satisfy* (3.15) *and*

$$\nu_N(\Phi) \in \big(\mu\bar{\varrho}_{\mathrm{s}} V(\phi)(1-\phi)(\delta_N - \boldsymbol{\delta}^{\mathrm{T}}\Phi), -\mu\bar{\varrho}_{\mathrm{s}} V(\phi)(1-\phi)\boldsymbol{\delta}^{\mathrm{T}}\Phi\big). \tag{6.8}$$

Proof. We set $\widetilde{\gamma}^{\infty} := -V(\phi)(1-\phi)\boldsymbol{\delta}^{\mathrm{T}}\Phi$ such that $\delta(\widetilde{\gamma}^{\infty}) = 0$. Using (3.11), we get

$$\begin{aligned} P(\widetilde{\gamma}^{\infty}) &= \left\{ V(\phi)(1-\phi) + \sum_{m=1}^{N} \frac{\phi_m}{\delta_m} \Big[-\delta_m V(\phi)(1-\phi) + \big(V(\phi)(1-\phi)\big)'(1-\phi)\delta_m \Big] \right\} \\ &\qquad \times \big(V(\phi)(1-\phi)\big)^{N-1} \delta_1 \cdots \delta_N \\ &= \Big\{ V(\phi)(1-\phi) - \phi V(\phi)(1-\phi) + \big(V(\phi)(1-\phi)\big)'\phi(1-\phi) \Big\} \\ &\qquad \times \big(V(\phi)(1-\phi)\big)^{N-1} \delta_1 \cdots \delta_N \\ &= \Big\{ V(\phi)(1-\phi) + \big(V(\phi)(1-\phi)\big)'\phi \Big\} (1-\phi)^N \big(V(\phi)\big)^{N-1} \delta_1 \cdots \delta_N. \end{aligned}$$

For $V(\phi) = (1-\phi)^{n-2}$, the expression in curled brackets is given by

$$\begin{aligned} (1-\phi)^{n-1} - (n-1)(1-\phi)^{n-2}\phi &= (1-\phi)^{n-2}\big((1-\phi) - (n-1)\phi\big) \\ &= (1-\phi)^{n-2}(1-n\phi), \end{aligned}$$

which is positive if and only if $\phi < 1/n$. In this case we thus have $P(\widetilde{\gamma}^{\infty}) > 0$. Since $\widetilde{\gamma}^{\infty} < \gamma^N$, $P(\gamma^N) < 0$, and we have now shown that $P(\widetilde{\gamma}^{\infty}) > 0$, the smallest eigenvalue λ_N of $\mathbf{J}$ satisfies $\widetilde{\gamma}^{\infty} < \lambda_N < \gamma^N$, which implies the statement of the lemma. ∎

Wherever $\phi < 1/n$, Lemma 6.1 can be used to estimate the local speeds of propagation (6.4) in the numerical method since, for our choice of $V(\phi)$, we then have that $\rho(\boldsymbol{\mathcal{J}}_{\mathbf{f}}(\Phi)) \le \mu\bar{\varrho}_{\mathrm{s}}(1-\phi)^{n-1} \max\{\boldsymbol{\delta}^{\mathrm{T}}\Phi, 1 - \boldsymbol{\delta}^{\mathrm{T}}\Phi\}$. Similarly, the eigenvalue bounds of Theorem 4.3 can be utilized to estimate the term $\max \rho(\boldsymbol{\mathcal{J}}_{\mathbf{f}})$ required in the CFL condition (6.3), which limits the step size ratio.

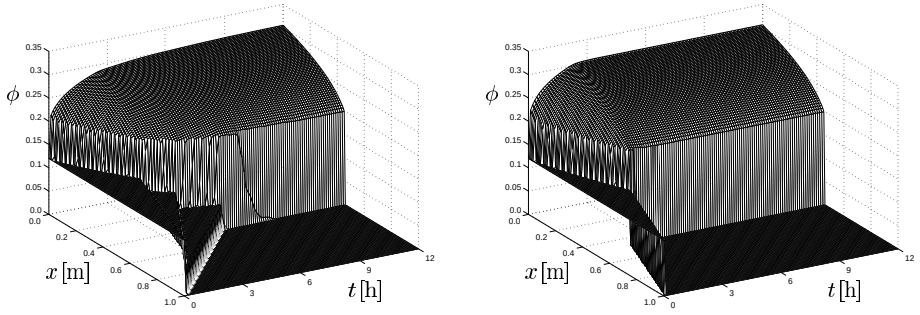


FIG. 1. *Simulations of the sedimentation-consolidation process of a tridisperse suspension (left) and a monodisperse suspension (right) using the same model functions $V(\phi)$ and $\sigma_e(\phi)$: plots of the (total) solids concentration ϕ.*

6.4. Numerical results. The numerical scheme is now employed to simulate settling processes of a tridisperse ($N = 3$) suspension forming compressible sediment. We consider here a (hypothetical) mixture described by the model functions (6.6) with $\phi_{\max} = 0.66$ and $n = 4.7$ (see [94]) and (6.7) with $\sigma_0 = 180\,\mathrm{Pa}$, $\phi_c = 0.2$, and $k = 6$. The remaining parameters are $\mu_f = 10^{-3}\mathrm{Pa{\cdot}s}$ (the dynamic viscosity of water), $d_1 = 1.19 \times 10^{-5}\,\mathrm{m}$, $\bar{\varrho}_s = 1800\,\mathrm{kg/m^3}$, and $g = 9.81\,\mathrm{m/s^2}$.

6.4.1. Settling of a tridisperse suspension. We consider an initially homogeneous suspension with $d_2/d_1 = \sqrt{0.5}$ and $d_3/d_1 = 0.5$, such that $\boldsymbol{\delta} = (1, 0.5, 0.25)^{\mathrm{T}}$, and $\Phi^0 = (0.04, 0.04, 0.04)^{\mathrm{T}}$ in a vessel of height $L = 1\,\mathrm{m}$. For the simulation, we chose $J = 1000$ and $\lambda = 0.0008\,\mathrm{h/m}$. The left diagram of Figure 1 shows the total volumetric solids concentration $\phi = \phi_1 + \phi_2 + \phi_3$ as a function of z and t, while Figure 2 displays the corresponding concentrations of the individual species.

To make the numerical results comparable to those obtained from the two existing models for monodisperse flocculated suspensions and for polydisperse suspensions of rigid spheres, we show in the right diagram of Figure 1 a simulation of the settling of a monodisperse suspension with $\phi_0 = 0.12$, and in Figure 3 the simulation of a tridisperse suspension of rigid particles (forming a sediment without compressibility effects) having the same parameters as the previously discussed case but with $\sigma_0 = 0$. The simulation shown in Figure 3 was made with $\lambda = 0.35\,\mathrm{h/m}$ and $J = 8000$. Note that the visual grid used in all diagrams is much coarser than the computational grid.

6.4.2. Effect of a third particle species on the settling of a bidisperse suspension. To study the effect of the size of a third species on the separation of two other species, we first consider a bidisperse suspension having the parameters given above and $\boldsymbol{\delta}^{\mathrm{T}} = (1.0, 0.5)$. The initial concentration is $\Phi^0 = (0.06, 0.06)^{\mathrm{T}}$. Other parameters for this simulation (and that of Figure 5) are $\lambda = 0.0008\,\mathrm{s/m}$, $J = 600$, and $L = 1\,\mathrm{m}$. Figure 4 shows a simulation of the settling of this suspension.

Next, we add a third species to this bidisperse mixture. The corresponding numerical results are shown in Figure 5. The left and right columns correspond to the size parameters $\delta_3 = 0.25$ and $\delta_3 = 0.1$, respectively. The initial concentrations of the tridisperse mixture are $\Phi^0 = (0.06, 0.06, 0.015)^{\mathrm{T}}$.

6.5. Discussion of the numerical results. In the left plot of Figure 1, three distinct zones are formed by the downwards-propagating concentration discontinuities, and, as expected, the concentration ϕ in the sediment bed increases more slowly than

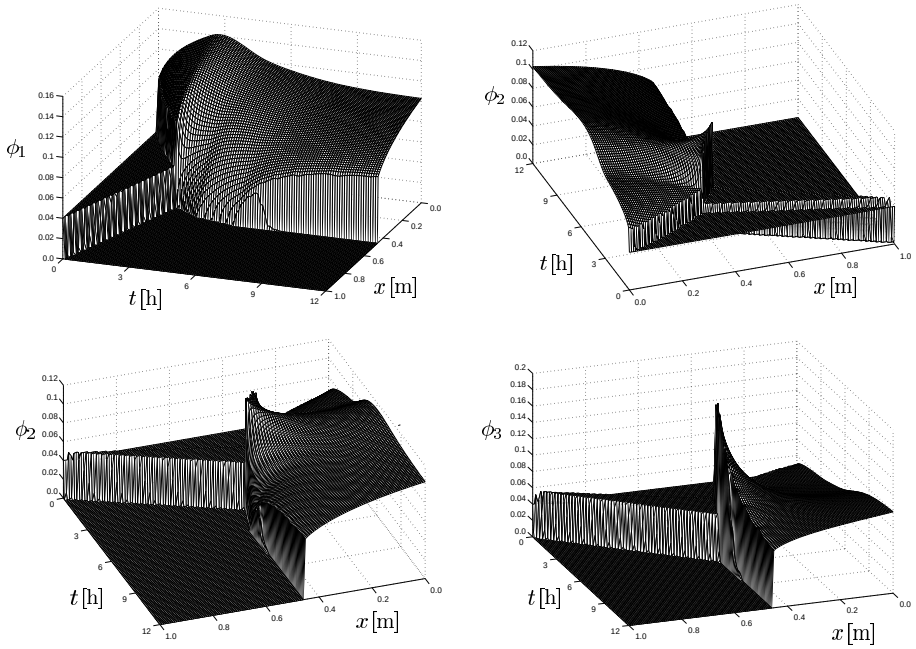


FIG. 2. *Simulation of the sedimentation-consolidation process of a tridisperse suspension: plots of the concentrations ϕ_1 of the largest (top left), ϕ_2 of the second-largest (top right and bottom left; two different views), and ϕ_3 of the smallest species (bottom right).*

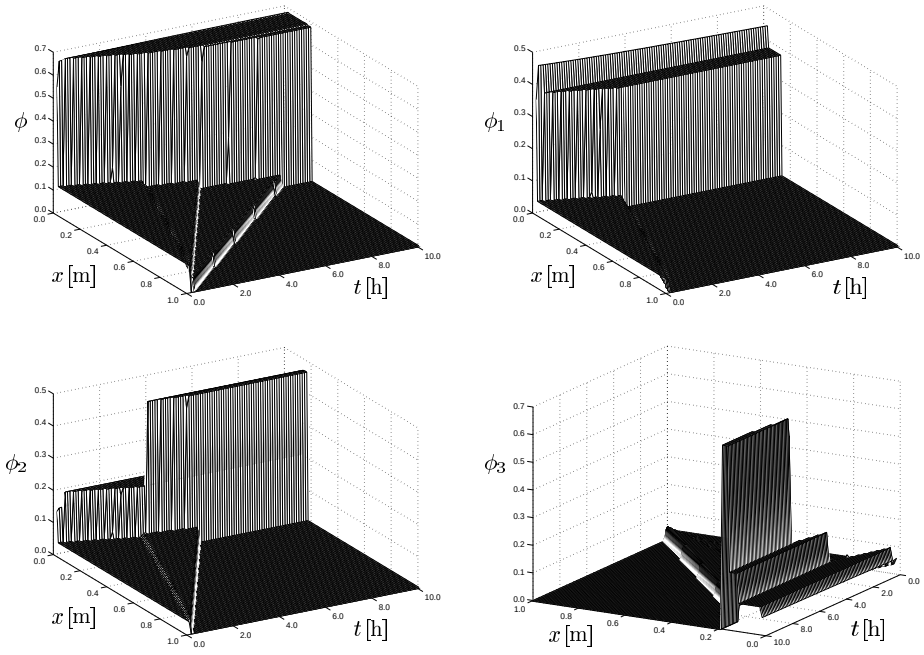


FIG. 3. *Simulation of the sedimentation of a tridisperse suspension of rigid particles (without compression, $\sigma_e \equiv 0$): plots of the cumulative concentration ϕ (top left) and the concentrations ϕ_1, ϕ_2, and ϕ_3 of the largest (top right), the second-largest (bottom left), and smallest species (bottom right).*

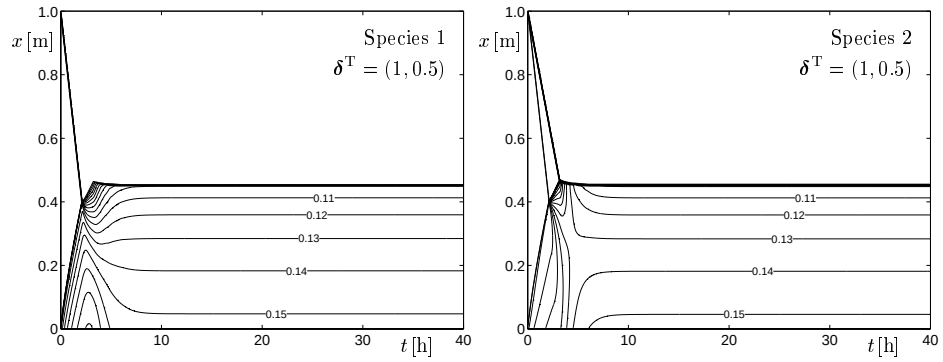


FIG. 4. *Simulation of the settling of an initially homogenous bidisperse suspension: iso-line of the concentrations ϕ_1 of the larger (left) and ϕ_2 of the smaller (right) species, corresponding to $\phi_{1,2} = 0, 0.01, 0.02, 0.03, \ldots$.*

in the monodisperse case. Comparing ϕ in the two tridisperse cases, we see that the zones formed in the first stages of sedimentation are still visible in the upper left plot of Figure 3, but have been entirely smoothed out in the left plot of Figure 1.

The two bottom plots of Figure 3 show the expected layering caused by differential sedimentation and the consequent enhancement of ϕ_2 and ϕ_3 above the lowest zone. Additional numerical examples illustrating the conventional model of sedimentation of suspensions of rigid spheres (when $\sigma_e = 0$) are given in [14, 19] (see also [48]). Figure 2 shows that the additional terms in the equation for suspensions forming compressible sediments result in the upward diffusion of the largest spheres and the downward diffusion of the smallest. Though these terms were expected to smooth the sharp boundaries found in suspensions of incompressible particles, the extent of the migration was unexpected.

The simulations described in (6.4.2) elucidate this phenomenon. We first simulated the sedimentation of an initially homogeneous bidisperse suspension and plotted the isolines of concentration. Figure 4 shows that these isolines ultimately have the same value for both species. This is a consequence of the assumption that $(\phi_i/\phi)\sigma_e(\phi)$ is the part of σ_e that acts on species i. For particles of equal density, and if we assume $V(\phi) > 0$, then the one-dimensional equilibrium form of (2.20) is

$$\bar{\varrho}_s(1-\phi) + \frac{\sigma_e(\phi)}{g\phi_i}\frac{d}{dz}\left(\frac{\phi_i}{\phi}\right) + \frac{1-\phi}{g\phi}\frac{d\sigma_e(\phi)}{dz} = 0, \quad i = 1, \ldots, N, \tag{6.9}$$

which can be rearranged to

$$\frac{d}{dz}\ln\left(\frac{\phi_i}{\phi}\right) = -\frac{1-\phi}{\sigma_e(\phi)}\left(\bar{\varrho}_s g\phi + \frac{d\sigma_e(\phi)}{dz}\right), \quad i = 1, \ldots, N. \tag{6.10}$$

From (3.53) of [27] with $u = 0$, or by setting $\phi_i = \phi$ in (6.9), we see that the expression in large parentheses is zero. Thus, ϕ_i/ϕ is constant and we have

$$\lim_{t\to\infty}\frac{\phi_i(z,t)}{\phi(z,t)} = \frac{\phi_i^0}{\phi_1^0 + \cdots + \phi_N^0}, \quad i = 1, \ldots, N. \tag{6.11}$$

The same phenomenon is also clear in Figure 5. Here the isolines of species 1 and 2 have the same ultimate values, while those of species 3 are proportional to its initial concentration.

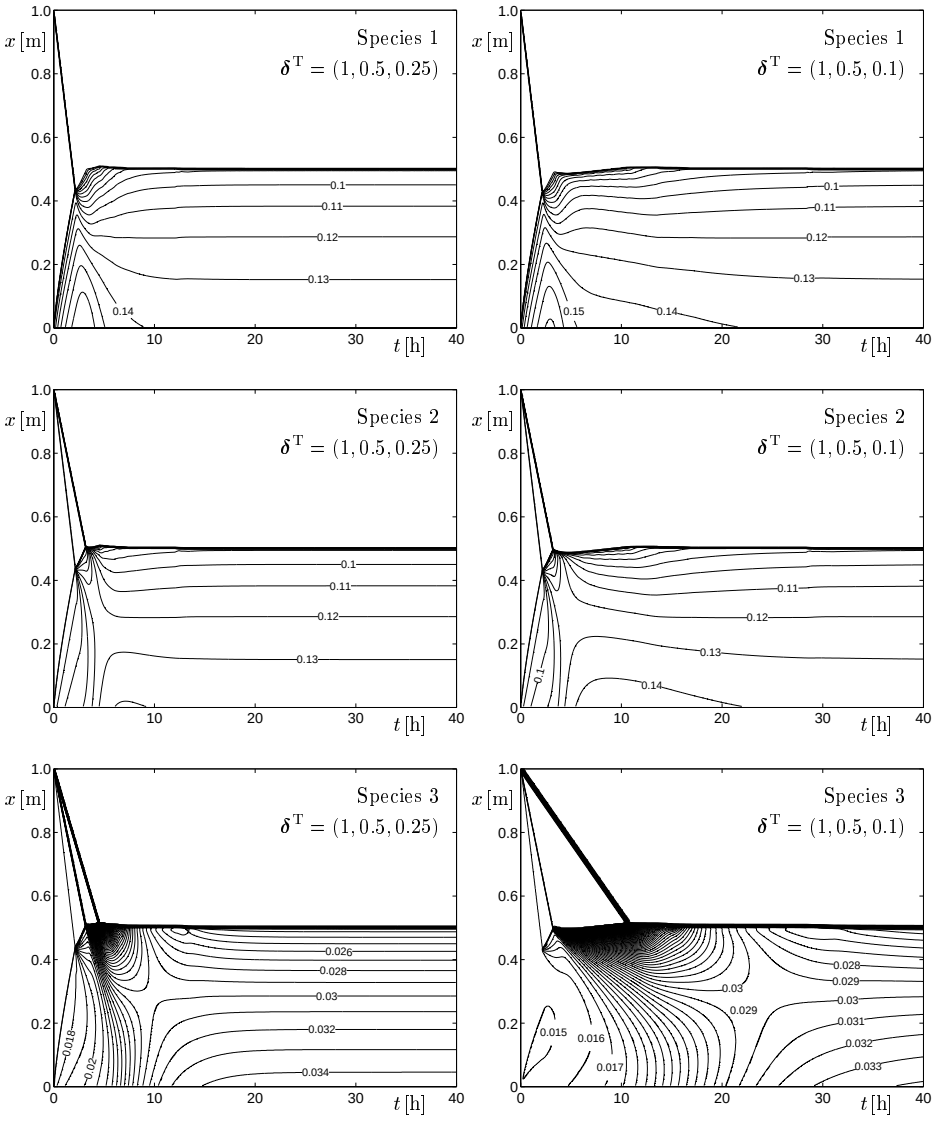


FIG. 5. *Simulations of the settling of tridisperse suspensions with different sizes of species* 3 *($d_3 = 0.5d_1$ in the left and $d_3 = 0.0.316d_1$ in the right column): iso-lines of the concentrations ϕ_1 of the larger (top) and ϕ_2 of the medium-sized (middle) species, corresponding to $\phi_{1,2} = 0, 0.01, 0.02, 0.03, \ldots$, and ϕ_3 of the smallest species (bottom) for $\phi_3 = 0, 0.001, 0.002, \ldots$.*

The complicated structures of the isolines at lower values of t arise from the resolution of the disparity between the segregation that occurs early in the sedimentation process and the ultimate uniformity with respect to species. The details of the process depend sensitively on the values of the terms in a_i. However, certain features are common.

We first consider a bidisperse suspension. When $\phi < \phi_{\mathrm{c}}$, the largest species settles most quickly and predominates in the lower region. In the consolidation phase ($\phi > \phi_{\mathrm{c}}$), the increase in ϕ tends to increase the concentration of both species in

the lower region. Figure 4 shows that species 1 reaches a concentration of 0.18 at the bottom while species 2 is still settling into the top (monodisperse) layer of the solids in compression. However, the larger particles diffuse into this layer, and the smaller particles diffuse out of it. This diffusion continues until the equilibrium state is reached.

In the tridisperse case shown in Figure 5, species 1 diffuses upward while species 2 diffuses both upward into the initially monodisperse upper layer of small particles and downward into the lower layer where large particles initially predominate. Species 3 diffuses downward from the top layer. In addition to reducing the final concentrations of the two larger species, the introduction of the smallest particles delays the evolution to the equilibrium state by introducing a segregated layer at the top of the suspension. In the example on the left, species 3 settles fairly quickly, and the change from segregated to uniform state occurs much earlier than in the example on the right, where species 3 settles very slowly. Further discussion of the phenomenon of sediment diffusivity seen in our simulations is provided in section 7.5.

7. Discussion.

7.1. Type analysis in several space dimensions. The type analysis confirms that the model is well-posed in that the one-dimensional system (1.2) is not of "general" type but has desirable algebraic properties. The analysis in sections 3 and 4 has been limited to one space dimension for notational convenience and since only in that case does the system (2.29), supplemented by the initial and boundary conditions (2.30) and (2.31), completely describe the sedimentation-consolidation process. In $D > 1$ space dimensions, not only system (2.23) for the concentrations of the solids species but also (2.24) and (2.25) for the motion of the mixture have to be solved. These equations are strongly coupled and probably will have to be solved alternately. Although, for $D > 1$, (2.23) no longer completely describes the sedimentation-consolidation process, this multidimensional system still is strongly degenerate parabolic-hyperbolic. To see this, consider first the case $\phi \le \phi_c$, for which the right-hand side of (2.23) vanishes. On the other hand, we recall that a D-dimensional $N \times N$ system of conservation laws

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \boldsymbol{\varphi}_1(\mathbf{u})}{\partial x_1} + \cdots + \frac{\partial \boldsymbol{\varphi}_D(\mathbf{u})}{\partial x_D} = 0,$$

with $\mathbf{u} \in \mathcal{D} \subset \mathbb{R}^N$ and flux vectors $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_D : \mathcal{D} \to \mathbb{R}^N$, is called hyperbolic if any linear combination $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}, \mathbf{u}) := \beta_1 \boldsymbol{\mathcal{J}}_{\boldsymbol{\varphi}_1}(\mathbf{u}) + \cdots + \beta_D \boldsymbol{\mathcal{J}}_{\boldsymbol{\varphi}_D}(\mathbf{u})$ of the Jacobians of the flux vectors is diagonalizable with real eigenvalues. The nonlinear fluxes $f_1^{\mathrm{M}}(\Phi), \ldots, f_N^{\mathrm{M}}(\Phi)$ in (2.23) are effective in the vertical direction of the z coordinate only. Considering $D = 3$ (the case $D = 2$ is analogous) and $\mathbf{q} = (q_x, q_y, q_z)^{\mathrm{T}}$, we obtain from (2.23) $\boldsymbol{\varphi}_1(\Phi) = q_x \Phi$, $\boldsymbol{\varphi}_2(\Phi) = q_y \Phi$, and $\boldsymbol{\varphi}_3(\Phi) = q_z \Phi + \mathbf{f}^{\mathrm{M}}(\Phi)$. Thus, the relevant linear combinations are $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}; \Phi) := (\beta_1 q_x + \beta_2 q_y + \beta_3 q_z)\mathbf{I} + \beta_3 \boldsymbol{\mathcal{J}}_{\mathbf{f}^{\mathrm{M}}}(\Phi)$, where $\boldsymbol{\mathcal{J}}_{\mathbf{f}^{\mathrm{M}}}(\Phi)$ is the Jacobian considered in section 3. Since $\boldsymbol{\mathcal{J}}_{\mathbf{f}^{\mathrm{M}}}(\Phi)$ has N pairwise-distinct eigenvalues and is thus diagonalizable, $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}; \Phi)$ is also diagonalizable with real eigenvalues, and (2.23) is therefore hyperbolic for $\phi \le \phi_c$. Of course, this statement is true under the same conditions as in the one-dimensional case, that is, for equal-density spheres and vectors $\Phi \in \mathcal{D}^0_{\phi_{\max}}$.

Next, we show that the system (2.23) is parabolic for $\phi > \phi_c$. More precisely, we show that it satisfies the classical definition of parabolicity in the sense of Petrovsky

[42, 49, 66, 104]. We do not state this condition in its most general form but limit the discussion to equations of the form

$$\frac{\partial u_i}{\partial t} + F_i(\mathbf{x}, t, \mathbf{u}, \nabla \mathbf{u}) = \sum_{m,n=1}^{D} \sum_{j=1}^{N} A_{ij}^{mn}(\mathbf{x}, t, \mathbf{u}) \frac{\partial^2 u_j}{\partial x_m \partial x_n}, \quad i = 1, \ldots, N. \tag{7.1}$$

Consider the matrix $\boldsymbol{\mathcal{A}}(\mathbf{x}, t, \mathbf{u})^{mn} := (A_{ij}^{mn})_{1 \le i,j \le N}$. Then (7.1) is called *parabolic in the sense of Petrovsky* (or simply *parabolic*) at a point $(\mathbf{x}, t, \mathbf{u}) \in Q_T \times \mathcal{D} \subset \mathbb{R}^D \times \mathbb{R}^+ \times \mathbb{R}^N$ if, for all vectors $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_D)^{\mathrm{T}}$ with $|\boldsymbol{\xi}| = 1$, the roots $\lambda = \lambda(\mathbf{x}, t, \mathbf{u}, \boldsymbol{\xi})$ of $\det(\boldsymbol{\mathcal{A}}(\mathbf{x}, t, \mathbf{u}, \boldsymbol{\xi}) - \lambda \mathbf{I}) = 0$, where

$$\boldsymbol{\mathcal{A}}(\mathbf{x}, t, \mathbf{u}, \boldsymbol{\xi}) := \sum_{m,n=1}^{D} -\boldsymbol{\mathcal{A}}^{mn}(\mathbf{x}, t, \mathbf{u}) \xi_n \xi_m,$$

satisfy $\mathrm{Re}\,(\lambda(\mathbf{x}, t, \mathbf{u}, \boldsymbol{\xi})) < -\delta(\mathbf{x}, t, \mathbf{u})$ for a constant $\delta > 0$. We now consider the right-hand part of (2.25). From (4.1) and (4.2) we get $\mathbf{a}_i(\Phi, \nabla\Phi) = \eta_{i1}(\Phi)\nabla\phi_1 + \cdots + \eta_{iN}(\Phi)\nabla\phi_N$ and therefore

$$\begin{aligned} \nabla \cdot \mathbf{a}_i(\Phi, \nabla\Phi) &= \sum_{m=1}^{D} \sum_{j=1}^{N} \frac{\partial}{\partial x_m} \left(\eta_{ij}(\Phi) \frac{\partial \phi_j}{\partial x_m} \right) \\ &= \sum_{m=1}^{D} \sum_{j=1}^{N} \eta_{ij}(\Phi) \frac{\partial^2 \phi_j}{\partial x_m^2} + \sum_{m=1}^{D} \sum_{j=1}^{N} \frac{\partial \eta_{ij}(\Phi)}{\partial \phi_j} \left(\frac{\partial \phi_j}{\partial x_m} \right)^2. \end{aligned}$$

Defining

$$F_i(\mathbf{x}, t, \Phi, \nabla\Phi) := \nabla \cdot \left(\phi_i \mathbf{q} + f_i^{\mathrm{M}}(\Phi) \right) - \sum_{m=1}^{D} \sum_{j=1}^{N} \frac{\partial \eta_{ij}(\Phi)}{\partial \phi_j} \left(\frac{\partial\ \phi_j}{\partial x_m} \right)^2, \quad i = 1, \ldots, N,$$

we can rewrite (2.23) in the form (7.1). We then obtain $\boldsymbol{\mathcal{A}}^{mn}(\mathbf{x}, t, \Phi) = \mathbf{A}(\Phi)$ for all $1 \le m, n \le D$ if $m = n$, and $\boldsymbol{\mathcal{A}}^{mn}(\mathbf{x}, t, \Phi) = 0$ otherwise, where $\mathbf{A}(\Phi)$ was introduced in section 4. This implies $\boldsymbol{\mathcal{A}}(\mathbf{x}, t, \Phi, \boldsymbol{\xi}) = -\mathbf{A}(\Phi)$ for all $\boldsymbol{\xi} \in \mathbb{R}^D$ with $|\boldsymbol{\xi}| = 1$. From Theorem 4.3 we see that $\boldsymbol{\mathcal{A}}(\mathbf{x}, t, \Phi, \boldsymbol{\xi})$ has N distinct real eigenvalues $-\Lambda_1, \ldots, -\Lambda_N$ for $\Phi \in \mathcal{D}^0_{\phi_{\max}} \backslash \mathcal{D}_{\phi_c}$. This implies that the parabolicity condition $\mathrm{Re}\,(\lambda(\Phi, \boldsymbol{\xi})) < -\delta(\Phi)$ holds with $\delta(\Phi) = W(\phi)\delta_N \min\{\sigma_e(\phi), \phi(1-\phi)^2\sigma_e'(\phi)\}$. Thus, system (2.23) is parabolic on $\mathcal{D}^0_{\phi_{\max}} \backslash \mathcal{D}_{\phi_c}$, and we conclude that the hyperbolicity and parabolicity properties obtained in sections 3 and 4 remain valid in an arbitrary number of space dimensions.

7.2. Extension to particles with different densities. The model equations established in section 2 admit that the solids species differ in both size and density. The analysis of section 3 is valid for the case of equal-density particles only, while the matrix $\mathbf{A}(\Phi)$ is independent of the particle densities. In [22] it was demonstrated that different densities lead to hyperbolic-elliptic or (for $N \ge 3$) nonhyperbolic systems. Thus, it is tempting to conclude that the model framework of section 2 leads to systems having even more interesting properties (like a second-order parabolic system for $N = 2$ degenerating into a first-order hyperbolic-elliptic one). However, since particles of different densities consist of different materials, the assumption $\sigma_e(\Phi) = \sigma_e(\phi)$, stating that the effective stress depends only on the sediment porosity $1 - \phi$, is very unlikely to remain valid.

7.3. Physical explanation of Lemma 6.1. Recall that $f_i^{\mathrm{M}}(\Phi) = \phi_i v_i$, where v_i is the phase velocity of species i, that is, the settling velocity of a particle of species i. In view of (3.1), Theorem 3.4 states that the eigenvalues ν_1 to ν_{N-1} of $\boldsymbol{\mathcal{J}}_{\mathbf{f}^{\mathrm{M}}}(\Phi)$ satisfy $v_i \le \nu_i \le v_{i+1}$; i.e., the propagation of the characteristic information associated with the eigenvalue ν_i is bounded by the physical velocities of particles of species i and $i+1$ for $i = 1, \dots, N-1$. The upper bound on ν_N given by the parameter γ^∞ of Theorem 3.4, which is valid for any admissible hindered settling function $V(\phi)$, has no obvious physical interpretation, but Theorem 3.4 already provides further support for the MLB model wherever $\sigma_{\mathrm{e}} = 0$ since all waves should travel at bounded finite speeds and, for a given particle size distribution, γ^∞ is uniformly bounded with respect to Φ. However, the upper bound of ν_N in (6.8) also has a physical meaning. From (3.1), the total solids flux is

(7.2)

$$f^{\mathrm{M}}(\Phi) := \phi_1 v_1 + \cdots + \phi_N v_N = f_1^{\mathrm{M}}(\Phi) + \cdots + f_N^{\mathrm{M}}(\Phi) = \mu V(\phi)(1-\phi)^2 \bar{\varrho}_{\mathrm{s}} \boldsymbol{\delta}^{\mathrm{T}} \Phi.$$

On the other hand, we recall from the definition of $\mathbf{q}$ that

$$v_{\mathrm{f}} = \frac{1}{1-\phi}\big(q - (\phi_1 v_1 + \cdots + \phi_N v_N)\big) = \frac{q - f^{\mathrm{M}}(\Phi)}{1-\phi},$$

where v_{f} is the fluid phase velocity. Since we consider $q = 0$, we obtain $-\mu\bar{\varrho}_{\mathrm{s}} V(\phi)(1-\phi)\boldsymbol{\delta}^{\mathrm{T}}\Phi = v_{\mathrm{f}}$. Thus Lemma 6.1 states that, for relatively dilute suspensions (when $\phi < 1/n$), all eigenvalues (and therefore wave velocities) are bounded by the local velocities of the solid and fluid "particles." In the examples in section 6, we chose $1/n = 0.2128 > \phi_{\mathrm{c}} = 0.2$, such that the model equations are either hyperbolic with the sharp estimates of Lemma 6.1 holding or parabolic.

7.4. Hydrodynamic diffusion. The MLB model (like all other equations for polydisperse suspensions) assumes that $\mathbf{v}_i(\Phi)$ is the velocity of every particle of the ith species at that concentration. Of course, it has long been recognized that identical spheres at the same concentration can have very different velocities. See [111] and [114] for references to early work on this topic. More recently, Segrè, Herbolzheimer, and Chaikin [95] and Guazzelli and colleagues [78, 79, 83] used advanced technology to follow the paths of individual spheres and thereby determine their velocities.

There are essentially three methods of introducing this variability into a model. Historically, the first was the three-parameter Markov model [107, 112], which used the variance and autocorrelation of velocity as additional parameters. A decade later, a model was developed [55] (see also [37]) that combined the variance and autocorrelation in a coefficient of self-induced hydrodynamic diffusion. Thus, the two models are related, but not identical [106]. In both, the parameters must be determined experimentally or computationally. Velocity fluctuations appear to depend on wall effects [111, 114] and density stratification [75, 111] as well as on both the distant [110, 111] and local values of Φ. Theoretical [111], computational [65], and some experimental studies [111] indicate that the variance increases with the size of the container, while other experimental studies [78] show no increase. Recent work by Segrè, Herbolzheimer, and Chaikin [95] and Mucha et al. [75] has gone some way towards resolving this contradiction.

The variability of the velocities of the smaller spheres is considerably increased by the presence of larger or denser spheres [56, 83]. Since the hydrodynamic diffusion

coefficient varies with Φ and $\nabla\Phi$, the diffusion model becomes very complicated [109] for polydisperse suspensions. The Markov model is more tractable [109], but both models require data that are currently lacking. The final method of introducing variability is to use one of several numerical techniques [19, 56, 65]; these solve a specific case and demand considerable computational effort.

Fortunately, the overall behavior of suspensions is usually determined primarily by the mean velocity [84, 113] and does not require the determination of the trajectories of individual spheres [107]. Simulations show that the principal effect of hydrodynamic diffusion is a blurring of the interfaces [19, 108]. In many cases, however, these remain fairly sharp. Experimentally, interfaces are readily detected and, owing to self-sharpening [37, 69], closely approximate discontinuities. Thus, the omission of hydrodynamic diffusion terms at this stage is justified by practical limitations, theoretical considerations, computational comparisons, and experimental results.

7.5. Sediment diffusivity. In the examples discussed in section 6.5, ϕ increases fairly quickly during the consolidation phase, and hence the diffusion of species is highlighted. We recall that the material parameters chosen for the simulations do not correspond to a real suspension; rather, the parameters of the function $\sigma_e(\phi)$ have been chosen such that the numerical simulation produces some clearly visible, distinct effects within a relatively thick sediment layer. The latter point requires that the suspension be highly compressible and therefore that σ_0 and k be relatively large. Thus, strength and rapidity of the diffusion processes are to some extent a consequence of our deliberate choice of parameters, and these effects may be less pronounced for real materials. In fact, it is not clear whether the predicted behavior actually occurs in real suspensions. The assumption that $(\phi_i/\phi)\sigma_e(\phi)$ is that part of σ_e that acts on species i appears to be the obvious choice. Also, (1.2) describes nonlinear diffusion with drift, so it is not surprising that species diffuse to regions of lower concentration.

We mention that nonlinear diffusion in polydisperse suspensions has been considered by Esipov [43] and is postulated as part of a general "competition" mechanism for multispecies granular mixtures by Braun [13]. However, the terms considered in [43] account for hydrodynamic diffusion, and the consequences of the nonlinearity do not appear, since (apparently, for simplicity) these diffusivities are replaced by constants, and cross-diffusivities (e.g., the dependence of the flux of particle species 1 on the flux of species 2) are ignored, while in [13] the nonlinearity is retained, but cross-diffusivities are equally neglected, and no physical interpretation of the origin of nonlinear diffusion is given. In our case, it is difficult to imagine a physical process that leads to the predicted results discussed in section 6.5. In compression, the particles touch each other and support those above. This would appear to make relative movement difficult.

One way out of this dilemma is restricting the movement of particles at very high concentrations. In fact, it has long been held that, at very high concentrations, the particles are locked in place and all species move at the same velocity. This should certainly be true in compressible suspensions. The problem may be not that the diffusion coefficient is much too high in general (which could be fixed, for example, by an appropriate choice of the model functions $V(\phi)$ and $\sigma_e(\phi)$), but that differential diffusion, driven by the gradient $\nabla(\phi_i/\phi)$ in (2.20), becomes dominant when sedimentation is very slow. This is quite unphysical. The first part of the simulations appears reasonable. It is the differential movement at the end that is not.

One way to amend this would be to adopt an idea of Shih, Gidaspow, and Wasan [98], who utilize an expression for the portion of the effective solid stress *gradient*

for each species [98, eq. (10)] that is equivalent to leaving out the term involving $\nabla(\phi_i/\phi)$ in our approach. Unfortunately, the presentation of their numerical solution of a bidisperse system with $\bar{\varrho}_1 = \bar{\varrho}_2$ and $\delta_2 = 0.1766$ is limited to just one profile [98, Figs. 4–6] taken at a time at which the uppermost particles of neither species have reached the sediment layer, a situation that roughly corresponds to $t = 1.5\,\mathrm{h}$ in our Figure 4. However, their solution is similar to ours at that stage, since Figure 4 of [98] shows a concentrated sediment formed by the larger particles with a small portion (actually, only slightly different from the initial concentration) of the smaller. It should be pointed out, of course, that no steady-state prediction of the relative volume fractions ϕ_i/ϕ such as (6.10), (6.11) exists when there are no terms involving that same quantity.

Another way to solve our dilemma, which would go even a step further, would be to change to a common rate of sedimentation at some concentration ϕ^* with $\phi_\mathrm{c} < \phi^* < \phi_\mathrm{max}$. For values of ϕ with $\phi^* < \phi < \phi_\mathrm{max}$, we could eliminate the term in ϕ_i/ϕ in (2.20) and treat the suspension as if all particles were the same size, probably using the average value of δ_i. The best guess for ϕ^* could be found from the simulations by noting the concentrations at which the differential sedimentation dominates. (A possibly more realistic alternative would be to introduce a collective movement gradually, but this would be much more complicated.) Though this solution may seem arbitrary, it does have empirical support. For compressible suspensions, differential sedimentation occurs at medium concentrations. When the concentration is sufficiently high, even dense particles settle at the same speed as the particles of lower density. Thus, the final sediment shows no evidence of segregation; see Been and Sills [5]. When all flocs have the same density, there is a concentration at which initial floc size is unimportant. Essentially, we have a connected structure that is being compressed.

Some more treatments that less closely refer to a particular mathematical model support the similar idea of "en masse" sedimentation of multispecies suspensions at high concentrations [5, 29, 121]. Zeng and Lowe [121] consider rigid-sphere suspensions (not forming compressible sediments) but postulate the existence of a "critical concentration," in the sense of the quantity ϕ^* (*not* ϕ_c) introduced above, at which change in sedimentation behavior from differential settling (size fractionation) to "en-masse settling of the entire suspension" occurs [121], and they indicate that values of ϕ^* ranging from 0.3 to 0.55 are suitable, depending on the material. Related experimental findings were reported much earlier by Shannon and coworkers [96, 97], who observed that for equal-density spheres (normally distributed in diameters plus a tail of fines), the rise of the packed bed showed that the solids flux remained constant throughout (in contrast to sedimentation of dilute suspensions, in which the flux decreases after the larger particles have settled out).

The previous discussion shows that there is no obvious unique way to reduce the sediment diffusivity seen in our numerical examples. Published experimental information to which the numerical predictions could be compared is scarce (see the references cited in this section and [9, 116]), and a definite solution of the problems discussed here cannot be suggested. Basically, there seem to exist three alternatives.

Our approach is based on a rigorous derivation and establishes a polydisperse sedimentation model that is "well-posed" in the sense that strict parabolicity is about the best property we can expect system (1.2) to have whenever the right-hand part is different from zero (i.e., for $\phi_\mathrm{c} < \phi < \phi_\mathrm{max}$). This property, combined with the hyperbolicity of the first-order system, makes the model amenable to numerical so-

lution and is conserved when we vary the model functions $V(\phi)$ and $\sigma_e(\phi)$ to reduce sediment diffusivity. In the monodisperse case, it turned out that using the expression $V(\phi) = (1-\phi)^{n-2}$ for all ranges of concentration values (as, for simplicity, done here) leads to an overestimation of particle diffusivity in the sediment, and better agreement was obtained by using piecewise definition of $V(\phi)$ or of the resulting flux density function $f^M(\phi)$; see [15, 16]. The emphasis here is on a *gradual* variation of the parameters, which leaves the nature of the model unaltered.

The next step of modification would be "switching off" the term $\nabla(\phi_i/\phi)$ in (2.20) on the interval $[\phi^*, \phi_{\max}]$, where we admit the limiting case $\phi^* = \phi_c$. The mathematical consequences of such a reduction can be derived easily, since in the derivation of section 4, $\sigma_e(\phi)$ and its derivative $\sigma_e'(\phi)$ are formally treated as independent functions. Thus for the parabolicity analysis, deleting $\nabla(\phi_i/\phi)$ in (2.20) corresponds to sending σ_e to zero and leaving the occurrences of $\sigma_e'(\phi)$ unchanged. From (4.1) we see that then $\mathbf{A}(\Phi)$ is a rank-one matrix having $N-1$ eigenvalues that vanish. The system is then no longer strictly parabolic for $\phi \in [\phi^*, \phi_{\max}]$. This case is explicitly excluded in the analysis of certain schemes [68] but is admitted in others [60] and still has the advantage that explicit tracking of the sediment-suspension interface is unnecessary.

The most radical modification would be to change to an "en masse" sedimentation model for $\phi \in [\phi^*, \phi_{\max}]$. In particular, this would imply that $\phi = \phi^*$ denotes an interface across which we change from the system of N convection-diffusion equations (1.2) to one scalar equation, i.e., between two different models. This idea is viable when we a priori do not wish to differentiate between size classes in the sediment. In fact, this is the main idea of the model advanced by Stamatakis and Tien [102]. There is an advantage in computation time when there is a region in which a scalar equation instead of an $N \times N$ system has to be solved, but formulating transition conditions across the model change interface and tracking it during computation may become complicated.

7.6. Applications. Been and Sills [5] measured local changes in particle size distribution due to the relative movement of particles of different sizes under the influence of effective solid stress and at different initial concentrations. Their experiments were performed with estuarine mud, a natural mixture composed of many different materials for which the constitutive model equations are difficult to determine. More precise knowledge of these functions can be expected in chemical engineering applications, where the settling solids are usually formed by a product having more homogeneous material properties. The model outlined herein should thus be useful for simulations in any of the industrial applications cited in section 1. In particular, the model can also be applied to centrifugal configurations and pressure filtration (see [15, 21] for the monodisperse cases) and thereby be employed to simulate the manufacturing and final composition of ceramic materials with functionally graded material properties (see [9, 10, 11]). Comparing our Figures 2 and 3 illustrates that the effective stress is a decisive factor when the variation of sediment composition should be continuous.

Acknowledgments. We thank the referees for valuable comments and suggestions.

REFERENCES

[1] H. Arastoopour, S.C. Lin, and S.A. Weil, *Analysis of vertical pneumatic conveying of solids using multiphase flow models*, AIChE J., 28 (1982), pp. 467–473.

[2] D. Aregba-Driollet, R. Natalini, and S. Tang, *Diffusive Kinetic Explicit Schemes for*

Nonlinear Degenerate Parabolic Systems, preprint, Norwegian University of Science and Technology Conservation Laws Preprint Server, Tondheim, Norway, 2000; available online at http://www.math.ntnu.no/conservation/.

[3] L.G. Austin, C.H. Lee, and F. Concha, *Hindered settling and classification partition curves*, Minerals & Metallurgical Process., 9 (1992), pp. 161–168.

[4] G.K. Batchelor and R.W. Janse van Rensburg, *Structure formation in bidisperse sedimentation*, J. Fluid Mech., 166 (1986), pp. 370–407.

[5] K. Been and G.C. Sills, *Self-weight consolidation of soft soils: An experimental and theoretical study*, Géotechnique, 31 (1981), pp. 519–535.

[6] P. Bénilan and H. Touré, *Sur l'équation générale* $u_t = \varphi(u)_{xx} - \psi(u)_x + v$, C. R. Acad. Sci. Paris Sér. I Math., 299 (1984), pp. 919–922.

[7] P. Bénilan and H. Touré, *Sur l'équation générale* $u_t = a(\cdot, u, \phi(\cdot, u)_x)_x + v$ *dans* L^1. I. *Étude du problème stationnaire*, in Evolution Equations, Proceedings of the International Conference Held at Louisiana State University (Baton Rouge, LA, 1992), G. Ferreyra, G. Ruiz Goldstein, and F. Neubrander, eds., Marcel Dekker, New York, 1995, pp. 35–62.

[8] P. Bénilan and H. Touré, *Sur l'équation générale* $u_t = a(\cdot, u, \phi(\cdot, u)_x)_x + v$ *dans* L^1. II. *Le problème d'évolution*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 727–761.

[9] P.M. Biesheuvel, *Particle segregation during pressure filtration for cast formation*, Chem. Engrg. Sci., 55 (2000), pp. 2595–2606.

[10] P.M. Biesheuvel and H. Verweij, *Calculation of the composition profile of a functionally graded material by centrifugal casting*, J. Amer. Ceram. Soc., 83 (2000), pp. 743–749.

[11] P.M. Biesheuvel, V. Breedveld, A.P. Higler, and H. Verweij, *Graded membrane supports produced by centrifugal casting of a slightly polydisperse suspension*, Chem. Engrg. Sci., 56 (2001), pp. 3517–3525.

[12] F. Bouchut, F.R. Guarguaglini, and R. Natalini, *Diffusive BGK approximations for nonlinear multidimensional parabolic equations*, Indiana Univ. Math. J., 49 (2000), pp. 723–749.

[13] J. Braun, *Segregation of granular media by diffusion and convection*, Phys. Rev. E, 64 (2001), paper 011307.

[14] R. Bürger, F. Concha, K.-K. Fjelde, and K.H. Karlsen, *Numerical simulation of the settling of polydisperse suspensions of spheres*, Powder Technol., 113 (2000), pp. 30–54.

[15] R. Bürger, F. Concha, and K.H. Karlsen, *Phenomenological model of filtration processes:* 1. *Cake formation and expression*, Chem. Engrg. Sci., 56 (2001), pp. 4537–4553.

[16] R. Bürger, F. Concha, and F.M. Tiller, *Applications of the phenomenological theory to several published experimental cases of sedimentation processes*, Chem. Engrg. J., 80 (2000), pp. 105–117.

[17] R. Bürger, S. Evje, and K.H. Karlsen, *On strongly degenerate convection-diffusion problems modeling sedimentation-consolidation processes*, J. Math. Anal. Appl., 247 (2000), pp. 517–556.

[18] R. Bürger, S. Evje, K. H. Karlsen, and K.-A. Lie, *Numerical methods for the simulation of the settling of flocculated suspensions*, Chem. Engrg. J., 80 (2000), pp. 91–104.

[19] R. Bürger, K.-K. Fjelde, K. Höfler, and K.H. Karlsen, *Central difference solutions of the kinematic model of settling of polydisperse suspensions and three-dimensional particle-scale simulations*, J. Engrg. Math., 41 (2001), pp. 167–187.

[20] R. Bürger and K.H. Karlsen, *On some upwind difference schemes for the phenomenological sedimentation-consolidation model*, J. Engrg. Math., 41 (2001), pp. 145–166.

[21] R. Bürger and K.H. Karlsen, *A strongly degenerate convection-diffusion problem modeling centrifugation of flocculated suspensions*, in Hyperbolic Problems: Theory, Numerics, Applications (Proceedings of the Eighth International Conference, Magdeburg, Germany, 2000), Vol I, H. Freistühler and G. Warnecke, eds., Internat. Ser. Numer. Math. 140, Birkhäuser-Verlag, Basel, 2001, pp. 207–216.

[22] R. Bürger, K.H. Karlsen, E.M. Tory, and W.L. Wendland, *Model equations and instability regions for the sedimentation of polydisperse suspensions of spheres*, Z. Angew. Math. Mech., 82 (2002), pp. 699–722.

[23] R. Bürger and M. Kunik, *A critical look at the kinematic-wave theory for sedimentation-consolidation processes in closed vessels*, Math. Methods Appl. Sci., 24 (2001), pp. 1257–1273.

[24] R. Bürger, C. Liu, and W.L. Wendland, *Existence and stability for mathematical models of sedimentation-consolidation processes in several space dimensions*, J. Math. Anal. Appl., 264 (2001), pp. 288–310.

[25] R. Bürger and W. L. Wendland, *Existence, uniqueness, and stability of generalized solu-*

tions of an initial-boundary value problem for a degenerating quasilinear parabolic equation, J. Math. Anal. Appl., 218 (1998), pp. 207–239.

[26] R. Bürger, W.L. Wendland, and F. Concha, *Model equations for gravitational sedimentation-consolidation processes*, Z. Angew. Math. Mech., 80 (2000), pp. 79–92.

[27] M.C. Bustos, F. Concha, R. Bürger, and E.M. Tory, *Sedimentation and Thickening*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[28] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[29] J.C. Chang, B.V. Velakamanni, F.F. Lange, and D.S. Pearson, *Centrifugal consolidation of Al_2O_3 and Al_2O_3/ZrO_2 composite slurries vs. interparticle potentials: Particle packing and mass segregation*, J. Amer. Ceram. Soc., 74 (1991), pp. 2201–2204.

[30] G.-Q. Chen and E. DiBenedetto, *Stability of entropy solutions to the Cauchy problem for a class of nonlinear hyperbolic-parabolic equations*, SIAM J. Math. Anal., 33 (2001), pp. 751–762.

[31] G.-Q. Chen and B. Perthame, *Well-posedness for anisotropic degenerate parabolic-hyperbolic equations*, Ann. Inst. H. Poincaré Anal. Nonlinéaire, 20 (2003), pp. 645–668.

[32] G.-Q. Chen and D. Wang, *The Cauchy problem for the Euler equations for compressible fluids*, in Handbook of Mathematical Fluid Dynamics, Vol. 1, S.J. Friedlander and D. Serre, eds., Elsevier Science, Amsterdam, 2002, pp. 421–543.

[33] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[34] F. Concha and M.C. Bustos, *Theory of sedimentation of flocculated fine particles*, in Flocculation, Sedimentation and Consolidation, B.M. Moudgil and P. Somasundaran, eds., American Institute of Chemical Engineers, New York, 1986, pp. 275–284.

[35] F. Concha, M.C. Bustos, and A. Barrientos, *Phenomenological theory of sedimentation*, in Sedimentation of Small Particles in a Viscous Fluid, E.M. Tory, ed., Computational Mechanics Publications, Southampton, UK, 1996, pp. 51–96.

[36] P. D'Ancona and S. Spagnolo, *The Cauchy problem for weakly parabolic systems*, Math. Ann., 309 (1997), pp. 307–330.

[37] R.H. Davis, *Hydrodynamic diffusion of suspended particles: A symposium*, J. Fluid Mech., 310 (1996), pp. 325–335.

[38] R.H. Davis and H. Gecol, *Hindered settling function with no empirical parameters for polydisperse suspensions*, AIChE J., 40 (1994), pp. 570–575.

[39] R. De Boer, *Theory of Porous Media*, Springer-Verlag, Berlin, 2000.

[40] P. Diplas and A.N. Papanicolaou, *Batch analysis of slurries in zone settling regime*, J. Environ. Engrg., 123 (1999), pp. 659–667.

[41] D.A. Drew and L.A. Segel, *Averaged equations for two-phase flows*, Stud. Appl. Math., 50 (1971), pp. 205–231.

[42] S.D. Eidel'man, *Parabolic Systems*, North-Holland, Amsterdam, London, 1969.

[43] S.E. Esipov, *Coupled Burgers equations: A model of polydispersive sedimentation*, Phys. Rev. E, 52 (1985), pp. 3711–3718.

[44] M.S. Espedal and K.H. Karlsen, *Numerical solution of reservoir flow models based on large time step operator splitting algorithms*, in Filtration in Porous Media and Industrial Applications, Lecture Notes of the C.I.M.E course (Cetraro, Italy, 1998), A. Fasano, ed., Lecture Notes in Math. 1734, Springer-Verlag, Berlin, 2000, pp. 9–77.

[45] S. Evje and K.H. Karlsen, *Monotone difference approximations of BV solutions to degenerate convection-diffusion equations*, SIAM J. Numer. Anal., 37 (2000), pp. 1838–1860.

[46] R. Eymard, T. Gallouët, R. Herbin, and A. Michel, *Convergence of a finite volume scheme for nonlinear degenerate parabolic equations*, Numer. Math., 92 (2002), pp. 41–82.

[47] A. Fitt, *Mixed systems of conservation laws in industrial mathematical modelling*, Surveys Math. Indust., 6 (1996), pp. 21–53.

[48] X. Flotats, *Mathematical modeling of polydisperse suspensions sedimentation*, Hungarian J. Indust. Chem., 23 (1995), pp. 215–221.

[49] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[50] D. Gidaspow, *Multiphase Flow and Fluidization*, Academic Press, San Diego, 1994.

[51] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer-Verlag, New York, 1996.

[52] M.D. Green, M. Eberl, and K.A. Landman, *Compressive yield stress of flocculated suspensions: Determination via experiment*, AIChE J., 42 (1996), pp. 2308–2318.

[53] M. Gurtin, *An Introduction to Continuum Mechanics*, Acadamic Press, San Diego, 1981.
[54] K. Gustavsson, J. Oppelstrup, and J. Eiken, *Consolidation of concentrated suspensions—Shear and irreversible floc structure rearrangement*, Comput. Visual. Sci., 4 (2001), pp. 61–66.
[55] J.M. Ham and G.M. Homsy, *Hindered settling and hydrodynamic dispersion in quiescent sedimenting suspensions*, Int. J. Multiphase Flow, 14 (1988), pp. 533–546.
[56] K. Höfler, *Simulation and Modeling of Mono- and Bidisperse Suspensions*, Doctoral thesis, Institute for Computer Applications 1, University of Stuttgart, Stuttgart, Germany, 2000.
[57] H. Holden, K.H. Karlsen, and N.H. Risebro, *On uniqueness and existence of entropy solutions of weakly coupled systems of nonlinear degenerate parabolic equations*, Electron. J. Differential Equations, No. 46 (2003), pp. 1–31.
[58] F. Hubert, *Global existence for hyperbolic-parabolic systems with large periodic initial data*, Differential Integral Equations, 11 (1998), pp. 69–83.
[59] B.L. Keyfitz, *Multiphase saturation equations, change of type and inaccessible regions*, in Flow in Porous Media, J. Douglas and U. Hornung, eds., Birkhäuser-Verlag, Basel, Boston, Berlin 1993, pp. 103–116.
[60] K.H. Karlsen, K.-A. Lie, J.R. Natvig, H.F. Nordhaug, and H.K. Dahle, *Operator splitting methods for systems of convection-diffusion equations: Nonlinear error mechanisms and correction strategies*, J. Comput. Phys., 173 (2001), pp. 636–663.
[61] K.H. Karlsen and M. Ohlberger, *A note on the uniqueness of entropy solutions of nonlinear degenerate parabolic equations*, J. Math. Anal. Appl., 275 (2002), pp. 439–458.
[62] K.H. Karlsen and N.H. Risebro, *On the uniqueness and stability of entropy solutions of nonlinear degenerate parabolic equations with rough coefficients*, Discrete Contin. Dyn. Syst. Ser. A, 9 (2003), pp. 1081–1104.
[63] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier–Stokes Equations*, Academic Press, Boston, MA, 1989.
[64] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 241–282.
[65] A.J.C. Ladd, *Sedimentation of homogeneous suspensions of non-Brownian spheres*, Phys. Fluids, 9 (1997), pp. 491–499.
[66] O.A. Ladyženskaja, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, AMS Trans. Math. Monogr. 23, AMS, Providence, RI, 1968.
[67] K.A. Landman and L.R. White, *Solid/liquid separation of flocculated suspensions*, Adv. Colloid Interface Sci., 51 (1994), pp. 175–246.
[68] C. Lattanzio and R. Natalini, *Convergence of diffusive BGK approximations for nonlinear strongly parabolic systems*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 341–358.
[69] S. Lee, Y. Jang, C. Choi, and T. Lee, *Combined effect of sedimentation velocity fluctuation and self-sharpening on interface broadening*, Phys. Fluids A, 4 (1992), pp. 2601–2606.
[70] I-S. Liu, *Continuum Mechanics*, Springer-Verlag, Berlin, 2002.
[71] M.J. Lockett and K.S. Bassoon, *Sedimentation of binary particle mixtures*, Powder Technol., 24 (1979), pp. 1–7.
[72] C. Mascia, A. Porretta, and A. Terracina, *Non-homogeneous Dirichlet problems for degenerate parabolic-hyperbolic equations*, Arch. Ration. Mech. Anal., 163 (2002), pp. 87–124.
[73] J.H. Masliyah, *Hindered settling in a multiple-species particle system*, Chem. Engrg. Sci., 34 (1979), pp. 1166–1168.
[74] M. Massoudi, K.R. Rajagopal, J.M. Ekmann, and M.P. Mathur, *Remarks on the modeling of fluidized systems*, AIChE J., 38 (1992), pp. 471–472.
[75] P.J. Mucha, S.-Y. Tee, D.A. Weitz, B.I. Shraiman, and M.P. Brenner, *A unifying theory for velocity fluctuations in sedimentation*, J. Fluid Mech., submitted.
[76] K. Nakamura and C.E. Capes, *Vertical pneumatic conveying of binary particle mixtures*, in Fluidization Technology, Vol. 2, D.L. Keairns, ed., Hemisphere Publishing, Washington, DC, 1976, pp. 159–184.
[77] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.
[78] H. Nicolai and E. Guazzelli, *Effect of vessel size on the hydrodynamic diffusion of sedimenting spheres*, Phys. Fluids, 7 (1995), pp. 3–5.
[79] H. Nicolai, B. Herzhaft, E.J. Hinch, L. Oger, and E. Guazzelli, *Particle velocity fluctuations and hydrodynamic self-diffusion of sedimenting non-Brownian spheres*, Phys. Fluids, 8 (1995), pp. 12–23.
[80] M. Ohlberger, *A posteriori error estimates for vertex centered finite volume approxima-*

tions of convection-diffusion-reaction equations, Math. Model. Numer. Anal., 35 (2001), pp. 355–387.

[81] V. Pane and R.L. Schiffman, *A note on sedimentation and consolidation*, Géotechnique, 35 (1985), pp. 69–72.

[82] S.L. Passman and D.A. Drew, *A simple multicomponent fluid theory with accurate physics*, in Recent Developments in Structured Continua, D. De Kee and P.N. Kaloni, eds., Pitman Res. Notes Math. 229, Harlow, UK, 1990, pp. 293–300.

[83] Y. Peysson and E. Guazzelli, *Velocity fluctuations in a bidisperse sedimenting suspension*, Phys. Fluids, 11 (1999), pp. 1953–1955.

[84] D.K. Pickard and E.M. Tory, *Experimental implications of a Markov model for sedimentation*, J. Math. Anal. Appl., 72 (1979), pp. 150–176.

[85] M. Renardy, *A degenerate parabolic-hyperbolic system modeling the spreading of surfactants*, SIAM J. Math. Anal., 28 (1997), pp. 1048–1063.

[86] J.F. Richardson and W.N. Zaki, *Sedimentation and fluidization: Part* I, Trans. Inst. Chem. Engrs. (London), 32 (1954), pp. 35–53.

[87] F. Rosso and G. Sona, *Gravity-driven separation of oil-water dispersions*, Adv. Math. Sci. Appl., 11 (2001), pp. 127–151.

[88] A. Rushton, A.S. Ward, and R.G. Holdich, *Solid-Liquid Filtration and Sedimentation Technology*, 2nd ed., Wiley-VCH, Weinheim, Germany, 2000.

[89] W.K. Sartory, *Three-component analysis of blood sedimentation by the method of characteristics*, Math. Biosci., 33 (1977), pp. 145–165.

[90] U. Schaflinger, *Enhanced centrifugal separation with finite Rossby numbers in cylinders with compartment-walls*, Chem. Engrg. Sci., 42 (1987), pp. 1197–1205.

[91] U. Schaflinger, A. Köppl, and G. Filipczak, *Sedimentation in cylindrical centrifuges with compartments*, Ingenieur-Archiv, 56 (1986), pp. 321–331.

[92] R.L. Schiffman, V. Pane, and R.E. Gibson, *The theory of one-dimensional consolidation of saturated clays:* IV. *An overview of nonlinear finite strain sedimentation and consolidation*, in Sedimentation-Consolidation Models: Predictions and Validations, R.N. Yong and F.C. Townsend, eds., American Society of Civil Engineers, New York, 1984, pp. 1–29.

[93] W. Schneider, *Kinematic-wave theory of sedimentation beneath inclined walls*, J. Fluid Mech., 120 (1982), pp. 323–346.

[94] W. Schneider, G. Anestis, and U. Schaflinger, *Sediment composition due to settling of particles of different sizes*, Int. J. Multiphase Flow, 11 (1985), pp. 419–423.

[95] P.N. Segrè, E. Herbolzheimer, and P.M. Chaikin, *Long-range correlations in sedimentation*, Phys. Rev. Lett., 79 (1997), pp. 2574–2577.

[96] P.T. Shannon, R.D. DeHaas, E.P. Stroupe, and E.M. Tory, *Batch and continuous thickening*, Industrial and Engineering Chemistry Fundamentals, 3 (1964), pp. 250–260.

[97] P.T. Shannon, E.P. Stroupe, and E.M. Tory, *Batch and continuous thickening*, Industrial and Engineering Chemistry Fundamentals, 2 (1963), pp. 203–211.

[98] Y.T. Shih, D. Gidaspow, and D.T. Wasan, *Hydrodynamics of sedimentation of multisized particles*, Powder Technol., 50 (1987), pp. 201–215.

[99] G.F. Smith, *On isotropic functions of symmetric tensors, skew-symmetric tensors and vectors*, Int. J. Engrg. Sci., 9 (1971), pp. 899–916.

[100] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, New York, 1994.

[101] A. Spannenberg and K.P. Galvin, *Continuous differential sedimentation of a binary suspension*, Chem. Engrg. in Australia, 21 (1996), pp. 7–11.

[102] K. Stamatakis and C. Tien, *Batch sedimentation calculations—The effect of compressible sediment*, Powder Technol., 72 (1992), pp. 227–240.

[103] E. Tadmor, *Approximate solutions of nonlinear conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, B. Cockburn, C. Johnson, C.-W. Shu, and E. Tadmor, eds., Lecture Notes in Math. 1697, Springer-Verlag, Berlin, 1998, pp. 1–149.

[104] M.E. Taylor, *Partial Differential Equations* III, Springer-Verlag, New York, 1996.

[105] F.M. Tiller, R. Lu, J.H. Kwon, and D.J. Lee, *Variable flow rate in compactible filter cakes*, Water Res., 33 (1999), pp. 15–22.

[106] E.M. Tory, *Stochastic sedimentation and hydrodynamic diffusion*, Chem. Engrg. J., 80 (2000), pp. 81–89.

[107] E.M. Tory, M. Bargieł, and R.L. Honeycutt, *A three-parameter Markov model for sedimentation* III. *A stochastic Runge–Kutta method for computing first-passage times*, Powder Technol., 80 (1994), pp. 133–146.

[108] E.M. Tory and R.A. Ford, *Simulation of sedimentation of bidisperse suspensions*, Int. J.

Mineral Process., to appear.

[109] E.M. Tory and R.A. Ford, *Simulation of sedimentation of monodisperse and polydisperse suspensions*, in Analysis and Simulation of Multifield Problems, W. Wendland and M. Efendiev, eds., Springer-Verlag, Berlin, 2003, pp. 343–348.

[110] E.M. Tory and M.T. Kamel, *On the divergence problem in calculating particle velocities in dilute suspensions of identical spheres* II. *Effect of a plane wall*, Powder Technol., 55 (1988), pp. 51–59; *Erratum*, Powder Technol., 94 (1997), p. 265.

[111] E.M. Tory, M.T. Kamel, and C.F. Chan Man Fong, *Sedimentation is container-size dependent*, Powder Technol., 73 (1992), pp. 219–238.

[112] E.M. Tory and D.K. Pickard, *A three-parameter Markov model for sedimentation*, Can. J. Chem. Engrg., 55 (1977), pp. 655–665.

[113] E.M. Tory and D.K. Pickard, *Extensions and refinements of a Markov model for sedimentation*, J. Math. Anal. Appl., 86 (1982), pp. 442–470.

[114] E.M. Tory and D.K. Pickard, *Experimental evidence for a stochastic approach to sedimentation*, in Flocculation, Sedimentation and Consolidation, B.M. Moudgil and P. Somasundaran, eds., American Institute of Chemical Engineers, New York, 1986, pp. 297–306.

[115] C.-C. Wang, *A new representation theorem for isotropic functions: An answer to Professor G.F. Smith's criticism of my papers on representations for isotropic functions*. I. *Scalar-valued isotropic functions*, Arch. Ration. Mech. Anal., 36 (1970), pp. 166–197.

[115A] C.-C. Wang, *A new representation theorem for isotropic functions: An answer to Professor G.F. Smith's criticism of my papers on representations for isotropic functions*. II. *Vector-valued isotropic functions, symmetric tensor-valued isotropic functions, and skew-symmetric tensor-valued isotropic functions*, Arch. Ration. Mech. Anal., 36 (1970), pp. 198–223.

[115B] C.-C. Wang, *Corrigendum to my recent papers on "Representations for isotropic functions,"* Arch. Ration. Mech. Anal., 43 (1971), pp. 392–395.

[116] D.J. Wedlock, I.J. Fabris, and J. Grimsey, *Sedimentation in polydisperse particulate suspensions*, Colloids and Surfaces, 43 (1990), pp. 67–81.

[117] R.H. Weiland, Y.P. Fessas, and B.V. Ramarao, *On instabilities arising during sedimentation of two-component mixtures of solids*, J. Fluid Mech., 142 (1984), pp. 383–389.

[118] A.I. Vol'pert and S.I. Hudjaev, *Cauchy's problem for degenerate second order quasilinear parabolic equations*, Math. USSR Sb., 7 (1969), pp. 365–387.

[119] Z. Wu and J. Yin, *Some properties of functions in BV_x and their applications to the uniqueness of solutions for degenerate quasilinear parabolic equations*, Northeastern Math. J., 5 (1989), pp. 395–422.

[120] Z. Wu, J. Zhao, J. Yin, and H. Li, *Nonlinear Diffusion Equations*, World Scientific, Singapore, 2001.

[121] J. Zeng and D.R. Lowe, *A numerical model for sedimentation from highly-concentrated multi-sized suspensions*, Math. Geol., 24 (1992), pp. 393–415.

[122] Y. Zimmels, *Theory of density separation of particulate systems*, Powder Technol., 43 (1985), pp. 127–139.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 81–93

DIVERGENCE CRITERION FOR GENERIC PLANAR SYSTEMS*

SERGEI S. PILYUGIN† AND PAUL WALTMAN‡

Abstract. The divergence criterion has been shown to be helpful in distinguishing between sub- and supercritical Hopf bifurcations, but its applicability is limited to systems whose divergence is sign definite. A step-by-step computational procedure which allows one to extend the applicability of the divergence criterion is derived by altering the system to an equivalent one with sign definite divergence. The procedure is based on multiplying the original vector field by a positive quadratic function in a neighborhood of the bifurcating rest point. This procedure is then applied to several examples of planar systems that exhibit the Hopf bifurcation. Specifically, it is demonstrated that only supercritical bifurcations occur in a system modeling specific immune responses with handling time. It is also shown that the FitzHugh–Nagumo equations and the chemostat equations with substrate inhibition and linear yield coefficient may exhibit both sub- and supercritical Hopf bifurcations. In both cases, simple analytic criteria for determining the criticality of the bifurcation are presented.

Key words. divergence criterion, subcritical Hopf bifurcation, chemostat, FitzHugh–Nagumo equations

AMS subject classifications. 34C23, 37G10, 92D25

DOI. 10.1137/S0036139902418419

1. Introduction.

The bifurcation of a rest point for a system of ordinary differential equations to a periodic solution has been an intriguing area of research for the past half-century. The early work of Hopf [10] is usually referenced as the beginning point of research in this area, and this type of bifurcation bears his name.[1] The theory has been developed very extensively since. Several textbooks cover the subject, including those of Marsden and McCracken [14], Hassard, Kazarinoff, and Wan [8], Chow and Hale [4], and Kuznetsov [13]. The general subject of bifurcations has been developed to a sophisticated level, and it is now a proper part of nonlinear functional analysis.

Bifurcations are important in physical and biological systems because they represent the points at which the dynamics of the system undergoes a qualitative change. In terms of the parameters of the model system, the bifurcation points can frequently be expressed as thresholds. In many instances, experiments can be designed to detect such thresholds to test a particular model and/or theory. We refer the reader to [11] for an expository article on bifurcations in mathematical biology.

Many population models are described by planar dynamical systems, and simply detecting the existence of a Hopf bifurcation is not difficult. However, determining the direction of bifurcation, whether the bifurcation is subcritical or supercritical (i.e., determining the *criticality* of the bifurcation), is a more delicate problem, as the calculations in the above cited textbooks show. The subcritical bifurcations are especially important in biological systems because they show the existence of (often

*Received by the editors November 25, 2002; accepted for publication (in revised form) April 1, 2003; published electronically October 14, 2003.
http://www.siam.org/journals/siap/64-1/41841.html
†Corresponding author. Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (pilyugin@math.ufl.edu).
‡Department of Mathematics and Computer Science, Emory University, Atlanta, GA 30322 (waltman@emory.edu).
[1]Depending on the source, the Hopf bifurcation may be also referred to as the Andronov–Hopf bifurcation.

unexpected) periodic solutions and multiple periodic solutions in dissipative systems [19, 9, 16].

In previous work [16], a planar bifurcation theorem which determined the criticality of bifurcation was established using the divergence criterion. In particular, it was shown that a subcritical Hopf bifurcation produces at least two limit cycles in a planar, dissipative system. The applicability of the theorem was restricted to systems whose divergence was of one sign (except for a set of measure zero) in a neighborhood of the bifurcation point. In this work, we develop a general approach for determining the criticality of Hopf bifurcations in planar dynamical systems. We show that, for a generic system, one can multiply the vector field by a positive quadratic function and obtain a system whose divergence is sign definite near the bifurcating rest point. Since the resulting system has the same set of trajectories as the original system, the divergence criterion will determine the criticality of the bifurcation. This approach makes Theorem 2.1 in [16] applicable to a wide class of problems.

The divergence criterion is a generalization of the Dulac criterion. This criterion was used by Hofbauer and So [9] to determine the criticality of the Hopf bifurcation for a class of predator-prey equations, and was later generalized by Pilyugin and Waltman [16]. See also [17] for an earlier planar bifurcation theorem in this direction, and see Wolkowicz [19] and Zhu, Campbell, and Wolkowicz [20] for bifurcation analysis of predator-prey systems using the Lyapunov coefficient method. The change in the vector field simplifies the calculations and often renders them amenable to direct computation or to symbolic algebraic processors such as Mathematica [18] or Maple [7]. Sometimes, the simplification can be truly significant. In [16], the use of the divergence criterion resulted in the correction of mistakes found in a series of papers [1, 5, 6] that used the Lyapunov coefficient criterion.

This paper is organized as follows. We describe the construction of the quadratic function in section 2 and formulate the criterion for determining the criticality of the Hopf bifurcation. In section 3, we illustrate the procedure using two important biological problems. In section 4, we introduce a nonlinear rescaling of the vector field, which further simplifies the divergence criterion for a specific set of planar systems including the chemostat (also known as a bio-reactor or a CSTR), and study the Hopf bifurcation in the chemostat with variable yield and substrate inhibition. We conclude with a discussion section.

2. Divergence criterion for generic systems. Consider a planar dynamical system

$$x' = f(x, y), \qquad y' = g(x, y), \tag{1}$$

where f and g are sufficiently smooth, and assume that $(0, 0)$ is a center, that is, $f^0 = g^0 = 0$, $f^0_x + g^0_y = 0$, and $f^0_x g^0_y - f^0_y g^0_x > 0$, where we adopt the notation $F^0 = F(0, 0)$. Necessarily, $f^0_y g^0_x < 0$. We remark that all of the subsequent calculations and conclusions will remain valid if $(0, 0)$ is replaced by (x^0, y^0) and all derivatives are computed at (x^0, y^0).

The *divergence criterion* states that if the divergence of the vector field of (1) is negative (positive) almost everywhere in some neighborhood of $(0, 0)$, then $(0, 0)$ is a stable (unstable) spiral point. In our earlier work [16], we showed that the Hopf bifurcation is supercritical (subcritical) if the bifurcating rest point is a stable (unstable) spiral point. Therefore, we demonstrated that the criticality of the bifurcation can be determined from the stability of the bifurcating rest point. In this paper, we use

the divergence criterion to distinguish between stable and unstable spiral points or, equivalently, between super- and subcritical Hopf bifurcations.

The divergence criterion may not apply directly to the original system (1), because the divergence of its vector field may not be sign definite near the origin. In this section, we show that for a generic vector field (1), one can choose a quadratic function $a(x, y)$ so that $a(0, 0) = 1$ and the divergence of (af, ag) given by

$$\phi(x, y) = (af)_x + (ag)_y = a(f_x + g_y) + a_x f + a_y g \tag{2}$$

is sign definite in some neighborhood of $(0, 0)$. Since $a(x, y)$ is necessarily positive in some neighborhood of $(0, 0)$, the trajectories of (1) coincide with the trajectories of

$$x' = a(x, y)f(x, y), \qquad y' = a(x, y)g(x, y) \tag{3}$$

near $(0, 0)$. Consequently, systems (1) and (3) have the same orbital structure in a neighborhood of $(0, 0)$.

We begin by formally expanding ϕ, using the Taylor polynomial of second order

$$\phi = \phi^0 + \phi^0_x x + \phi^0_y y + \frac{1}{2}\left(\phi^0_{xx}x^2 + 2\phi^0_{xy}xy + \phi^0_{yy}y^2\right) + H.O.T., \tag{4}$$

where $H.O.T.$ denotes higher order terms. Evaluating (2) at $(0, 0)$, we find that $\phi^0 = 0$. Differentiating (2) yields

$$\phi_x = a(f_{xx} + g_{yx}) + a_x(f_x + g_y) + a_{xx}f + a_x f_x + a_{yx}g + a_y g_x, \tag{5}$$

$$\phi_y = a(f_{xy} + g_{yy}) + a_y(f_x + g_y) + a_{xy}f + a_x f_y + a_{yy}g + a_y g_y. \tag{6}$$

Setting $a(0, 0) = 1$, it follows that

$$\phi^0_x = (f^0_{xx} + g^0_{yx}) + a^0_x f^0_x + a^0_y g^0_x, \quad \phi^0_y = (f^0_{xy} + g^0_{yy}) + a^0_x f^0_y + a^0_y g^0_y. \tag{7}$$

Since $f^0_x g^0_y - f^0_y g^0_x > 0$, equations (7) uniquely define a^0_x and a^0_y. Our primary interest is, of course, to eliminate the first order terms in (4). Thus we set $\phi^0_x = \phi^0_y = 0$ in (7) and solve for a^0_x and a^0_y to obtain

$$a^0_x = \frac{(f^0_{xx} + g^0_{yx})g^0_y - (f^0_{xy} + g^0_{yy})g^0_x}{f^0_y g^0_x - f^0_x g^0_y}, \tag{8}$$

$$a^0_y = \frac{-(f^0_{xx} + g^0_{yx})f^0_y + (f^0_{xy} + g^0_{yy})f^0_x}{f^0_y g^0_x - f^0_x g^0_y}. \tag{9}$$

Subsequent differentiation of (5) and (6) yields

$$\begin{aligned}\phi_{xx} = {} & a_x(f_{xx} + g_{yx}) + a(f_{xxx} + g_{yxx}) + a_{xx}(2f_x + g_y) + a_x(2f_{xx} + g_{yx}) \\ & + a_{yx}g_x + a_y g_{xx} + a_{xxx}f + a_{xx}f_x + a_{yxx}g + a_{yx}g_x,\end{aligned}$$

$$\begin{aligned}\phi_{xy} = {} & a_y(f_{xx} + g_{yx}) + a(f_{xxy} + g_{yxy}) + a_{xy}(2f_x + g_y) + a_x(2f_{xy} + g_{yy}) \\ & + a_{yy}g_x + a_y g_{xy} + a_{xxy}f + a_{xx}f_y + a_{yxy}g + a_{yx}g_y,\end{aligned}$$

$$\begin{aligned}\phi_{yy} = {} & a_y(f_{xy} + g_{yy}) + a(f_{xyy} + g_{yyy}) + a_{yy}(f_x + 2g_y) + a_y(f_{xy} + 2g_{yy}) \\ & + a_{xy}f_y + a_x f_{yy} + a_{xyy}f + a_{xy}f_y + a_{yyy}g + a_{yy}g_y.\end{aligned}$$

Assuming that the functions f and g are sufficiently smooth, the mixed derivatives are independent of the order of differentiation. Therefore, evaluating the above expression at the center $(0,0)$ yields

$$\phi^0_{xx} = Q^{xx} + 2f^0_x a^0_{xx} + 2g^0_x a^0_{xy}, \tag{10}$$

$$\phi^0_{xy} = Q^{xy} + f^0_y a^0_{xx} + g^0_x a^0_{yy}, \tag{11}$$

$$\phi^0_{yy} = Q^{yy} + 2f^0_y a^0_{xy} + 2g^0_y a^0_{yy}, \tag{12}$$

where

$$Q^{xx} = (f^0_{xxx} + g^0_{yxx}) + a^0_x(3f^0_{xx} + 2g^0_{xy}) + a^0_y g^0_{xx},$$

$$Q^{xy} = (f^0_{xxy} + g^0_{yxy}) + a^0_x(2f^0_{xy} + g^0_{yy}) + a^0_y(f^0_{xx} + 2g^0_{xy}),$$

$$Q^{yy} = (f^0_{xyy} + g^0_{yyy}) + a^0_x f^0_{yy} + a^0_y(2f^0_{xy} + 3g^0_{yy}).$$

Thus far, the linear terms in (4) have been eliminated by choosing appropriate values for a^0_x and a^0_y. In what follows, we seek to choose the values a^0_{xx}, a^0_{xy}, and a^0_{yy} so as to make the second order terms in (4) sign definite. The second order terms in (4) are sign definite whenever the discriminant

$$\mathcal{D} = \phi^0_{xx}\phi^0_{yy} - (\phi^0_{xy})^2$$

is positive. We set

$$a^0_{xx} = -\frac{1}{2}\frac{Q^{xy}}{f^0_y}, \qquad a^0_{yy} = -\frac{1}{2}\frac{Q^{xy}}{g^0_x}, \tag{13}$$

because such a choice yields $\phi^0_{xy} = 0$. The discriminant then can be written as

$$\mathcal{D} = \left(Q^{xx} - \frac{Q^{xy}f^0_x}{f^0_y} + 2g^0_x a^0_{xy}\right)\left(Q^{yy} - \frac{Q^{xy}g^0_y}{g^0_x} + 2f^0_y a^0_{xy}\right) = (\beta_1 + \alpha_1 z)(\beta_2 + \alpha_2 z), \tag{14}$$

where $z = a^0_{xy}$, $\beta_1 = Q^{xx} - (Q^{xy}f^0_x/f^0_y)$, $\beta_2 = Q^{yy} - (Q^{xy}g^0_y/g^0_x)$, $\alpha_1 = 2g^0_x$, and $\alpha_2 = 2f^0_y$. Since the product $\alpha_1\alpha_2 = 4g^0_x f^0_y < 0$, the discriminant $\mathcal{D}$ is positive for any z located strictly between the roots $z_1 = -\beta_1/\alpha_1$ and $z_2 = -\beta_2/\alpha_2$. For generic functions f and g, $z_1 \neq z_2$. Thus we choose

$$z^* = -\frac{1}{2}\left(\frac{\beta_1}{\alpha_1} + \frac{\beta_2}{\alpha_2}\right) = -\frac{1}{2}\frac{Q^{xx}f^0_y + Q^{yy}g^0_x}{2g^0_x f^0_y} \tag{15}$$

and set $a^0_{xy} = z^*$.

At this point, we have determined all coefficients of the quadratic function

$$a(x,y) = 1 + a^0_x x + a^0_y y + \frac{1}{2}(a^0_{xx}x^2 + 2a^0_{xy}xy + a^0_{yy}y^2). \tag{16}$$

Finally, we transform the original vector field (f, g) into a vector field (af, ag) with sign definite divergence ϕ near the origin. The divergence ϕ is positive (negative) if ϕ^0_{xx} is positive (negative). Substituting (13) and (15) into (10), we find that

$$\phi^0_{xx} = \frac{1}{2}\left(Q^{xx} - Q^{yy}\frac{g^0_x}{f^0_y}\right) - Q^{xy}\frac{f^0_x}{f^0_y}. \tag{17}$$

For a given planar system that undergoes a Hopf bifurcation, we evaluate appropriate partial derivatives of its vector field at the bifurcation point and compute the quantity (17). The Hopf bifurcation is supercritical (subcritical) if (17) is negative (positive).

Unfortunately, for a generic system, expression (17) may become too complicated for symbolic applications. In this case, our method will have no advantage over the standard normal form computation. However, our approach can, sometimes, have a clear advantage over the standard method. To illustrate this, we treat several examples in subsequent sections. Expression (17) will be greatly simplified if the divergence of the vector field essentially involves only one of the state space variables x or y. It is therefore helpful to introduce a preliminary change of variables to achieve this, whenever possible. A specific change of variables that applies to chemostats is discussed in section 4.

3. Applications. In this section, we apply the change of vector fields to two examples in biological literature and determine the criticality of bifurcation. Before beginning, we note two changes from the usual presentation of bifurcation results.

- It is possible, and the theory is usually presented this way, to change variables so that the bifurcating rest point is always at the origin. Such a change, however, complicates the calculations for a specific problem, and we do not make it. The reader should be cautioned that, as parameters vary, the coordinates of the rest point vary.
- The traditional approach is to fix all of the parameters except one (usually designated as the bifurcation parameter) and let that parameter determine the bifurcation. We choose instead to present a *bifurcation locus*, which is defined as a hypersurface in the parameter space on which the bifurcation occurs. We have two reasons for doing this. First of all, biological problems frequently have many parameters, and it would be artificial to select a single one unless there is a specific experiment which can vary it. Secondly, our technique for determining the criticality of bifurcation depends only on the stability of the rest point at the critical parameter value(s). This implies that any parametric path crossing the bifurcation locus will produce a bifurcation whose criticality is determined exclusively by its crossing point on the bifurcation locus. In particular, any two parametric paths crossing the bifurcation locus via the same point will produce Hopf bifurcations of the same criticality. Of course, one has to ascertain that a bifurcation does indeed occur, that is, that the rest point does change its stability along the parametric path. On the other hand, our result does not require that the parametric path be strictly transverse to the bifurcation locus or, equivalently, that the pair of complex eigenvalues cross the imaginary axis with nonzero velocity. For more details, we refer the reader to the proof of the original Theorem 2.1 in [16].

For any crossing point on the bifurcation locus, the linearization of a planar system has purely imaginary eigenvalues. Such a rest point for the nonlinear system can be a stable or an unstable spiral, or a center, the choice being determined by the

nonlinear terms. Our quadratic factor determines whether the rest point is a stable spiral or an unstable spiral, depending on the sign of the (sign specific) divergence. It is also possible that the quadratic terms in (4) vanish, in which case our technique does not apply. If this is the case, then the rest point could still be a center or a spiral determined by nonlinear terms of higher order (and hence such a case would be nongeneric).

In the next two subsections, we study the criticality of Hopf bifurcations in two biological problems, where we can add to results already in the literature. These examples also illustrate the ease with which the technique can be applied to biological problems.

3.1. Specific immune responses with handling time. In this section, we apply the general divergence criterion to the model of specific immunity studied by Pilyugin and Antia in [15]. The authors reported the existence of Hopf bifurcation in the system

$$x' = rx - \frac{hx}{k+x}y, \tag{18}$$

$$y' = a + \left(\frac{\rho x}{k+x} - d\right) y, \tag{19}$$

where $r,\ h,\ k,\ a,\ \rho,\ d$ are positive parameters. Here x and y are dimensionless variables that represent the abundance of parasite (i.e., the number of infected cells) and the magnitude of the specific (cytotoxic) immune response, respectively. In this model, both the proliferation rate of immune cells $\frac{\rho x}{k+x}$ and the killing rate of infected cells $\frac{hx}{k+x}$ saturate as the number of infected cells x becomes large. The quantities r, a, and d represent the (per capita) rate of parasite replication, the input of immune cells from an external source, and the (per capita) death rate of immune cells, respectively. We restrict the bifurcation analysis to the biologically relevant case $x, y > 0$.

To simplify computations, we multiply the vector field of (18)–(19) by a positive function $k + x$ and consider the new system of the form

$$x' = rx(k+x) - hxy = f(x,y), \quad x(0) > 0, \tag{20}$$

$$y' = a(k+x) + (\rho x - d(k+x))y = g(x,y), \quad y(0) > 0. \tag{21}$$

Since $k + x > 0$, the phase portraits of (18)–(19) and (20)–(21) are identical.

The bifurcating rest point of (20)–(21) has coordinates

$$x^0 = \frac{rdk - ah}{r(\rho - d)} > 0, \quad y^0 = \frac{kr\rho - ah}{h(\rho - d)} > 0. \tag{22}$$

We compute the partial derivatives of f and g to find

$$f_x = rk + 2rx - hy, \quad f_y = -hx, \quad f_{xx} = 2r, \quad f_{xy} = -h, \quad f_{yy} = 0,$$

$$g_x = a + (\rho - d)y, \quad g_y = (\rho - d)x - dk, \quad g_{xy} = \rho - d, \quad g_{xx} = g_{yy} = 0.$$

Consequently,

$$f_x^0 = rx^0, \quad f_y^0 = -hx^0, \quad f_{xx}^0 = 2r, \quad f_{xy}^0 = -h, \quad f_{yy}^0 = 0,$$

$$g_x^0 = \frac{kr\rho}{h}, \quad g_y^0 = -\frac{ah}{r}, \quad g_{xy}^0 = \rho - d, \quad g_{xx}^0 = g_{yy}^0 = 0.$$

The bifurcating rest point must necessarily satisfy $0 = f_x^0 + g_y^0 = rx^0 - \frac{ah}{r}$, and thus

$$x^0 = \frac{ah}{r^2}.$$

Equating this value with that of (22), we find that the bifurcation locus is a subset of the hypersurface

$$d(r^2k + ah) = ah(\rho + r). \tag{23}$$

The determinant of the variational matrix is given by

$$\det(J) = -rx^0\frac{ah}{r} + hx^0\frac{kr\rho}{h} = x^0(kr\rho - ah).$$

A necessary condition for the Hopf bifurcation is that $\det(J) > 0$. Since $x^0 > 0$, it follows that $kr\rho - ah > 0$, and inequalities (22) further imply that $\rho - d > 0$ and $rdk - ah > 0$. Since $\rho - d > 0$ and $rdk - ah > 0$ together imply $kr\rho - ah > 0$, the bifurcation locus can be described as the subset of (23) restricted by two inequalities

$$rdk - ah > 0, \quad \rho - d > 0. \tag{24}$$

Using (8)–(9), we compute

$$a_x^0 = \frac{kr\rho - \frac{ah}{r}(2r + (\rho - d))}{\frac{ah}{r^2}(ah - kr\rho)}, \quad a_y^0 = \frac{(r + (\rho - d))h}{ah - kr\rho}. \tag{25}$$

The quantities Q^{**} are

$$Q^{xx} = a_x^0(6r + 2(\rho - d)), \quad Q^{xy} = -2ha_x^0 + a_y^0(2r + 2(\rho - d)), \quad Q^{yy} = -2ha_y^0.$$

Therefore,

$$\phi_{xx}^0 = a_x^0(r + (\rho - d)) + a_y^0\left(h + \frac{2r}{h}(r + (\rho - d))\right),$$

which can be simplified to

$$\phi_{xx}^0 = -\frac{(r + (\rho - d))^2}{k(r + \rho)}. \tag{26}$$

Clearly, $\phi_{xx}^0 < 0$. Since the divergence of the rescaled vector field is negative definite at any point on the bifurcation locus, the bifurcation is always supercritical.

3.2. Diffusionless FitzHugh–Nagumo equations. Several numerical examples of supercritical and subcritical Hopf bifurcations were presented by Kostova, Ravindran, and Schonbek [12] in the context of the classical FitzHugh–Nagumo equations. They derived a complicated expression to determine the criticality of the Hopf bifurcation using the normal form calculation. In this section, we use the divergence criterion to derive a simple analytic criterion to determine the criticality of the Hopf bifurcation.

The FitzHugh–Nagumo equations for a single neuron are

$$x' = F(x) - y + I = f(x, y), \tag{27}$$

$$y' = \quad x - wy \quad = g(x, y), \tag{28}$$

where $F(x) = \varepsilon x(1-x)(x-\lambda)$ and $\varepsilon > 0,\ 0 < \lambda < 1,\ w > 0$, and I are parameters. The variable x represents the membrane potential, y is the recovery variable that represents a negative feedback, and I is the membrane current.

Computing the partial derivatives of f and g, we find that

$$f_x = F'(x), \quad f_y = -1, \quad f_{xx} = F''(x), \quad f_{xy} = f_{yy} = 0,$$

$$g_x = 1, \quad g_y = -w, \quad g_{xy} = g_{xx} = g_{yy} = 0.$$

Consequently, at any rest point (x^0, y^0), we have

$$f^0_x = F'(x^0), \quad f^0_y = -1, \quad f^0_{xx} = F''(x^0), \quad f^0_{xy} = f^0_{yy} = 0,$$

$$g^0_x = 1, \quad g^0_y = -w, \quad g^0_{xy} = g^0_{xx} = g^0_{yy} = 0.$$

The bifurcation locus consists of rest points (x^0, y^0) such that $f^0_x + g^0_y = 0$ and $f^0_x g^0_y - f^0_y g^0_x > 0$. The former condition implies that $F'(x^0) = w$. The latter condition then implies that $1 - w^2 > 0$. Hence the bifurcation locus is the set of rest points (x^0, y^0) such that

$$f^0 = g^0 = 0, \quad F'(x^0) = w, \quad w^2 < 1. \tag{29}$$

Since $F'(x) = \varepsilon(-3x^2 + 2(1+\lambda)x - \lambda)$, the second condition in (29) implies that x^0 must satisfy the quadratic equation

$$3(x^0)^2 - 2(1+\lambda)x^0 + \lambda + \frac{w}{\varepsilon} = 0$$

or, equivalently,

$$x^0_{1,2} = \frac{(1+\lambda) \pm \sqrt{(1+\lambda)^2 - 3(\lambda + \frac{w}{\varepsilon})}}{3}. \tag{30}$$

Using (8)–(9), we compute

$$a^0_x = \frac{wF''(x^0)}{1-w^2}, \quad a^0_y = -\frac{F''(x^0)}{1-w^2}. \tag{31}$$

The quantities Q^{**} are

$$Q^{xx} = F'''(x^0) + \frac{3w(F''(x^0))^2}{1-w^2} = -6\varepsilon + \frac{3w(F''(x^0))^2}{1-w^2},$$

$$Q^{xy} = -\frac{(F''(x^0))^2}{1-w^2}, \quad Q^{yy} = 0.$$

Therefore,

$$\phi^0_{xx} = \frac{1}{2}\left(-6\varepsilon + \frac{3w(F''(x^0))^2}{1-w^2}\right) + \frac{(F''(x^0))^2}{1-w^2}\frac{F'(x^0)}{-1},$$

which can be further simplified to

$$\phi^0_{xx} = \frac{1}{2}\left(-6\varepsilon + \frac{w(F''(x^0))^2}{1-w^2}\right). \tag{32}$$

Substituting $F''(x^0) = 2\varepsilon(1+\lambda-3x^0)$ into (32), we rewrite the quantity ϕ^0_{xx} as

$$\phi^0_{xx} = \varepsilon\left(-3 + \frac{2\varepsilon w(1+\lambda-3x^0)^2}{1-w^2}\right). \tag{33}$$

Using (30), we can rewrite the quantity $(1+\lambda-3x^0)^2$ as

$$(1+\lambda-3x^0)^2 = (1+\lambda)^2 - 3\left(\lambda + \frac{w}{\varepsilon}\right).$$

Substituting this expression into (33), we finally obtain

$$\phi^0_{xx} = \frac{\varepsilon}{1-w^2}\left(2\varepsilon w(1-\lambda+\lambda^2) - 3(1+w^2)\right). \tag{34}$$

The multiplier $\frac{\varepsilon}{1-w^2}$ is positive due to (29). Consequently, ϕ^0_{xx} has the same sign as the quantity $2\varepsilon w(1-\lambda+\lambda^2) - 3(1+w^2)$. The Hopf bifurcation in the FitzHugh–Nagumo equations is subcritical if $\phi^0_{xx} > 0$, that is, if it occurs on the part of the bifurcation locus where

$$\varepsilon > \frac{3(1+w^2)}{2w(1-\lambda+\lambda^2)},$$

and supercritical if it occurs on the part of the bifurcation locus where the reversed strict inequality holds.

In [12], two numerical examples were presented: a Figure 2 with $\varepsilon = 14.0,\ w = 0.38,\ \lambda = 0.1$ and a Figure 3 with $\varepsilon = 14.0,\ w = 0.06,\ \lambda = 0.5$. In the former case,

$$\varepsilon = 14.0 > \frac{3(1+0.38^2)}{2\cdot 0.38(1-0.1+0.1^2)} = 4.964,$$

and the bifurcation is subcritical. In the latter case,

$$\varepsilon = 14.0 < \frac{3(1+0.06^2)}{2\cdot 0.06(1-0.5+0.5^2)} = 33.453,$$

and the bifurcation(s) are supercritical.

4. Rescaling for chemostat equations. The method presented in this section for rescaling the vector field is a generalization of the technique used by Hofbauer and So [9]. Specifically, we consider the system

$$x' = f(x) - q_1(y)g(x), \quad y' = q_2(y)h(x), \tag{35}$$

where f, g, h, q_i are sufficiently smooth and such that positive solutions of (35) remain positive. Also, suppose that $g(x) > 0$ and $q_2(y) > 0$ for $x, y > 0$. We multiply the vector field (35) by a positive function $Q(y)/g(x)$ to obtain a new system

$$x' = Q(y)f(x)/g(x) - q_1(y)Q(y) = Q(y)G(x) - q_1(y)Q(y), \tag{36}$$

$$y' = Q(y)q_2(y)h(x)/g(x) = Q(y)q_2(y)H(x), \tag{37}$$

where $(Qq_2)' = \beta Q$ and β is a real number to be determined later. The explicit expression for $Q(y)$ is

$$Q(y) = \frac{\exp\left(\beta\int \frac{dy}{q_2(y)}\right)}{q_2(y)} > 0.$$

Any positive rest point (x^0, y^0) of (36)–(37) must satisfy $H(x^0) = 0$. The divergence of the new vector field (36)–(37) is given by

$$D(x,y) = Q(y)G'(x) + (Q(y)q_2(y))'H(x) = Q(y)\big(G'(x) + \beta H(x)\big). \tag{38}$$

Now suppose that (x^0, y^0) is the bifurcating rest point, that is, that $D(x^0, y^0) = 0$. Since $Q(y^0) > 0$ and $H(x^0) = 0$, we necessarily have that

$$G'(x^0) + \beta H(x^0) = G'(x^0) = 0.$$

We choose $\beta = -G''(x^0)/H'(x^0)$, so that in a small neighborhood of (x^0, y^0),

$$G'(x) + \beta H(x) = \frac{\delta}{2}(x - x^0)^2 + H.O.T.,$$

where

$$\delta = G'''(x^0) - \frac{G''(x^0)}{H'(x^0)}H''(x^0) = H'(x^0)\left(\frac{G''}{H'}\right)'(x^0). \tag{39}$$

Since $Q(y) > 0$, the sign of $D(x,y)$ near (x^0, y^0) is the same as the sign of δ. Consequently, the application of the divergence criterion to systems of type (35) can be greatly simplified. For example, the predator-prey models analyzed in [2, 9, 19, 20] fall into this category. The criteria for the criticality of Hopf bifurcations obtained in these works can be directly compared to the expression (39). Various models of the chemostat are also of type (35). In the following subsection, we illustrate this simplified approach by treating a special case of the chemostat with substrate inhibition and a linear yield coefficient.

4.1. Chemostats with substrate inhibition and linear yields. In this section, we study the Hopf bifurcation in the model of a chemostat with linear yield coefficient which also features substrate inhibition of growth at higher substrate levels. For a general study of the chemostat with inhibition, constant yield, and several competitors, see Butler and Wolkowicz [3]. The original model presented in Agrawal et al. [1], takes the following form:

$$x' = 1 - x - y\frac{\mu(x)}{1 + cx}, \tag{40}$$

$$y' = y(\mu(x) - 1), \tag{41}$$

where x and y denote the dimensionless substrate and biomass concentrations, and $\mu(x) = mx\exp(-x/K)$ is the microbial growth rate. The function $1 + cx$ represents the yield coefficient,[2] which is assumed to increase linearly with substrate concentration; thus $c > 0$.

Equations (40)–(41) are of the form (35) with

$$f(x) = 1 - x, \qquad g(x) = \frac{mx\exp(-x/K)}{1 + cx},$$

$$h(x) = mx\exp(-x/K) - 1, \qquad q_1(y) = q_2(y) = y.$$

[2]The yield coefficient is defined as the ratio of the amount of substrate consumed to the amount of biomass produced at a steady state. There is strong biological evidence that the yield may increase with substrate concentration. For details, see [16] and the references therein.

Equations (40)–(41) admit up to two positive rest points. The x-coordinate of a positive rest point must satisfy $\mu(x) = 1$ with $0 < x < 1$. The function $\mu(x)$ has a maximum at $x = K$, and its maximal value is given by $\mu_{\max} = mKe^{-1}$. Consequently, if $mK > e$, then there exist two positive solutions of $\mu(x) = 1$, $x_1^0 < K < x_2^0$. It is easy to verify that the rest point with $x = x_2^0$ (if it is feasible, that is, if $x_2^0 < 1$) is always a saddle. Consequently, the Hopf bifurcation can occur only at the rest point (x_1^0, y_1^0) with $0 < x_1^0 < \min(1, K)$ and $y_1^0 = (1 - x_1^0)(1 + cx_1^0)$. The bifurcation occurs when the trace of the variational matrix of (40)–(41) at (x_1^0, y_1^0) equals zero:

$$-1 - y_1^0 \frac{d}{dx}\Big(\frac{\mu(x)}{1+cx}\Big)(x_1^0) = 0.$$

Consequently, on the bifurcation locus, the value of c is given by

$$c = \frac{K - x_1^0 + (x_1^0)^2}{(x_1^0)^2(1 - K - x_1^0)}. \tag{42}$$

The bifurcation occurs in the feasible region if the value of c given by (42) is positive, that is, if $x_1^0 < 1 - K$ and $0 < K < 1$.

To determine the criticality of the Hopf bifurcation, we computed the functions $G(x)$ and $H(x)$ as defined in (36)–(37) and found that

$$G(x) = \frac{(1-x)(1+cx)\exp(x/K)}{mx}, \tag{43}$$

$$H(x) = \frac{(1+cx)(mx - \exp(x/K))}{mx}. \tag{44}$$

Then we created a Mathematica [18] notebook to compute the quantity δ defined in (39), and found that

$$\delta(x(m,K),K) = \frac{P_0(x) + KP_1(x) + K^2P_2(x) + K^3P_3(x) + K^4P_4(x)}{K^2(K-x)(1-x)(1-K-x)x^3}, \tag{45}$$

where $P_0(x) = 3(1-x)^3x^3$, $P_1(x) = 2x^2(1-x)^2(x-6)$, $P_2(x) = 2x(x-1)(x^2 + x - 8)$, $P_3(x) = 2(5x^2 - 4x - 3)$, $P_4(x) = 2(3-x)$, and $x_1^0 = x(m,K)$. The Hopf bifurcation in (40)–(41) is subcritical if $\delta > 0$ and supercritical if $\delta < 0$.

The existence of both sub- and supercritical Hopf bifurcations in (40)–(41) is illustrated in Figure 1. For $(K,m) \in A$, no rest point with $0 < x < 1$ exists. The curve between A and B is given by $m = e/K$. Region B ($\delta > 0$) corresponds to subcritical bifurcations. The curve between B and C is the implicit plot of $\delta(x(m,K),K) = 0$. Region C ($\delta < 0$) corresponds to supercritical bifurcations. The curve between C and D is the implicit plot of $x(m,K) = 1 - K$. For $(K,m) \in D$, no bifurcations with $c > 0$ occur. The region B terminates at $K = 0.413$, and the region C terminates at $K = 0.5$.

5. Discussion. We have developed a unifying approach for studying the Hopf bifurcation for generic planar systems, which stems from the divergence criterion. Specifically, we presented a step-by-step computational procedure which can be used to distinguish between sub- and supercritical bifurcations. This procedure can be easily programmed in any standard computer algebra system such as Maple [7] or Mathematica [18] so that all of the necessary computations can be automated. This work may serve as a good example of the analytic approach that involves computerized

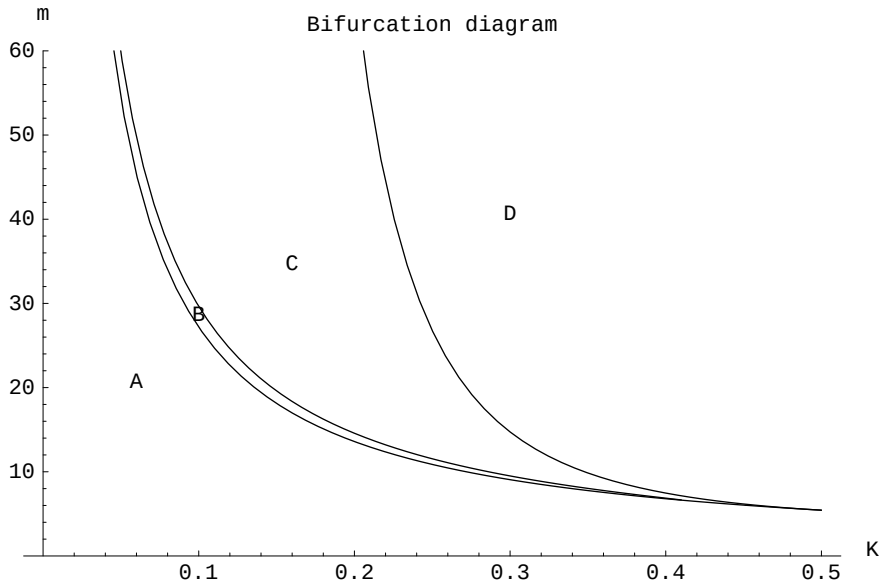


FIG. 1. *Existence and criticality of Hopf bifurcations in the* (K, m) *plane.*

symbolic calculations. The technique is generic in the sense that it can be applied to a generic planar vector field.

We applied our procedure to several important biological systems and obtained new results on the criticality of Hopf bifurcation. Interestingly, we found that in several instances—for example, with diffusionless FitzHugh–Nagumo equations—we were able to perform all calculations by hand in a reasonable amount of time, which illustrated that our method may have an advantage over the calculation of the third Lyapunov coefficient and/or normal form calculation for the Hopf bifurcation. In particular, our method does not require such computational steps as

- changing coordinates to place the bifurcating rest point at the origin,
- finding eigenvalues and eigenvectors of the variational matrix,
- transforming the linear part of the vector field to the canonical form.

In certain examples, our method produces analytic expressions that are easier to simplify.

We have also presented a specific change of variables that works well with a whole class of planar systems including the equations of the chemostat. Performing this change of variables essentially eliminates one of the phase variables from the expression for divergence and thus greatly simplifies the analysis of Hopf bifurcation.

Acknowledgments. The authors are grateful to the anonymous referees and the handling editor for their valuable comments and suggestions.

REFERENCES

[1] R. AGRAWAL, C. LEE, H. C. LIM, AND D. RAMKRISHNA, *Theoretical investigations of dynamic behavior of isothermal continuous stirred tank biological reactors*, Chem. Engrg. Sci., 37 (1982), pp. 453–462.

[2] G. BUTLER AND P. WALTMAN, *Bifurcation from a limit cycle in a two predator-one prey ecosystem modeled on a chemostat*, J. Math. Biol., 12 (1981), pp. 295–310.

[3] G. J. Butler and G. S. K. Wolkowicz, *A mathematical model of the chemostat with a general class of functions describing nutrient uptake*, SIAM J. Appl. Math., 45 (1985), pp. 138–151.
[4] S. N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.
[5] P. S. Crooke, C.-J. Wei, and R. D. Tanner, *The effect of the specific growth rate and yield expressions on the existence of oscillatory behavior of a continuous fermentation model*, Chem. Engrg. Commun., 6 (1980), pp. 333–347.
[6] P. S. Crooke and R. D. Tanner, *Hopf bifurcations for a variable yield continuous fermentation model*, Internat. J. Engrg. Sci., 20 (1982), pp. 439–443.
[7] F. Garvan, *The Maple Book*, Chapman & Hall/ CRC, Boca Raton, FL, 2002.
[8] B. D. Hassard, N. D. Kazarinoff, and Y.-H. Wan, *Theory and Applications of Hopf Bifurcation*, Cambridge University Press, Cambridge, UK, 1980.
[9] J. Hofbauer and J. W.-H. So, *Multiple limit cycles for predator-prey models*, Math. Biosci., 99 (1990), pp. 71–75.
[10] E. Hopf, *Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems*, Berichten der Mathematisch-Physischen Klasse der Sächsischen Akademie der Wissenschaften zu Lepizig, 94 (1942), pp. 1–22. (An English translation of this paper can be found in Marsden and McCracken [14].)
[11] F. Hoppensteadt and P. Waltman, *Did something change? Thresholds in population models*, in Trends in Nonlinear Analysis, M. Kirkilionis, S. Kroemker, R. Rannacher, and F. Tomi, eds., Springer-Verlag, Berlin, 2002, pp. 341–374.
[12] T. Kostova, R. Ravindran, and M. Schonbek, *FitzHugh-Nagumo revisited: Types of bifurcations, periodic forcing and stability regions by a Lyapunov functional*, Internat. J. Bifurc. Chaos, to appear.
[13] Y. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 1995.
[14] J. E. Marsden and M. McCracken, *The Hopf Bifurcation and Its Applications*, Springer-Verlag, New York, 1976.
[15] S. S. Pilyugin and R. Antia, *Modeling immune responses with handling time*, Bull. Math. Biol., 62 (2000), pp. 869–890.
[16] S. S. Pilyugin and P. Waltman, *Multiple limit cycles in the chemostat with variable yield*, Math. Biosci., 182 (2003), pp. 151–166.
[17] P. Waltman, *A bifurcation theorem*, Proc. Amer. Math. Soc., 15 (1964), pp. 627–631.
[18] S. Wolfram, *The Mathematica Book*, Cambridge University Press, New York, 1999.
[19] G. S. K. Wolkowicz, *Bifurcation analysis of a predator-prey system involving group defense*, SIAM J. Appl. Math., 48 (1988), pp. 592–606.
[20] H. Zhu, S. A. Campbell, and G. S. K. Wolkowicz, *Bifurcation analysis of a predator-prey system with nonmonotonic functional response*, SIAM J. Appl. Math., 63 (2002), pp. 636–682.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 94–107

A METHOD FOR SOLVING DYNAMICALLY ACCELERATING CRACK PROBLEMS IN LINEAR VISCOELASTICITY*

TANYA L. LEISE† AND JAY R. WALTON‡

Abstract. We develop a general solution technique for a dynamically accelerating crack in a linear viscoelastic material, based on a transform method developed by Slepyan (see [*Models and Phenomena in Fracture Mechanics*, Springer-Verlag, New York, 2002]) for solution of dynamic elastic fracture problems. We review the elastic fracture solution method and then treat the viscoelastic mode III fracture problem for crack tip speeds less than the short-time shear wave speed. The analysis includes an exact, closed-form expression for the stress intensity factor for an arbitrary time dependent crack face traction. As examples, we apply this solution method to the Achenbach–Chao and standard linear solid viscoelastic models.

Key words. dynamic fracture, viscoelastic material

AMS subject classifications. 74D05, 74H05, 74H35, 74R10, 74R15, 74R20, 47G20

DOI. 10.1137/S003613990241953X

1. Introduction.

While dynamic elastic fracture is a well-studied and fairly mature subject (see [4], for example), the literature devoted to dynamic fracture in viscoelastic material is limited. The first exact, closed-form solution for a dynamically accelerating crack in a viscoelastic material was derived by Bourne and Walton [2], who assumed a semi-infinite, antiplane shear crack in an Achenbach–Chao linear viscoelastic solid in the limiting case of a vanishing equilibrium shear modulus. (Also see [13] for an exposition of this viscoelastic solution, and [14] for the inclusion of a Dugdale zone.) With the exception of an approximate analysis by Goleniewski [5], the authors know of no other solutions to date for dynamically accelerating cracks in viscoelastic material.

The first major contributions to the solution of dynamically accelerating cracks in an elastic material were due to Kostrov [6], whose work motivated others in the field such as Eshelby [3]. However, the approach taken by this paper most closely follows the techniques for an elastic material developed by Slepyan [11], which are related to but distinct from the methods developed by Kostrov or Eshelby. We also make use of the ideas developed by Bourne and Walton [2] for a crack in a linear viscoelastic material in the special case of a vanishing equilibrium shear modulus. Our goal is to generalize Bourne and Walton's exact, closed-form solution for this special case to the general case of a linear viscoelastic material. In addition, we derive an explicit expression for the mode III stress intensity factor for a dynamically accelerating crack in a general linear viscoelastic material. One could also use the equations derived in this paper to numerically compute the displacement and stress along the entire boundary (the crack line), facilitating the inclusion of a cohesive zone.

In the following section, we briefly outline the solution method for an accelerating semi-infinite crack in an elastic material for comparison to that for the viscoelastic

*Received by the editors December 12, 2002; accepted for publication (in revised form) March 17, 2003; published electronically October 14, 2003.
http://www.siam.org/journals/siap/64-1/41953.html
†Department of Mathematics, Rose-Hulman Institute of Technology, Terre Haute, IN 47803-3999 (leise@rose-hulman.edu).
‡Department of Mathematics, Texas A&M University, College Station, TX 77843-3368 (jwalton@math.tamu.edu).

problem, to highlight both what is analogous and what is very different. We then extend this solution method to the general case of a dynamically accelerating semi-infinite crack in a linear viscoelastic material in sections 3 and 4. In sections 5 and 6 we apply the solution method to the Achenbach–Chao and standard linear solid viscoelastic models and then make some concluding remarks in section 7.

2. Review of the dynamic elastic problem. The following integral transform method for a dynamically accelerating semi-infinite crack in an elastic material was first developed by Slepyan (see [11]) for all fracture modes. (In fact, in both [11] and [9], a more general problem is considered, where the displacement in front of the crack need not vanish.) It is closely related to the method developed in [15] by Walton and Herrmann, and we extended the method to the cases of finite length and multiple cracks in an elastic material in [8] and [7]. We will review the method as applied to a semi-infinite mode III crack in an elastic material for comparison with the viscoelastic equations that we will derive in sections 3 and 4.

Consider a dynamically accelerating, semi-infinite, antiplane shear crack for a general infinite homogeneous and isotropic linearly elastic body. The governing equations for this elastic fracture problem are

$$\rho\ddot{u}(t) = \mu\triangle u \quad \text{and} \quad \sigma(x,y,t) = \mu\frac{\partial}{\partial y}u(x,y,t), \tag{2.1}$$

where μ is the elastic shear modulus, $u(x,y,t)$ is the out-of-plane displacement $u_3(x,y,t)$, and $\sigma(x,y,t)$ is the shear stress $\sigma_{23}(x,y,t)$. Let $a(t)$ be the crack tip position at time t, as in Figure 2.1. The initial and boundary conditions for this problem are

$$u(x,y,0) = 0 = \dot{u}(x,y,0), \tag{2.2}$$

$$u(x,0,t) = 0 \quad \text{for } x > a(t), \tag{2.3}$$

$$\sigma(x,0,t) = \Lambda(x,t) \quad \text{for } x < a(t), \tag{2.4}$$

$$u(x,y,t) \to 0 \quad \text{as } |y| \to \infty, \tag{2.5}$$

where $a(t)$ is the position of the crack tip as it propagates to the right and $\Lambda(x,t)$ is assumed to be a known loading of the crack face. Furthermore we assume that the crack speed remains subsonic, $0 \le \dot{a}(t) < c$, where $c = \sqrt{\mu/\rho}$ is the shear wave speed.

The key to the solution method is the observation that, after applying Laplace and Fourier transforms and taking the limit as $y \to 0^+$, the governing equations

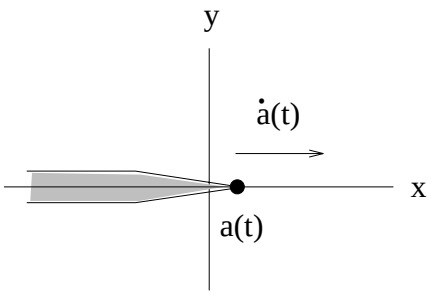


FIG. 2.1. *A semi-infinite dynamically accelerating crack. We consider the antiplane shear case, and thus the crack face displacement is actually perpendicular to the page.*

reduce to a simple transfer map (see [13] for the details of this derivation for both the elastic and viscoelastic cases):

$$\hat{\bar{\sigma}}(p,s) = \hat{\bar{T}}(p,s)\hat{\bar{u}}(p,s), \tag{2.6}$$

where the transfer function is

$$\hat{\bar{T}}(p,s) = -\mu\sqrt{p^2 + s^2/c_s^2}. \tag{2.7}$$

We use Fourier and Laplace transforms defined as follows:

$$\hat{f}(p) = \int_{-\infty}^{\infty} e^{ipx} f(x)dx, \tag{2.8}$$

$$\bar{g}(s) = \int_{0}^{\infty} e^{-st} g(t)dt. \tag{2.9}$$

Decompose the transfer function $T(x,t)$ as $\hat{\bar{T}}(p,s) = \hat{\bar{T}}_+(p,s)\hat{\bar{T}}_-(p,s)$ in such a manner that the functions $T_\pm(x,t)$ and $S_\pm(x,t)$, where $\hat{\bar{S}}_\pm = 1/\hat{\bar{T}}_\pm$, satisfy the following conditions:

$$T_+(x,t) = S_+(x,t) = 0 \quad \text{for } x < ct, \tag{2.10}$$

$$T_-(x,t) = S_-(x,t) = 0 \quad \text{for } x > -ct. \tag{2.11}$$

We also decompose the stress and displacement into parts with support either in front of the crack tip or on the crack face:

$$\sigma_+(x,t) = \begin{cases} \sigma(x,t) & \text{if } x > a(t), \\ 0 & \text{otherwise,} \end{cases} \tag{2.12}$$

$$\sigma_-(x,t) = \begin{cases} \sigma(x,t) & \text{if } x < a(t), \\ 0 & \text{otherwise,} \end{cases} \tag{2.13}$$

and similarly for the displacement $u(x,t)$. Expand (2.6) using these decompositions, and multiply by $\hat{\bar{S}}_+$:

$$\hat{\bar{S}}_+\hat{\bar{\sigma}}_+ + \hat{\bar{S}}_+\hat{\bar{\sigma}}_- = \hat{\bar{T}}_-\hat{\bar{u}}_+ + \hat{\bar{T}}_-\hat{\bar{u}}_-. \tag{2.14}$$

Assuming the properties (2.10) and (2.11), the convolution $T_- * *u_-$ vanishes for $x > a(t)$:

$$T_- * *u_-(x,t) \tag{2.15}$$

$$= \int_{-\infty}^{\infty} dr \int_0^t T_-(r,\tau)u_-(x-r,t-\tau)d\tau \tag{2.16}$$

$$= \int_0^t d\tau \int_{-c\tau}^{x-a(t-\tau)} T_-(r,\tau)u_-(x-r,t-\tau)dr. \tag{2.17}$$

For $x > a(t)$, the two functions in this double integral have no region of support in common, so that the integral vanishes. Similarly, the convolution $S_+ * *\sigma_+$ vanishes for $x < a(t)$. Apply Fourier and Laplace transformations again to the relations

$$T_- * *u_-(x,t) = H(a(t)-x)(S + * * \sigma_- - T_- * *u_+), \tag{2.18}$$

$$S_+ * *\sigma_+(x,t) = -H(x-a(t))(S + * * \sigma_- - T_- * *u_+), \tag{2.19}$$

multiply by $\hat{\tilde{S}}_-$ and $\hat{\tilde{T}}_+$, respectively, and invert the transformations to obtain

$$u_- = S_- * *[H(a(t)-x)(S_+ * *\sigma_- - T_- * *u_+)], \tag{2.20}$$

$$\sigma_+ = -T_+ * *[H(x-a(t))(S_+ * *\sigma_- - T_- * *u_+)]. \tag{2.21}$$

Note that these relations give the unknown quantities in terms of the known quantities (from the boundary conditions) for a semi-infinite crack. These formulas first appeared in [10] (and hold for all three fracture modes).

Let's look more closely at these maps for an antiplane shear crack in elastic material. For this case we decompose the transfer map into two square root functions in the complex plane:

$$\hat{\tilde{T}}_+(p,s) = \sqrt{-ip + s/c}, \tag{2.22}$$

$$\hat{\tilde{T}}_-(p,s) = -\mu\sqrt{ip + s/c}, \tag{2.23}$$

where the branch cuts for the square roots are taken along the negative imaginary axis for (2.22) and along the positive imaginary axis for (2.23). The functions $S_\pm$ are straightforward to find and have the desired support properties:

$$S_+(x,t) = \delta\left(t - \frac{x}{c}\right)\frac{H(x)}{\sqrt{\pi x}}, \tag{2.24}$$

$$S_-(x,t) = -\mu\delta\left(t + \frac{x}{c}\right)\frac{H(-x)}{\sqrt{-\pi x}}, \tag{2.25}$$

where $\delta(\tau)$ denotes the Dirac delta function and $H(x)$ denotes the Heaviside function. Also note that

$$\mathcal{F}^{-1}\circ\mathcal{L}^{-1}\left[\left(-ip + \frac{s}{c}\right)\hat{\tilde{u}}(p,s)\right](x,t) = \frac{1}{c}\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}, \tag{2.26}$$

$$\mathcal{F}^{-1}\circ\mathcal{L}^{-1}\left[\left(ip + \frac{s}{c}\right)\hat{\tilde{u}}(p,s)\right](x,t) = \frac{1}{c}\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x}. \tag{2.27}$$

Substituting these functions and the boundary conditions into the relations (2.20) and (2.21) leads to expressions for the crack face displacement and stress in front of the crack that depend only on the crack face load $\Lambda(x,t)$ and the crack tip path $a(t)$:

$$u_-(\eta,\xi) = -\frac{1}{2\pi}\int_{\max\{-\eta, b_0\circ b_1^{-1}(\eta)\}}^{\xi}\frac{dq}{\sqrt{\xi - q}}\int_{\max\{-q,0\}}^{\eta}\Lambda(r,q)\frac{dr}{\sqrt{\eta - r}}, \tag{2.28}$$

$$\sigma_+(\eta,\xi) = \frac{1}{\pi}\frac{\partial}{\partial\eta}\int_{-\xi}^{b_1\circ b_0^{-1}(\xi)}\frac{dr}{\sqrt{\eta - r}}\int_{-\xi}^{r}\Lambda(s,\xi)\frac{ds}{\sqrt{r - s}}. \tag{2.29}$$

Here we have used characteristic coordinates $\eta = t + x/c$ and $\xi = t - x/c$ in place of the original (x,t) coordinates, and retarded and advanced time scales $b_0(t)$ and $b_1(t)$, respectively, defined via

$$b_0(t) = t - a(t)/c, \qquad b_1(t) = t + a(t)/c. \tag{2.30}$$

(Since we impose subsonic crack propagation, $0 \leq \dot{a}(t) < c$, the functions $b_0(t)$ and $b_1(t)$ will be strictly increasing and hence invertible.) Taking the limit $x \to a(t)^+$ of (2.29) after reversing the order of integration and then performing a change of variables leads to the usual mode III stress intensity factor:

$$K_{III}(t) = -\frac{1}{\pi}\sqrt{c - \dot{a}(t)} \int_0^t \Lambda(c(r - b_0(t)), r) \frac{dr}{\sqrt{t-r}}. \tag{2.31}$$

See [15], [13], [8], and [7] for more details on this solution method for antiplane shear cracks in elastic material.

3. General framework for viscoelastic case. We will consider a dynamically accelerating, semi-infinite, antiplane shear crack for a general infinite homogeneous and isotropic linearly viscoelastic body. The governing equations for this viscoelastic fracture problem are

$$\rho\ddot{u}(x,y,t) = \mu * \triangle u \quad \text{and} \quad \sigma(x,y,t) = \frac{\partial}{\partial y}(\mu * du), \tag{3.1}$$

where $\mu(t)$ is the shear relaxation function, $u(x,y,t)$ is the out-of-plane displacement u_3, $\sigma(x,y,t)$ is the shear stress σ_{23}, and the convolutions are with respect to t. Let $a(t)$ be the crack tip position at time t, as in Figure 2.1. The initial and boundary conditions for this problem are

$$u(x,y,0) = 0 = \dot{u}(x,y,0), \tag{3.2}$$

$$u(x,0,t) = 0 \quad \text{for } x > a(t), \tag{3.3}$$

$$\sigma(x,0,t) = \Lambda(x,t) \quad \text{for } x < a(t), \tag{3.4}$$

$$u(x,y,t) \to 0 \quad \text{as } |y| \to \infty. \tag{3.5}$$

See [13] for the details of the derivation of the transfer map corresponding to this problem.

Finding a solution method for a crack in a viscoelastic material is, of course, complicated by the dependence on time histories of the displacement and stress. The decomposition of the transfer function must be carefully chosen to result in useful support properties. In general, we will assume that the transfer function has form

$$\hat{\tilde{T}}(p,s) = -\sqrt{p^2 + s^2/\tilde{c}(s)^2} \tag{3.6}$$

for the transfer map

$$\hat{\tilde{\sigma}}(p,s) = \hat{\tilde{T}}(p,s)\hat{\tilde{D}}(p,s), \tag{3.7}$$

where

$$\hat{\tilde{D}}(p,s) = \tilde{\mu}(s)\hat{\tilde{u}}(p,s), \tag{3.8}$$

and $\tilde{\mu}(s) = \mu_0 + \int_0^\infty e^{-ts}\dot{\mu}(t)dt$ is the s-multiplied Laplace transform of the shear relaxation function. A natural choice for a decomposition $\hat{\tilde{T}}(p,s) = \hat{\tilde{T}}_+(p,s)\hat{\tilde{T}}_-(p,s)$ (with $\hat{\tilde{S}}_\pm = 1/\hat{\tilde{T}}_\pm$ as before) is

$$\hat{\tilde{T}}_+(p,s) = \sqrt{-ip + s/\tilde{c}(s)}, \tag{3.9}$$

$$\hat{\tilde{T}}_-(p,s) = -\sqrt{ip + s/\tilde{c}(s)}, \tag{3.10}$$

with the branch cut for the square roots taken along the negative real axis. The complication in finding the inverse Laplace transform lies in the nonconstant transverse wave speed function $\tilde{c}(s) = \sqrt{\tilde{\mu}(s)/\rho}$.

The key to analyzing the transfer function is to examine the inverse Laplace transform of $\exp[-|x|s/\tilde{c}(s)]$, a crucial component of the Dirichlet-to-Neumann map. Observe that since

$$\hat{\tilde{S}}_{\pm}(p,s) = \frac{\pm 1}{\sqrt{\mp ip + s/\tilde{c}(s)}}, \tag{3.11}$$

the functions $S_{\pm}(x,t)$ will be found via

$$S_{+}(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ipx}dp \frac{1}{2\pi i}\int_{\Gamma} e^{st}\frac{ds}{\sqrt{-ip + s/\tilde{c}(s)}} \tag{3.12}$$

$$= \frac{1}{2\pi i}\int_{\Gamma} e^{s(t-x/\tilde{c}(s))}ds \frac{1}{2\pi}\int_{-\infty+is/\tilde{c}(s)}^{\infty+is/\tilde{c}(s)} e^{-ipx}\frac{dp}{\sqrt{-ip}} \tag{3.13}$$

$$= \frac{H(x)}{\sqrt{\pi x}}\mathcal{L}^{-1}\big[e^{-xs/\tilde{c}(s)}\big], \tag{3.14}$$

with $S_{-}(x,t) = -S_{+}(-x,t)$. By definition of the inverse Laplace transform, the contour Γ is a vertical line placed to the right of any singularities. Hence this decomposition will place any poles and branch cuts in the left half of the complex plane (relative to Γ), and the inverse transform will have support $t - |x|/c > 0$.

Define $\breve{c}(s)$ via $\tilde{c}(s) = c\breve{c}(s)$, where $c = \sqrt{\mu_0/\rho}$ is the glassy (short-time) shear wave speed and $c_\infty = \sqrt{\mu_\infty/\rho}$ is the equilibrium (long-time) shear wave speed. The properties of $\breve{c}(s)$ are assumed to be as follows:

1. $\breve{c}(s) \to 1$ as $s \to \infty$.
2. $\breve{c}(s) \to \delta$ as $s \to 0$, where $\delta = c_\infty/c \le 1$.
3. $1/\breve{c}(s) = 1 + \gamma(s)$, with $s\gamma(s) = O(1)$ as $|s| \to \infty$.

(For example, the Achenbach–Chao and standard linear solid models both satisfy these conditions.) These conditions should ensure that the inverse Laplace transform in (3.14) is well defined. To see this, decompose the exponential function into two pieces:

$$e^{-\frac{|x|s}{\tilde{c}(s)}} = e^{-\frac{|x|\tilde{\gamma}}{c}}e^{-\frac{|x|s}{c}} + e^{-\frac{|x|\tilde{\gamma}}{c}}e^{-\frac{|x|s}{c}}\big(e^{-\frac{|x|}{c}(s\gamma(s)-\tilde{\gamma})} - 1\big), \tag{3.15}$$

where $\tilde{\gamma} = \lim_{|s|\to\infty} s\gamma(s)$. In terms of the shear relaxation function $\mu(t) = \mu_0 m(t/\tau)$, where $m(t)$ is nondimensional, we have $\tilde{\gamma} = |m'(0)|/2$, which equals $(1-\delta)/\tau$ for the Achenbach–Chao model and $(1-\delta^2)/2\tau$ for standard linear solid model.

The first term on the right-hand side of (3.15) will be similar to the elastic case, while the second term will lead to a dependence on the time history.

$$\mathcal{L}^{-1}\left[e^{-\frac{|x|\tilde{\gamma}}{c}}e^{-\frac{|x|s}{c}}\right] = e^{-\frac{|x|\tilde{\gamma}}{c}}\delta(t-|x|/c), \tag{3.16}$$

while the function $F(x,t)$ is defined via

$$\mathcal{L}^{-1}\left[e^{-\frac{|x|s}{c}}\big(e^{-\frac{|x|}{c}(s\gamma(s)-\tilde{\gamma})} - 1\big)\right](t) = \mathcal{L}^{-1}\big[e^{-\frac{|x|}{c}(s\gamma(s)-\tilde{\gamma})} - 1\big]\left(t - \frac{|x|}{c}\right) \tag{3.17}$$

$$= H\left(t - \frac{|x|}{c}\right)F\left(|x|, t - \frac{|x|}{c}\right). \tag{3.18}$$

Under the given assumptions, the transfer functions for the viscoelastic case will take the form

$$S_+(x,t) = \frac{H(x)}{\sqrt{\pi x}} e^{-\frac{x\tilde{\gamma}}{c}} \left(\delta\left(t - \frac{x}{c}\right) + H\left(t - \frac{x}{c}\right) F\left(x, t - \frac{x}{c}\right)\right), \tag{3.19}$$

$$S_-(x,t) = -\frac{H(-x)}{\sqrt{-\pi x}} e^{\frac{x\tilde{\gamma}}{c}} \left(\delta\left(t + \frac{x}{c}\right) + H\left(t + \frac{x}{c}\right) F\left(-x, t + \frac{x}{c}\right)\right). \tag{3.20}$$

Compare these to the functions $S_\pm$ given in (2.24) and (2.25) found for the elastic case. These transfer functions result (when convolved) in operators that are the sum of an elastic-like operator plus one involving the time history. With this in mind, define the elastic-like part as

$$S_{E\pm} = \pm \frac{H(\pm x)}{\sqrt{\pm \pi x}} e^{\mp \frac{x\tilde{\gamma}}{c}} \delta(t \mp x/c) \tag{3.21}$$

(noting that this elastic-like part involves only instantaneous moduli), while the component involving the time history is

$$S_{H\pm} = \pm \frac{H(\pm x)}{\sqrt{\pm \pi x}} e^{\mp \frac{x\tilde{\gamma}}{c}} H\left(t \mp \frac{x}{c}\right) F\left(\pm x, t \mp \frac{x}{c}\right). \tag{3.22}$$

Since

$$e^{-\frac{|x|}{c}(s\gamma(s) - \tilde{\gamma})} - 1 = \sum_{n=1}^{\infty} \frac{(-|x|)^n (s\gamma(s) - \tilde{\gamma})^n}{c^n n!}, \tag{3.23}$$

we have that $F(|x|, t - |x|/c) = O(|x|)$ as $|x| \to 0$. Convolutions with these functions have the following form in characteristic coordinates $\eta = t + x/c$ and $\xi = t - x/c$:

$$S_{E+} * * f = \sqrt{\frac{c}{2\pi}} \int_{-\xi}^{\eta} e^{-\frac{\tilde{\gamma}}{2}(\eta - r)} \tilde{f}(r, \xi) \frac{dr}{\sqrt{\eta - r}}, \tag{3.24}$$

(3.25)

$$S_{H+} * * f = \sqrt{\frac{c}{2\pi}} \int_{\frac{\xi - \eta}{2}}^{\xi} dq \int_{\xi - 2q}^{\eta} F\left(\frac{c}{2}(\eta - r), \xi - q\right) \tilde{f}(r - \xi + q, q) \frac{e^{-\frac{\tilde{\gamma}}{2}(\eta - r)} dr}{\sqrt{\eta - r}}.$$

Convolutions with S_- are similar but with roles of η and ξ exchanged and with a minus sign in front.

To find $T_\pm(x,t)$, rewrite the square roots as

$$\hat{\tilde{T}}_\pm(x,t) = \pm \frac{\mp ip + s/\tilde{c}(s)}{\sqrt{\mp ip + s/\tilde{c}(s)}} \tag{3.26}$$

and observe that

$$\mp ip + \frac{s}{\tilde{c}(s)} = \mp ip + \frac{s}{c} + \frac{\tilde{\gamma}}{c} + \frac{s\gamma(s) - \tilde{\gamma}}{c}, \tag{3.27}$$

whose inverse transform is the sum of derivatives, an identity operator, and a smoothing operator, as compared to the elastic case, which involves only derivatives (2.26)

and (2.27). The expression for these functions can be simplified into the sum of an elastic-like part $T_{E\pm}$ and a time history part $T_{H\pm}$:

$$T_{E+} * * f = \sqrt{\frac{2}{c\pi}} e^{-\frac{\tilde{\gamma}}{2}\eta} \frac{\partial}{\partial \eta} \int_{-\xi}^{\eta} e^{\frac{\tilde{\gamma}}{2} r} \tilde{f}(r, \xi) \frac{dr}{\sqrt{\eta - r}}, \tag{3.28}$$

(3.29)

$$T_{H+} * * f = -\frac{1}{c} \sqrt{\frac{2}{c\pi}} \int_{\frac{\xi-\eta}{2}}^{\xi} dq \int_{\xi-2q}^{\eta} F\left(\frac{c}{2}(\eta - r), \xi - q\right) \tilde{f}(r - \xi + q, q) \frac{e^{-\frac{\tilde{\gamma}}{2}(\eta - r)} dr}{(\eta - r)^{3/2}}.$$

Observe that T_{E+} and S_{E+} are inverse Abel operators.

Convolution with T_- leads to similar expressions, with the roles of η and ξ reversed and a change of sign.

4. Construction of the maps for the viscoelastic case. Due to the dependence on time history, the supports of $S_\pm$ will be $0 \le x \le ct$ and $-ct \le x \le 0$, as seen in the previous section, so that conditions (2.10) and (2.11) from the elastic case are not satisfied. This fact leads us to decompose the displacement $u(x,t)$ into three parts, with supports separated in the xt-plane by the crack tip path $x = a(t)$ and the line $x = a_0$:

$$u_+(x,t) = \begin{cases} u(x,t) & \text{if } x > a(t), \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

$$u_c(x,t) = \begin{cases} u(x,t) & \text{if } a_0 < x < a(t), \\ 0 & \text{otherwise,} \end{cases} \tag{4.2}$$

$$u_-(x,t) = \begin{cases} u(x,t) & \text{if } x < a_0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

The corresponding functions $D_c(x,t)$ and $D_-(x,t)$ as defined by (3.8) will have the same supports as $u_c(x,t)$ and $u_-(x,t)$, respectively, and $D_+(x,t)$ will be identically zero since we assume that $u_+(x,t) \equiv 0$. The stress $\sigma(x,t)$ will be split as in (2.12) and (2.13), with $\sigma_+(x,t)$ having support in the region $x > a(t)$, and $\sigma_-(x,t) = \Lambda(x,t)$, the known crack face loading, for $x < a(t)$. See Figure 4.1.

Expand (3.7) using these decompositions and multiply by $\hat{\tilde{S}}_+$:

$$\hat{\tilde{S}}_+ \hat{\tilde{\sigma}}_+ + \hat{\tilde{S}}_+ \hat{\tilde{\sigma}}_- = \hat{\tilde{T}}_- \hat{\tilde{D}}_c + \hat{\tilde{T}}_- \hat{\tilde{D}}_-. \tag{4.4}$$

The convolution $S_+ * * \sigma_+$ has support $x > a_0$, while the convolution $S_+ * * \sigma_-$ has support $x < a_0 + ct$. The convolution $T_- * * D_c$ has support $a_0 - ct < x < a(t)$, while the convolution $T_- * * D_-$ has support $x < a_0$.

Apply Fourier and Laplace transformations again to the following three equations, where Heaviside functions have been included to indicate intervals of support:

$$S_+ * * \sigma_+(x,t) = H(x - a_0)(-S_+ * * \sigma_- + T_- * * D_c), \tag{4.5}$$

$$T_- * * D_c(x,t) = H(a(t) - x)(S_+ * * \sigma_+ + S_+ * * \sigma_- - T_- * * D_-), \tag{4.6}$$

$$T_- * * D_-(x,t) = H(a_0 - x)(S_+ * * \sigma_- - T_- * * D_c). \tag{4.7}$$

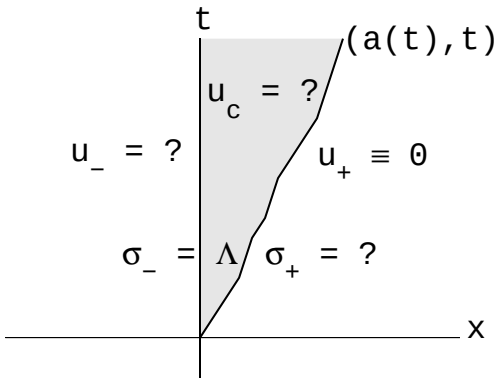


FIG. 4.1. *Division of the upper half of the xt-plane into three regions, separated by the crack tip path $x = a(t)$ and the line $x = a_0$ (the t-axis), for a semi-infinite, dynamically accelerating crack.*

Multiply the transform of (4.5) by $\hat{\tilde{T}}_+$, and the transforms of (4.6) and (4.7) by $\hat{\tilde{S}}_-$, then invert the transformations to obtain

$$\sigma_+(x,t) = T_+ * * \Big[H(x-a_0)\big[H(a(t)-x)(T_- * * D_c) - H(a_0+ct-x)(S_+ * * \Lambda)\big]\Big], \tag{4.8}$$

$$D_c(x,t) = S_- * * \Big[H(a(t)-x)H(x-a_0+ct)\big[H(x-a_0)(S_+ * * \sigma_+) + H(a_0+ct-x)(S_+ * * \Lambda) - H(a_0-x)(T_- * * D_-)\big]\Big], \tag{4.9}$$

$$D_-(x,t) = S_- * * \Big[H(a_0-x)\big[H(a_0+ct-x)(S_+ * * \Lambda) - H(a(t)-x)H(x-a_0+ct)(T_- * * D_c)\big]\Big]. \tag{4.10}$$

Unfortunately, D_c is unknown and cannot directly be solved using the above maps. However, if we can derive a formula for D_c that involves only the known loading $\Lambda(x,t)$, then we can use these maps to find the remaining crack face displacement D_- and the stress σ_+ in front of the crack tip.

The function $\sigma_+(x,t)$ vanishes for $x < a(t)$, so the right-hand side of (4.8) equals zero if $x < a(t)$. Define a new (completely known) function $t_0(x,t)$ with support $a_0 < x < a(t)$ as follows (analogous to what is done in [2]),

$$t_0(x,t) = H(a(t)-x)H(x-a_0)\left[T_+ * * \big[H(x-a_0)(S_+ * * \Lambda)\big]\right], \tag{4.11}$$

and then use (4.8) to form a new equation:

$$H(a(t)-x)\left[T_+ * * \big[H(x-a_0)H(a(t)-x)(T_- * * D_c)\big]\right] = t_0(x,t). \tag{4.12}$$

We could solve this new differo-integral equation for D_c using numerical methods, leading to a complete numerical solution for all boundary data by combining with (4.8) and (4.10). We will instead calculate asymptotic behavior near the crack tip; in particular, we wish to determine the stress intensity factor (SIF) $K(t)$, defined via $K(t) = \lim_{x\to a(t)^+}[\sqrt{x-a(t)}\sigma_+(x,t)]$.

We will consider two cases: a stationary crack $a(t) \equiv a_0$ and a moving crack $a(t) > a_0$. Start by assuming that the crack tip has not yet begun to move: $a(t) = a_0$. Since D_c has no support in this case, (4.8) reduces to (the dots refer to bounded terms not contributing to the SIF)

$$\sigma_+(x,t) = -T_+ * *\big[H(x-a_0)H(a_0+ct-x)(S_+ * *\Lambda)\big] \tag{4.13}$$

$$= -\left(\frac{\partial}{\partial\eta} + \frac{\tilde{\gamma}}{c}\right)[S_{E+} * *[H(x-a_0)H(a_0+ct-x)(S_+ * *\Lambda)]] + \cdots . \tag{4.14}$$

Multiplying by $\sqrt{x-a_0}$ and taking the limit as $x \to a_0^+$ yields the SIF corresponding to a stationary crack tip $a(t) \equiv a_0$:

$$K_{\text{stationary}}(t) = -\frac{\sqrt{c}}{\pi}\int_0^t \Lambda\left(c\left(r-t+\frac{a_0}{c}\right),r\right)\frac{e^{-\tilde{\gamma}(t-r)}dr}{\sqrt{t-r}} - \frac{\sqrt{c}}{\pi}\int_0^t du\int_0^u \Lambda\left(c\left(r-u+\frac{a_0}{c}\right),r\right)F(c(u-r),t-u)\frac{e^{-\tilde{\gamma}(u-r)}dr}{\sqrt{u-r}}. \tag{4.15}$$

Now consider the case where the crack has progressed beyond its initial position: $a(t) > a_0$. The term directly involving $\Lambda(x,t)$ in (4.8) will have the form of a bounded integral divided by $\sqrt{x-a_0}$, so that this term has no contribution to the SIF if $a(t) > a_0$. We need consider only the term involving $D_c(x,t)$ in (4.8) to derive the SIF for this case. Therefore, in order to determine the SIF for a moving crack tip, we will first derive an expression for the SIF in terms of the displacement $D_c(x,t)$.

Using (4.8) and observing that for a moving crack tip there is no singularity in the term involving $\Lambda(x,t)$ or the term involving T_{H+}, we obtain

$$K(t) = \lim_{x\to a(t)^+}\big[\sigma_+(x,t)\sqrt{x-a(t)}\big] \tag{4.16}$$

$$= \lim_{x\to a(t)^+}\sqrt{x-a(t)}\big[T_+ * *[H(x-a_0)H(a(t)-x)(T_- * *D_c)]\big] \tag{4.17}$$

$$= \sqrt{\frac{c-\dot{a}(t)}{c\pi}}\lim_{x\to a(t)^+} e^{-\frac{\tilde{\gamma}}{2}\eta}\sqrt{\eta - b_1\circ b_0^{-1}(\xi)} \cdot \frac{\partial}{\partial\eta}\int_{\xi+\frac{2a_0}{2}}^{b_1\circ b_0^{-1}(\xi)} e^{\frac{\tilde{\gamma}}{2}r}(T_- * *D_c)(r,\xi)\frac{dr}{\sqrt{\eta-r}} \tag{4.18}$$

$$= -\sqrt{\frac{c-\dot{a}(t)}{c\pi}}\lim_{x\to a(t)^-}(T_- * *D_c)(x,t). \tag{4.19}$$

In the above limit we have used the fact that

$$\lim_{x\to a(t)^+}\frac{x-a(t)}{\eta - b_1\circ b_0^{-1}(\xi)} = \frac{1}{2}(c-\dot{a}(t)). \tag{4.20}$$

Next we will make use of the supports of the various convolutions and (4.5)–(4.6) to construct an expression for $T_- * *D_c$. Using (4.6) and the facts that $T_- * *D_-$

vanishes for $x > a_0$ and $S_{E+} * *\sigma_+$ vanishes for $x < a(t)$, we have that

$$\lim_{x\to a(t)^-}(T_- * *D_c)(x,t) = \lim_{x\to a(t)^-}(S_+ * *\sigma_+ + S_+ * *\Lambda)(x,t) \tag{4.21}$$

$$= \lim_{x\to a(t)^-}(S_{H+} * *\sigma_+ + S_+ * *\Lambda)(x,t) \tag{4.22}$$

$$= (S_{H+} * *\sigma_+)(a(t),t) + (S_+ * *\Lambda)(a(t),t), \tag{4.23}$$

where

$$(S_+ * *\Lambda)(a(t),t) = \sqrt{\frac{c}{\pi}}\int_0^t e^{-\tilde{\gamma}(t-r)}\Lambda(c(r-b_0(t)),r)\frac{dr}{\sqrt{t-r}} \tag{4.24}$$
$$+\sqrt{\frac{c}{\pi}}\int_0^t dr\int_{b_0(r)}^{b_0(t)}\Lambda(c(r-q),r)F(c(q-r)+a(t),b_0(t)-q)\frac{e^{-\tilde{\gamma}(q-r+a(t)/c)}dq}{\sqrt{q-r+a(t)/c}}.$$

Now apply (4.8) to derive a new expression for $S_{H+} * *\sigma_+$ on the region $a_0 < x < a(t)$:

$$S_{H+} * *\sigma_+ = S_{H+} * *T_+ * *\big[H(x-a_0)H(a(t)-x)(T_- * *D_c) \tag{4.25}$$
$$- H(x-a_0)H(a_0+ct-x)(S_+ * *\Lambda)\big],$$

and then apply (4.6) as well as the fact that the support of $T_- * *D_-$ is $x < a_0$:

$$S_{H+} * *\sigma_+ = S_{H+} * *T_+ * *\big[H(x-a_0)H(a(t)-x)(S_+ * *\sigma_+ + S_+ * *\Lambda) \tag{4.26}$$
$$- H(x-a_0)H(a_0+ct-x)(S_+ * *\Lambda)\big].$$

After simplifying, we now have an operator equation for $S_{H+} * *\sigma_+$ for $a_0 < x < a(t)$:

$$S_{H+} * *\sigma_+ = S_{H+} * *T_+ * *\big[H(x-a_0)H(a(t)-x)(S_{H+} * *\sigma_+) \tag{4.27}$$
$$- H(x-a(t))H(a_0+ct-x)(S_+ * *\Lambda)\big].$$

To simplify calculations, the operator in (4.27) can be reduced to a much simpler form, using the facts that S_+ and T_+ are inverse operators when convolved, as are S_{E+} and T_{E+}:

$$S_{H+} * *T_+ * *f = f - S_{E+} * *T_+ * *f \tag{4.28}$$

$$= f - (S_{E+} * *T_{E+} * *f + S_{E+} * *T_{H+} * *f) \tag{4.29}$$

$$= -S_{E+} * *T_{H+} * *f. \tag{4.30}$$

Observe that this new expression involves no derivatives and fewer integrations than the original form.

Putting together (4.19), (4.23), (4.27), and (4.30), we can derive a closed-form expression for the SIF for a moving crack tip that involves only the known crack face loading $\Lambda(x,t)$:

$$K(t) = -\sqrt{\frac{c-\dot{a}(t)}{c\pi}}G(a(t)^+,t) \tag{4.31}$$
$$-\sqrt{\frac{c-\dot{a}(t)}{c\pi}}\sum_{n=0}^{\infty}\big[S_{E+} * *T_{H+} * *\big[H(x-a_0)H(a(t)-x)(-S_{E+} * *T_{H+} * *)\big]^n G\big](a(t),t),$$

where we have defined $G(x,t)$ to be

$$G(x,t) = H(x-a(t))H(a_0+ct-x)(S_+ **\Lambda)(x,t). \tag{4.32}$$

From (4.30), we observe that the operator in (4.27) is second kind Volterra, which implies that the Neumann expansion in (4.31) will converge. Alternatively, we could solve the second kind Volterra integral equation

$$J(x,t) + (S_{E+} **T_{H+} **J)(x,t) = (S_{E+} **T_{H+} **G)(x,t) \tag{4.33}$$

for the function $J(x,t) = H(x-a_0)H(a(t)-x)(S_{H+}**\sigma_+)$ and then use the following expression for the SIF:

$$K(t) = -\sqrt{\frac{c-\dot a(t)}{c\pi}}G(a(t)^+,t) - \sqrt{\frac{c-\dot a(t)}{c\pi}}J(a(t)^-,t). \tag{4.34}$$

Also note that the expression in (4.31) agrees with (4.15) in the case that $\dot a(t) = 0$ and hence is the general expression for the viscoelastic SIF; compare to the elastic SIF given by (2.31).

In the next two sections we will apply this solution method to two common models of linear viscoelasticity and, in particular, state the time history function for each. Numerical simulations and comparison to the steady-state solution will be done in a forthcoming paper.

5. The Achenbach–Chao model. The Achenbach–Chao viscoelastic model (introduced by Achenbach and Chao [1] as an approximation to the standard linear solid) is particularly convenient to work with, as it avoids the need to integrate around branch cuts when finding $S_\pm(x,t)$ (as must be done for the standard linear solid, for example). The s-multiplied Laplace transforms of the relaxation function $\mu(t)$ and wave speed $c(t)$ for an Achenbach–Chao viscoelastic material are defined to be

$$\tilde\mu(s) = \mu_0\left[\frac{\delta+\tau s}{1+\tau s}\right]^2, \tag{5.1}$$

$$\tilde c(s) = c\frac{\delta+\tau s}{1+\tau s}. \tag{5.2}$$

The modified displacement function $D(x,t)$ has the following form for the Achenbach–Chao model:

$$D(x,t) = \mu_0 u(x,t) + \mu_0\int_0^t u(x,t-u)\left[\delta^2 + (1-\delta^2)\left(1-\frac{1-\delta}{1+\delta}\frac{u}{\tau}\right)e^{-u/\tau}\right]du. \tag{5.3}$$

The time history function for the Achenbach–Chao model is

$$F(x,\xi) = e^{-\delta\xi/\tau}\sum_{k=0}^{\infty}\frac{\xi^k\big((1-\delta)\delta x/(c\tau^2)\big)^{k+1}}{k!(k+1)!} \tag{5.4}$$

$$= e^{-\delta\xi/\tau}\sqrt{\frac{\delta(1-\delta)x/c}{\xi\tau^2}}I_1\left(2\sqrt{\frac{\delta(1-\delta)}{\tau^2}\frac{x}{c}\xi}\right). \tag{5.5}$$

Here $I_1(z)$ is the modified Bessel function of the first kind. (Note that $\lim_{t\to|x|/c}F(|x|,t-|x|/c) = \delta(1-\delta)|x|/c\tau^2$, and thus $F(|x|,t-|x|/c)/\sqrt{|x|}$ is not singular.)

The SIF for the Achenbach–Chao model reduces in the case $\delta = 1$ to the elastic SIF given in (2.31), and in the case $\delta = 0$ to the expression found in [2] for the SIF of the Achenbach–Chao approximation to a Maxwell fluid.

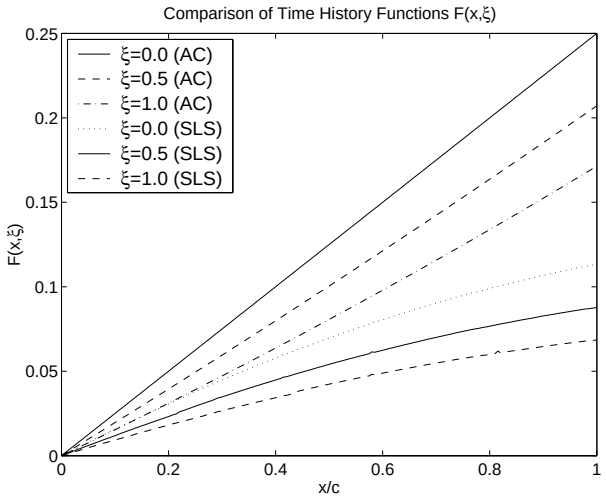


FIG. 6.1. *Comparison of time history functions $F(x,\xi)$ for the Achenbach–Chao (AC) and standard linear solid (SLS) models, for parameter values $\delta = 0.5$ and $\tau = 1$.*

6. Standard linear solid. As another example, we consider the standard linear solid viscoelastic model. The s-multiplied Laplace transforms of the relaxation function $\mu(t)$ and wave speed $c(t)$ are defined to be

$$\tilde{\mu}(s) = \mu_0 \left[\frac{\delta^2 + s}{1 + s} \right], \tag{6.1}$$

$$\tilde{c}(s) = c \sqrt{\frac{\delta^2 + s}{1 + s}}. \tag{6.2}$$

The time history function for the standard linear solid model is

$$F(x,\xi) = \frac{1}{\pi} e^{\frac{1-\delta^2}{2}\frac{|x|}{c}} \int_{\delta^2}^{1} e^{-s\xi} e^{-\frac{s|x|}{c}} \sin\left(\frac{|x|}{c} s \sqrt{\frac{1-s}{s-\delta^2}} \right) ds. \tag{6.3}$$

The time history functions for the two models are compared in Figure 6.1.

7. Concluding remarks. The key to this solution method is the observation that, after applying Laplace and Fourier transforms and taking the limit as $y \to 0^+$, the governing equations to the dynamic fracture problem reduce to a simple transfer map. In this paper we have focused on the mode III (antiplane shear) case, but the transfer maps corresponding to the mode I and II elastic and viscoelastic fracture problems share the same structure:

$$\hat{\tilde{\sigma}}(p,s) = \hat{\tilde{T}}(p,s) \hat{\tilde{u}}(p,s), \tag{7.1}$$

where $\sigma(x,t)$ is the appropriate component of the stress tensor and $u(x,t)$ is the appropriate displacement along the boundary $y = 0^+$. The elastic mode III transfer function $T(p,s)$ is quite simple,

$$\hat{\tilde{T}}_{III}(p,s) = -\mu \sqrt{p^2 + s^2/c_s^2}, \tag{7.2}$$

while that for the elastic mode I case is more complicated, requiring a more complicated decomposition:

$$\hat{\tilde{T}}_I(p,s) = \frac{\mu^2 R(p,s)}{\rho s^2 \sqrt{p^2 + s^2/c_L^2}}, \tag{7.3}$$

where $R(p,s)$ is the Rayleigh wave function. (The mode II transfer function is similar.) See [9] for the details of the mode I elastic case of subsonic fracture. For faster fracture speeds, this method could still be carried out by choosing appropriate decompositions as described in [12], although the problem of how to jump across speed barriers remains unaddressed.

The difficulty in applying this method to a crack in a viscoelastic material is due to the dependence on the time history. Based on the analysis of section 3, for any fracture mode the viscoelastic transfer function should lead to expressions for the stress and displacement that involve an elastic-like part T_E plus a time history part T_H, which greatly increases the computational difficulty in obtaining, for example, the crack face displacement, as compared to in the elastic case. In principle, however, one can carry out an analysis similar to the one given in this paper for an opening mode crack in a linear viscoelastic material, though the details will be considerably more complicated.

REFERENCES

[1] J. Achenbach and C. Chao, *A three-parameter viscoelastic model particularly suited for dynamic problems*, J. Mech. Phys. Solids, 10 (1962), pp. 245–252.

[2] J. P. Bourne and J. R. Walton, *On a dynamically accelerating crack in an Achenbach-Chao viscoelastic solid*, Internat. J. Engrg. Sci., 31 (1993), pp. 569–581.

[3] J. Eshelby, *The elastic field of a crack extending non-uniformly under general anti-plane loading*, J. Mech. Phys. Solids, 17 (1969), pp. 177–199.

[4] L. Freund, *Dynamic Fracture Mechanics*, Cambridge University Press, Cambridge, UK, 1990.

[5] G. Goleniewski, *Equations of Motion for Viscoelastic Moving Crack Problems*, Ph.D. dissertation, University of Bath, Bath, UK, 1988.

[6] B. Kostrov, *Unsteady propagation of longitudinal shear cracks*, Appl. Math. Mech. (Prikl. Mat. Mekh.), 30 (1966), pp. 1041–1049.

[7] T. Leise and J. Walton, *A general method for solving dynamically accelerating multiple colinear cracks*, Int. J. Fracture, 111 (2001), pp. 1–16.

[8] T. L. Leise and J. R. Walton, *Dynamically accelerating cracks.* II. *A finite length mode* III *crack in elastic material*, Quart. Appl. Math., 59 (2001), pp. 601–614.

[9] V. Saraikin and L. Slepyan, *Plane problem of the dynamics of a crack in an elastic solid*, Mechanics of Solids, 14 (1979), pp. 46–62.

[10] L. Slepyan, *Approximate model of crack dynamics*, in Dynamics of Continuous Media, Issue XIX–XX, Institute of Hydrodynamics of the Academy of Science, Novosibirsk, 1974, pp. 101–109.

[11] L. I. Slepyan, *Models and Phenomena in Fracture Mechanics*, Springer-Verlag, New York, 2002.

[12] L. I. Slepyan and A. L. Fishkov, *On the problem of crack spreading with intersonic velocity*, Dokl. Akad. Nauk SSSR, 261 (1981), pp. 1316–1319.

[13] J. Walton, *Dynamic viscoelastic fracture*, in Crack and Contact Problems for Viscoelastic Bodies, G. Graham and J. Walton, eds., CISM Courses and Lectures 356, Springer-Verlag, New York, 1995, pp. 259–311.

[14] J. Walton, *On a dynamically accelerating Dugdale-zone in elastic and viscoelastic material*, J. Mech. Phys. Solids, 44 (1996), pp. 1353–1370.

[15] J. R. Walton and J. M. Herrmann, *A new method for solving dynamically accelerating crack problems.* I. *The case of a semi-infinite mode* III *crack in elastic material revisited*, Quart. Appl. Math., 50 (1992), pp. 373–387.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 108–126

FAST OPTIMAL DESIGN OF SEMICONDUCTOR DEVICES*

MARTIN BURGER† AND RENÉ PINNAU‡

Abstract. This paper presents a new approach to the design of semiconductor devices, which leads to fast optimization methods whose numerical effort is of the same order as a single forward simulation of the underlying model, the *stationary drift-diffusion system*. The design goal we investigate is to increase the outflow current on a contact for fixed applied voltage; the natural design variable is the doping profile.

By reinterpreting the doping profile as a state variable and the electrostatic potential as the new design variable, we obtain a simpler optimization problem, whose Karush–Kuhn–Tucker conditions partially decouple. This property allows us to construct efficient iterative optimization algorithms, which avoid solving the fully coupled drift-diffusion system, and need only solves of the continuity equations and their adjoints. The efficiency and success of the new approach is demonstrated in several numerical examples.

Key words. semiconductor design, drift-diffusion, optimal control, dopant profiling

AMS subject classifications. 35J50, 49J20, 49K20

DOI. 10.1137/S0036139902420560

1. Introduction.

Optimal design and characterization of semiconductor devices is a field of growing interest in recent years, in engineering (cf., e.g., [5, 6, 13, 14, 18, 21, 22, 23]) as well as in the applied mathematics community (cf., e.g., [2, 3, 7, 8, 9, 10, 11]). A major objective in the optimal design of devices is to improve the current flow over some contact by modifying the device doping profile, which enters as a source term in the mathematical model used for semiconductor devices, the so-called *drift-diffusion system*.

The stationary drift-diffusion system in physical variables (cf. [15, 24]) consists of nonlinear elliptic equations for the electrostatic potential V, the electron density n, and the hole density p, in a bounded domain $\Omega \subset \mathbb{R}^N$, $N \leq 3$:

$$\begin{aligned} \operatorname{div}(\epsilon_s \nabla V) &= q(n - p - C) && \text{in } \Omega, \\ \operatorname{div}(D_n \nabla n - \mu_n n \nabla V) &= 0 && \text{in } \Omega, \\ \operatorname{div}(D_p \nabla p + \mu_p p \nabla V) &= 0 && \text{in } \Omega, \end{aligned}$$

where ϵ_s denotes the semiconductor permittivity, q the elementary charge, μ_n and μ_p are the electron and hole mobilities, and D_n and D_p are the electron and hole diffusion coefficients, respectively. This system is supplemented by homogeneous Neumann boundary conditions on a part $\partial\Omega_N$ of the boundary, modelling the insulating parts of the boundary, and Dirichlet conditions on the remaining part, which models the

*Received by the editors December 31, 2002; accepted for publication (in revised form) April 30, 2003; published electronically October 14, 2003. This research was supported by the Austrian Science Foundation FWF and project SFB F 013 / 08 and by the European Union under research network *Hyperbolic and Kinetic Equations*, contract HPRN-CT-2002-00282.

http://www.siam.org/journals/siap/64-1/42056.html

†Institut für Industriemathematik, Johannes Kepler Universität Linz, Altenbergerstr. 69, A-4040 Linz, Austria (burger@indmath.uni-linz.ac.at).

‡Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, D-64289 Darmstadt, Germany (pinnau@mathematik.tu-darmstadt.de).

Ohmic contacts of the device:

$$V(x) = V_D(x) = U(x) + V_{bi}(x) = U(x) + U_T \ln\left(\frac{n_D(x)}{n_i}\right) \qquad \text{on } \partial\Omega_D,$$

$$n(x) = n_D(x) = \frac{1}{2}\left(C(x) + \sqrt{C(x)^2 + 4n_i^2}\right) \qquad \text{on } \partial\Omega_D,$$

$$p(x) = p_D(x) = \frac{1}{2}\left(-C(x) + \sqrt{C(x)^2 + 4n_i^2}\right) \qquad \text{on } \partial\Omega_D.$$

Here n_i is the intrinsic density, U_T the thermal voltage, and U the applied biasing voltage.

Under usual conditions, the mobilities and diffusion coefficients are related by Einstein's relation, i.e., $D_{n/p} = \mu_{n/p} U_T$, which enables the transformation into the so-called *Slotboom variables* [19] defined by

$$n = n_i e^{V/U_T} u, \qquad p = n_i e^{V/U_T} v. \tag{1.1}$$

The assumption that ϵ_s and q are constant allows for the choice of an appropriate scaling, yielding the system

$$\lambda^2 \Delta V = (e^V u - e^{-V} v) - C \qquad \text{in } \Omega, \tag{1.2}$$

$$\operatorname{div}\left(\mu_n e^V \nabla u\right) = 0 \qquad \text{in } \Omega, \tag{1.3}$$

$$\operatorname{div}\left(\mu_p e^{-V} \nabla v\right) = 0 \qquad \text{in } \Omega, \tag{1.4}$$

where $\lambda^2 = (\epsilon_s U_T)/(q\, C_{max}\, L^2)$ is the scaled Debye length of the device (for details, see, e.g., [16]). The Dirichlet boundary conditions can be written as

$$V = V_D = U + V_{bi} \qquad \text{on } \partial\Omega_D, \tag{1.5}$$

$$u = u_D \qquad \text{on } \partial\Omega_D, \tag{1.6}$$

$$v = v_D \qquad \text{on } \partial\Omega_D, \tag{1.7}$$

where u_D and v_D are the transformations of n_D and v_D under (1.1). On the remaining part $\partial\Omega_N = \partial\Omega \setminus \partial\Omega_D$, the homogeneous Neumann conditions can be formulated on J_n and J_p, where J_n and J_p are the electron and hole current densities, which are related to the Slotboom variables by

$$J_n = \mu_n e^v \nabla u, \qquad J_p = -\mu_p e^{-V} \nabla v. \tag{1.8}$$

Hence, we have

$$\frac{\partial V}{\partial \nu} = 0 \qquad \text{on } \partial\Omega_N, \tag{1.9}$$

$$\frac{\partial u}{\partial \nu} = 0 \qquad \text{on } \partial\Omega_N, \tag{1.10}$$

$$\frac{\partial v}{\partial \nu} = 0 \qquad \text{on } \partial\Omega_N. \tag{1.11}$$

Throughout the whole paper, we shall assume that all Dirichlet boundary values V_D, u_D, and v_D are bounded in $H^{\frac{1}{2}}(\Omega) \cap L^\infty(\Omega)$, which is the basis for an existence proof of the drift-diffusion system in $(H^1(\Omega) \cap L^\infty(\Omega))^3$ (see [16]).

The objective of the optimization, the current flow over a contact Γ, is given by

$$I = \int_\Gamma J \cdot \nu = \int_\Gamma (J_n + J_p) \cdot d\nu. \tag{1.12}$$

An optimal control approach to the optimization of a functional related to the current density J or the current flow I was investigated in [10, 11], where the drift-diffusion system (1.2)–(1.11) was interpreted as an equation constraint determining the state (V, u, v). Consequently, a penalty term related to the control variable C was added to the objective in order to stabilize the system. To the penalizing problem, an iterative algorithm was applied, which as usual needed solutions of the drift-diffusion system and some adjoint system in each iteration. In this paper we investigate a completely different approach; namely, we reinterpret the potential V as the design variable, and the doping profile C as a state variable. For a given V satisfying appropriate boundary conditions, it is easy to show that the drift-diffusion system has a unique solution (u, v, C) and, moreover, that the partial differential equations (1.2)–(1.4) have a simple triangular structure in the new state variables. Corresponding to our interpretation of state and design variables we add to the objective functional a penalty term corresponding to V in order to stabilize the problem. As we shall see below, this yields a reasonably simple optimality system, from which a fast optimization algorithm can be constructed.

For the sake of simplicity and shortness of presentation, we assume that $\mu_n = \mu_p = 1$, but analogous reasoning is possible for general mobilities, even for energy dependent ones. Moreover, we ignore recombination-generation terms [19], noting that they could be incorporated into our analysis with only few modifications.

The paper is organized as follows: in section 2 we review the current state of semiconductor design and introduce our new optimization approach. Some basic analysis of the optimization problem under investigation (such as the existence of solutions and first-order optimality) is provided in section 3. Section 4 is devoted to the iterative solution of the optimal design problem, and in particular an efficient method based on a lower diagonal approximation of the Karush–Kuhn–Tucker system is introduced. Numerical results for some diodes and a metal-semiconductor field effect transistor (MESFET) are presented in section 5, and finally we give some conclusions in section 6.

2. Optimal design of semiconductor devices. In the following we discuss some basic problems in optimal semiconductor design and present a new approach for optimization problems at a single applied voltage.

Although the optimal design of semiconductor devices is of major importance in practical applications, the first systematic approaches to such optimization problems have been carried out only in the last few years (cf. [10, 11, 18, 21, 22, 23]). One of the main reasons for this late development is the computational difficulties and the complexity of such optimization problems. Even the numerical solution of the drift-diffusion system itself is not a simple task, and an optimization based on the drift-diffusion model therefore becomes quite involved. In the first optimization approaches to this problem, gradient-type methods were used, with gradient evaluations either by finite differencing (cf. [18, 21, 22, 23]) or by an adjoint method (cf. [10, 11]). Both approaches resulted in a very high numerical effort due to a large number of iterations needed. E.g., by finite differencing, around 4000 direct solutions of the drift-diffusion system were needed for the optimization of a metal-oxide-semiconductor field effect transistor (MOSFET), at a rather coarse discretization of the doping profile with 62

design parameters (cf. [18]). The adjoint approach, used in [11] for the minimization of a functional of the form

$$Q_\beta(C) := Q(n(C), p(C), V(C)) + \frac{\beta}{2}\|C - C^*\|^2 \to \min_C, \tag{2.1}$$

reduces the number of nonlinear solves, and adds few solves of an adjoint linear system. This reduces the numerical effort, but causes the need for an accurate discretization and numerical solution of the adjoint system, which is not well investigated so far.

We use a different approach, based on exchanging the interpretations of control and state between C and V. We interpret the potential V as the design variable and interpret the Poisson equation (1.2) as a state equation for the doping profile C. Consequently, we introduce a penalty dependent on $V - V^*$ rather than on $C - C^*$. As the initial guess V^* we use the one obtained from the solution of the drift-diffusion system with doping profile C^*. Since the Laplacian of $V - V^*$ is needed for the evaluation of $C - C^*$, it seems natural to use a penalty term dependent on

$$W := \Delta(V - V^*), \tag{2.2}$$

i.e., we minimize the functional

$$Q_\epsilon(u, u, V, W) := Q(u, v, V) + \frac{\epsilon}{2}\int_\Omega |W(x)|^2 \, dx, \tag{2.3}$$

subject to (2.2) and the drift-diffusion system. In order to ensure that C does not change its boundary values, W must satisfy homogeneous boundary conditions on $\partial\Omega_D$; on the remaining boundary we may use any homogeneous boundary condition. For simplicity we will carry out our analysis for

$$W = 0 \qquad \text{on } \partial\Omega. \tag{2.4}$$

In a numerical test (cf. section 5.3) we will use the boundary condition $W = 0$ on $\partial\Omega_D$, and $\frac{\partial W}{\partial \nu} = 0$ on $\partial\Omega_N$, which permits a similar analysis.

Of particular importance are functionals Q, which depend only on the values of the outflow current density on some contact Γ, i.e.,

$$Q(u, v, V) = R(J \cdot \nu|_\Gamma) = R\left(\left(e^V \frac{\partial u}{\partial \nu} - e^{-V}\frac{\partial u}{\partial \nu}\right)\Big|_\Gamma\right). \tag{2.5}$$

In [11], the functional under investigation was

$$R(J \cdot \nu|_\Gamma) = \frac{1}{2}\|(J - J^*) \cdot \nu\|^2_{H^{-\frac{1}{2}}(\Gamma)}, \tag{2.6}$$

corresponding to the objective of finding an outflow current density $J \cdot \nu$ close to a desired density $J^* \cdot \nu$. Since in most practical applications, one is rather interested in the total current flow on a contact, we rather consider the functional

$$R(J. \cdot \nu|_\Gamma) = \frac{1}{2}\left|\int_\Gamma J \cdot d\nu - I^*\right|^2 \tag{2.7}$$

(for some desired current flow I^*) as the motivation for the analysis in this paper, and we also use it for our numerical tests. We note that for one-dimensional diodes, which have been investigated in [11] and will be used for some of our numerical tests too, the above two functionals are equivalent, since the geometry of a contact corresponds to a boundary point of an interval.

3. Analysis of the optimization problem. In the following we provide some analysis of the optimization problem

$$Q_\epsilon(u,v,V,W) \to \min_{(n,p,V,W)\in\mathcal{D}_{ad}}, \tag{3.1}$$

with the admissible domain

$$\mathcal{D}_{ad} := \{(u,v,V,W) \in H^1(\Omega)^2 \times (H^1(\Omega)\cap L^\infty(\Omega)) \times L^2(\Omega) \\ \text{satisfying (1.3)–(1.11), (2.2)}\}.$$

We shall investigate the existence of minima as well as a derivation of the Karush–Kuhn–Tucker conditions, which allow us to deduce further regularity of minimizers.

3.1. Existence of a minimum. We start our analysis with a basic result on the existence of minima, for which we need two fundamental properties, namely, the weak lower semicontinuity of the objective functional and the weak closedness of the admissible domain. The weak lower semicontinuity of Q_ϵ is obviously obtained in $H^1(\Omega)^3 \times L^2(\Omega)$; the weak closedness of $\mathcal{D}_{ad}$ is obtained if $\Delta(V-V^*)$ remains in $L^2(\Omega)$. This leads us to the following result.

THEOREM 3.1 (existence). *Let $\epsilon > 0$. Then there exists a minimum*

$$(\overline{u},\overline{v},\overline{V},\overline{W}) \in H^1(\Omega)^2 \times (H^1(\Omega)\cap L^\infty(\Omega)) \times L^2(\Omega) \tag{3.2}$$

of (3.1).

Proof. Suppose that $(u^k,v^k,V^k,W^k)_{k\in\mathbb{N}}$ is a minimizing sequence; then we immediately conclude that W^k is bounded in $L^2(\Omega)$, and thus, by standard elliptic regularity, $V^k - V^*$ is uniformly bounded in $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$. Since the a priori guess V^* is in $L^\infty(\Omega)$, we obtain uniform boundedness of V^k in $L^\infty(\Omega)$. Standard energy arguments for the elliptic equations (1.3) and (1.4) consequently yield the boundedness of u^k and v^k in $H^1(\Omega)\cap L^\infty(\Omega)$. Thus, we may extract a weakly converging subsequence $(u_{k_\ell},v_{k_\ell},V_{k_\ell},W_{k_\ell})_{k_\ell\in\mathbb{N}} \in H^1(\Omega)^2 \times H^1(\Omega) \times L^2(\Omega)$, which also preserves the L^∞ bound (and such that $\Delta(V_{k_\ell}-V^*)$ converges weakly in $L^2(\Omega)$). The weak closedness of the admissible domain and the weak lower semicontinuity of the objective functional imply that the weak limit of this subsequence is a minimizer of (3.1). ∎

A direct consequence of the representation

$$C = C^* - \lambda^2 W + n - n^* - p + p^*$$

is the existence of a doping profile $C \in L^2(\Omega)$ such that (u,v,V) is a solution of the corresponding drift-diffusion system.

COROLLARY 3.2. *Let $\epsilon > 0$. Then there exists a minimum*

$$(\overline{u},\overline{v},\overline{V},\overline{W}) \in H^1(\Omega)^2 \times (H^1(\Omega)\cap L^\infty(\Omega)) \times L^2(\Omega) \tag{3.3}$$

of (3.1) *and a doping profile $\overline{C} \in L^2(\Omega)$ such that $(\overline{u},\overline{v},\overline{V})$ is a weak solution of the drift-diffusion system* (1.2)–(1.11) *with $C = \overline{C}$.*

3.2. First-order optimality. In order to derive the first-order optimality conditions, we define the Lagrangian given by

$$\mathcal{L}(u,v,V,W;\mu_1,\mu_2,\mu_3) = Q_\epsilon(u,v,V,W) + \int_\Omega \left(e^V\nabla u\cdot\nabla\mu_1 - e^{-V}\nabla v\nabla\mu_2\right)\,dx \\ + \int_\Omega \left(\nabla(V-V^*)\cdot\nabla\mu_3 + W\mu_3\right)\,dx. \tag{3.4}$$

One observes that the only nonlinear terms in the equation constraints (1.3)–(1.11), (2.2) are of the form $e^V \nabla u$ and $e^{-V}\nabla v$, and these are continuously Fréchet-differentiable in $\mathcal{D}_{ad}$ since $V \in H^1(\Omega) \cap L^\infty(\Omega)$ and $(u, v) \in H^1(\Omega)^2$. Hence, with little effort we obtain the following result.

PROPOSITION 3.3. *The Lagrangian $\mathcal{L}$ is continuously Fréchet-differentiable on $\mathcal{D}_{ad} \times H^1(\Omega)^3$.*

Each solution of the optimization problem is a saddle point of the Lagrangian, i.e., a solution of

$$\inf_{(u,v,V,W)} \sup_{(\mu_1,\mu_2,\mu_3)} \mathcal{L}(u, v, V, W; \mu_1, \mu_2, \mu_3). \tag{3.5}$$

For such saddle points we can derive the Karush–Kuhn–Tucker conditions by computing the variations of the Lagrangian $\mathcal{L}$ with respect to all primal and dual variables, which all must vanish. The variations with respect to the dual variables yield just the equality constraints, while from the variation with respect to the primal variables we deduce that

$$0 = \frac{\partial}{\partial u} Q_\epsilon(u, v, V, W)\hat{u} + \int_\Omega \left(e^V \nabla \hat{u} \cdot \nabla \mu_1\right) dx, \tag{3.6}$$

$$0 = \frac{\partial}{\partial v} Q_\epsilon(u, v, V, W)\hat{v} - \int_\Omega \left(e^V \nabla \hat{v} \cdot \nabla \mu_2\right) dx, \tag{3.7}$$

$$\begin{aligned} 0 = {} & \frac{\partial}{\partial V} Q_\epsilon(u, v, V, W)\hat{V} \\ & + \int_\Omega \left(\hat{V}\left(e^V \nabla u \cdot \nabla \mu_1 + e^{-V}\nabla v \nabla \mu_2\right) + \nabla \hat{V} \cdot \nabla \mu_3\right) dx, \end{aligned} \tag{3.8}$$

$$0 = \int_\Omega \hat{W}(\epsilon W - \mu_3) dx \tag{3.9}$$

holds for all variations $(\hat{u}, \hat{v}, \hat{V}, \hat{W}) \in H^1(\Omega)^3 \times L^2(\Omega)$.

One observes that the so-called *adjoint equations* (3.6)–(3.8) have a simple triangular structure with respect to the Lagrangian variables. Thus, the problem of proving existence and uniqueness of the Lagrangian variables $(\mu_1, \mu_2, \mu_3) \in H^1_{0,D}(\Omega)^3$ solving (3.6)–(3.8) for given primal variables (u, v, V) simplifies to analyzing subsequently three different variational problems, which turn out to be coercive in $H^1_{0,D}(\Omega)$, with

$$H^1_{0,D}(\Omega) := \{\varphi \in H^1(\Omega) \mid \varphi|_{\partial\Omega_D} = 0\}.$$

This yields another advantage of our approach with respect to the direct optimal control approach, where analyzing the adjoint problem is a difficult task, which is possible only close to thermal equilibrium (cf. [11]).

THEOREM 3.4. *Let $(u, v, V, W) \in \mathcal{D}_{ad}$ be given; then there exists a unique solution $(\mu_1, \mu_2, \mu_3) \in H^1_{0,D}(\Omega)^3$ of the variational problem* (3.6)–(3.8).

Proof. The variational problem (3.6) is of the form

$$A(\mu_3, \hat{u}) = \langle F, \hat{u}\rangle \qquad \forall\ \hat{u} \in H^1_{0,D}(\Omega),$$

with a continuous linear functional F on $H^1_{0,D}(\Omega)$ and a coercive, continuous bilinear form

$$A(u, v) = \int_\Omega e^V \nabla u \cdot \nabla v\ dx \quad \text{ on } H^1_{0,D}(\Omega)^2.$$

Thus, existence and uniqueness of μ_1 follow from the Lax–Milgram theorem. Since we can apply analogous reasoning to (3.7), we also obtain the existence and uniqueness of μ_2. Since μ_1 and μ_2 are determined by (3.6), (3.7), we may consider them as a given right-hand side in (3.8). The latter is now a scalar problem for μ_3, whose well-posedness can again be shown by a straightforward application of the Lax–Milgram theorem. Note that $L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$. ∎

Our subsequent analysis will be carried out for the important case of Q being the outflow current functional (2.5). In this case, the derivative of the functional Q is given by

$$Q'(u,v,V)(\hat{u},\hat{v},\hat{V}) = R'(J\cdot\nu|_\Gamma)\left(e^V\frac{\partial\hat{u}}{\partial\nu} - e^{-V}\frac{\partial\hat{v}}{\partial\nu} + \left(e^V\frac{\partial u}{\partial\nu} + e^{-V}\frac{\partial v}{\partial\nu}\right)\hat{V}\right), \tag{3.10}$$

and, noticing that $\hat{V} \in H^1_{0,D}(\Omega)$, we observe that the last term on the right-hand side vanishes. In the particular case of (2.7) the derivative simplifies to

$$Q'(u,v,V)(\hat{u},\hat{v},\hat{V}) = \left(\int_\Gamma J\cdot d\nu - I^*\right)\int_\Gamma\left(e^V\frac{\partial\hat{u}}{\partial\nu} - e^{-V}\frac{\partial\hat{v}}{\partial\nu}\right)ds. \tag{3.11}$$

Due to the simple form of (3.9), it seems obvious that we need to eliminate the Lagrangian variable $\mu_3 = \epsilon W$ and rewrite (3.8) as

$$0 = \int_\Omega\left(\hat{V}\left(e^V\nabla u\cdot\nabla\mu_1 + e^{-V}\nabla v\nabla\mu_2\right) + \epsilon\nabla\hat{V}\cdot\nabla W\right)dx. \tag{3.12}$$

This suggests the interpretation of W as the design variable and (3.12) as the optimality condition corresponding to the minimization of the functional Q, subject to the equality constraints (1.3), (1.4) for the state variables (u,v,V).

If we choose the Lagrangian variables μ_i, $i = 1,2$, such that $\mu_i = 0$ only on $\partial\Omega_D \setminus \Gamma$ and $\mu_1 = \mu_2 = \eta$ on Γ for some real constant η, then we can derive a simple form of the optimality system. With this choice, the Lagrangian becomes

$$\begin{aligned}\mathcal{L}(u,v,V,W;\mu_1,\mu_2,\mu_3) = Q_\epsilon(u,v,V,W) &+ \int_\Omega\left(e^V\nabla u\cdot\nabla\mu_1 - e^{-V}\nabla v\nabla\mu_2\right)\,dx\\ &+ \int_\Omega\left(\nabla(V-V^*)\cdot\nabla\mu_3 + W\mu_3\right)\,dx - \eta\int_\Gamma J\cdot d\nu,\end{aligned} \tag{3.13}$$

and the optimality with respect to u yields

$$\left(\int_\Gamma J\cdot d\nu - I^* - \eta\right)\int_\Gamma\left(e^V\frac{\partial\hat{u}}{\partial\nu}\right)ds + \int_\Omega\left(e^V\nabla\hat{u}\cdot\nabla\mu_1\right)dx = 0. \tag{3.14}$$

With the choice $\eta = \int_\Gamma J\cdot d\nu - I^*$, this reduces to the weak form corresponding to the elliptic partial differential equation

$$\text{div}\,\left(e^V\nabla\mu_1\right) = 0 \qquad \text{in } \Omega, \tag{3.15}$$

subject to the boundary conditions

$$\mu_1 - \int_\Gamma J\cdot d\nu + I^* = 0 \qquad \text{on } \Gamma, \tag{3.16}$$

$$\mu_1 = 0 \qquad \text{on } \partial\Omega_D\setminus\Gamma, \tag{3.17}$$

$$\frac{\partial\mu_1}{\partial\nu} = 0 \qquad \text{on } \partial\Omega_N. \tag{3.18}$$

Analogous reasoning yields the equation

$$\text{div } \left(e^{-V}\nabla\mu_2\right) = 0 \qquad \text{in } \Omega, \tag{3.19}$$

subject to the same boundary conditions as for μ_1, determining the Lagrangian variable μ_2. Finally, we determine the optimality condition with respect to W, which can be rewritten as the equation

$$\epsilon\Delta W = e^{V}\nabla u\cdot\nabla\mu_1 + e^{-V}\nabla v\nabla\mu_2 \qquad \text{in } \Omega, \tag{3.20}$$

subject to homogeneous Dirichlet conditions on $\partial\Omega_D$ and homogeneous Neumann conditions on $\partial\Omega_N$.

3.3. Regularity. In the following we use the Karush–Kuhn–Tucker system derived above, which must be fulfilled by any solution of the optimal design problem, to prove additional regularity of minimizers. First, since W satisfies the Poisson equation (3.20) with right-hand side in $L^1(\Omega)$ and subject to the homogeneous boundary conditions (2.4), we may conclude that $W \in W^{1,\rho}(\Omega)$, $\rho < \frac{N}{N-1}$ (cf. [20]).

For the primal variables u and v, which satisfy the homogeneous elliptic equations (1.3) and (1.4), respectively, we can apply a standard maximum principle as in [16], which implies $u \in L^\infty(\Omega)$ and $v \in L^\infty(\Omega)$. Analogous reasoning can be applied to the dual variables μ_1 and μ_2, which solve the same elliptic equations as u and v, and whose Dirichlet boundary data are uniformly bounded too (since μ_i is piecewise constant on $\partial\Omega_D$). Hence, we may conclude that $\mu_i \in L^\infty(\Omega)$, $i = 1, 2$. As a consequence of this type of regularity, we obtain that

$$\begin{aligned}\nabla(C - C^*) = -\lambda^2\nabla W + e^{V}(u\nabla V + \nabla u) + e^{-V}(v\nabla V - \nabla v)\\ - e^{V^*}(u^*\nabla V^* + \nabla u^*) - e^{-V^*}(v^*\nabla V^* - \nabla v^*)\end{aligned} \tag{3.21}$$

is bounded in $L^\rho(\Omega)$ (with ρ as above and $\rho \le 2$ for $N = 1$), since all the gradient terms on the right-hand side are in $L^\rho(\Omega)$ and the zero-order terms are in $L^\infty(\Omega)$. Thus, we have deduced the following type of regularity for minimizers.

THEOREM 3.5. *Let $(\overline{u}, \overline{v}, \overline{V}, \overline{W}) \in \mathcal{D}_{ad}$ be a minimizer of* (3.1). *Then,*

$$\overline{u} \in L^\infty(\Omega), \qquad \overline{v} \in L^\infty(\Omega), \qquad \overline{W} \in H^1(\Omega). \tag{3.22}$$

The Lagrangian variables $\overline{\mu}_i \in H^1(\Omega)$ associated with (1.3) *and* (1.4) *satisfy*

$$\overline{\mu}_1 \in L^\infty(\Omega), \qquad \overline{\mu}_2 \in L^\infty(\Omega). \tag{3.23}$$

Moreover, if $C^* \in W^{1,\rho}(\Omega)$, then the associated doping profile $\overline{C}$ (via (1.2)*) satisfies $\overline{C} \in W^{1,\rho}(\Omega)$ for $\rho < \frac{N}{N-1}$ ($\rho \le 2$ for $N = 1$).*

We finally give an interpretation of the optimality system with respect to the local regularity of the doping profile. If the initial doping profile C^* has a discontinuity (occurring typically at a pn (positive-negative) junction), then the corresponding solution V^* is locally not C^2 across the junction, but (via standard regularity) on every open set not containing the junction. Since, for a solution of the optimality system, $V - V^*$ satisfies a Poisson equation with homogeneous boundary data and a right-hand side W that we may expect to be smooth (W solves a Poisson equation itself), the solution $V - V^*$ should have higher regularity even across the discontinuity of the doping profile. Hence, the only source of lower regularity is contained in V^*, and thus its location must be the same for $V = (V - V^*) + V^*$.

4. A fast optimization method. In the following we discuss a simple optimization method, which allows the design of semiconductor devices by solving decoupled elliptic partial differential equations only. For simplicity we consider the case of (2.5), (2.7) in the following, but an analogous approach is possible for different objective functionals, too.

We start by discussing the Lagrange–Newton iteration for the primal variables (u^k, v^k, V^k, W^k) and the dual variables (μ_1^k, μ_2^k) given by

$$\Delta V^k = \Delta V^* + W^k, \tag{4.1}$$

$$\text{div } (e^{V^{k-1}}\nabla u^k) = -\text{div } (e^{V^{k-1}}(V^k - V_{k-1})\nabla u^{k-1}), \tag{4.2}$$

$$\text{div } (e^{-V^{k-1}}\nabla v^k) = \text{div } (e^{-V^{k-1}}(V^k - V_{k-1})\nabla u^{k-1}), \tag{4.3}$$

$$\text{div } (e^{V^{k-1}}\nabla \mu_1^k) = -\text{div } (e^{V^{k-1}}(V^k - V_{k-1})\nabla \mu_1^{k-1}), \tag{4.4}$$

$$\text{div } (e^{-V^{k-1}}\nabla \mu_2^k) = \text{div } (e^{-V^{k-1}}(V^k - V_{k-1})\nabla \mu_2^{k-1}), \tag{4.5}$$

$$\begin{aligned} -\epsilon W^k = &-e^{V^{k-1}}\left((V^k\nabla u^{k-1} + \nabla u^k)\cdot\nabla\mu_1^{k-1} + \nabla u^{k-1}\cdot\nabla\mu_1^k\right) \\ &+ e^{-V^{k-1}}\left((-V^k\nabla v^{k-1} + \nabla v^k)\cdot\nabla\mu_2^{k-1} + \nabla v^{k-1}\cdot\nabla\mu_2^k\right), \end{aligned} \tag{4.6}$$

subject to the boundary conditions (1.5)–(1.11). As for the solution of the drift-diffusion system, the full Lagrange–Newton method yields a sequence of systems of partial differential equations, which is in general convection-dominated due to the strong influence of first-order terms. As for the drift-diffusion system, we may expect the numerical solution of this system to be a difficult task, particularly for large applied voltages. Moreover, the advantage of our optimization approach, namely, the partial decoupling into scalar elliptic partial differential equations, is lost by using this Newton-type approach.

Therefore it seems favorable to use a different iterative method for the solution of the optimality system. Using a lower triangular approximation of the optimality system, we first solve (2.2) with given W for the potential V, and subsequently the continuity equations (1.3), (1.4) with given potential V for u and v. With given potential and given u and v, we solve the adjoint equations (3.15), (3.19) to obtain the Lagrangian variables μ_1 and μ_2. Finally, we can perform a gradient step with respect to the design variable W using the optimality equation (3.20). Due to the simple structure of this equation, it seems reasonable to discretize the Laplace term in an implicit way and thus to solve

$$-\epsilon\Delta W + \tau W = \tau W^* - e^V\nabla u\cdot\nabla\mu_1 + e^{-V}\nabla v\nabla\mu_2 \tag{4.7}$$

for an appropriately chosen damping parameter τ, where W^* is the old value of W. Putting this all together, we can write this iteration as

$$\Delta V^k = \Delta V^* + W^{k-1}, \tag{4.8}$$

$$\text{div } (e^{V^k}\nabla u^k) = 0, \tag{4.9}$$

$$\text{div } (e^{-V^k}\nabla v^k) = 0, \tag{4.10}$$

$$\text{div } (e^{V^k}\nabla \mu_1^k) = 0, \tag{4.11}$$

$$\text{div } (e^{-V^k}\nabla \mu_2^k) = 0, \tag{4.12}$$

$$-\epsilon W^k + \tau W^k = \tau W^{k-1} - e^{V^k}\nabla u^k\cdot\nabla\mu_1^k + e^{-V^k}\nabla v^k\nabla\mu_2^k, \tag{4.13}$$

subject to the above boundary conditions. The only coupling in the boundary conditions occurs in the condition $\mu_i = \int_\Gamma J^k\cdot d\nu$, but there also we can use the previously

computed values for u^k and v^k to obtain J^k. The corresponding value of the doping profile can be computed independently by

$$C^k - C^* = -\lambda^2 W^k + n^k - n^* - p^k + p^*, \tag{4.14}$$

where $n^k = e^{V^k} u^k$ and $p^k = e^{-V^k} v^k$.

5. Numerical examples. In the following we report numerical results for three different examples, two bipolar diodes and a unipolar MESFET device. For all the numerical experiments we use the physical parameters for silicon as given in Table 5.1, with a standard forward-bias scaling of the variables (cf. [16]). Moreover, we use the mobilities $\mu_n = \mu_p = \mu_0$. The objective functional is given by (2.5), (2.7), with the aim of increasing the current flow over a contact. This aim can in general be achieved for ϵ sufficiently small, but for some cases only with a doping profile deviating far from the original, so that one may rather use the result for a larger value of ϵ. All the numerical examples have been implemented within the software system MATLAB.

TABLE 5.1
Physical parameters for silicon.

Parameter	Physical meaning	Numerical value
q	elementary charge	$1.6 \cdot 10^{-19}$ As
n_i	intrinsic density	10^{10} cm^{-3}
ϵ_S	permittivity constant	10^{-12} As V^{-1}s^{-1}
μ_0	low field mobility	$1.5 \cdot 10^3$ cm^{2}V^{-1}s^{-1}
U_T	thermal voltage at $T = 300$K	0.0259 V

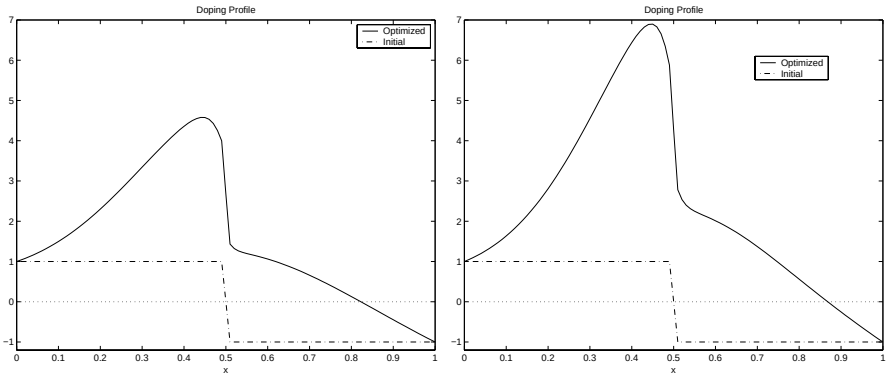


FIG. 5.1. *Initial (dash-dotted) and optimized (solid) doping profile in the example of section* 5.1, *for* $\epsilon = 10^{-2}$ *(left) and* $\epsilon = 10^{-3}$ *(right), with dimensionless units in the* x*- and* y*-axes according to the forward-bias scaling.*

5.1. A pn-diode. Our first example is a pn-diode, with the domain Ω scaled to the unit interval $(0, 1)$. The pn-diode is characterized by a doping profile that has exactly one positive and one negative region; we choose one of the simplest possibilities for the (scaled) initial doping profile, namely, a function jumping almost abruptly from the value 1 to -1 at the junction. This initial profile is shown as the dash-dotted function in Figure 5.1. The Debye length in this experiment is given by $\lambda^2 = 10^{-3}$, and the value of the applied voltage is $U = 10U_T = 0.259$. The optimization objective

is to increase the current flow (i.e., the current flow density, which is constant in the domain since $J_x = 0$) by 50%, and consequently we chose

$$I^* = 1.5 \cdot \int_\Gamma J_0 \cdot d\nu, \tag{5.1}$$

where $\Gamma = \{0\}$ and J_0 is the current flow density obtained with the initial doping.

For the numerical solution of the drift-diffusion system and the linear elliptic equations arising during the iterative solution of the optimization problem in this example (as well as the following one) we use an exponentially fitted scheme of Scharfetter–Gummel type (cf. [1, 4]); the fineness of the uniform spatial grid is given by $h = 10^{-2}$. For the discretization of all variables involved $(C, V, W, n, p, \mu_1, \mu_2)$ we use piecewise linear finite elements. The drift-diffusion system with given initial doping profile is solved using Newton's method and voltage continuation to obtain the initial value of the potential (cf. [17]).

The numerical experiments were performed for several values of ϵ, with the result that most changes appeared for ϵ between 10^{-3} and 10^{-2}, and thus we plot the results for these two values in Figure 5.1. (Note that for large values of ϵ the penalty does not allow enough change to the initial configuration, while for small values the observation tends to be almost zero in our case, so that further change in the solution is negligible.) Figure 5.1 shows the optimized doping profiles (solid) in both cases compared to the initial one (dash-dotted). The dashed line is the coordinate axis for x; its cut with the doping profile marks the pn-junction. One observes that the optimized doping profiles have similar shapes for both values of ϵ, but the magnitude of the doping in the n-region grows with decreasing ϵ. In both cases the n-region grows at the expense of the p-region; i.e., the pn-junction moves right, and the value of C is larger than the initial one in the whole domain. Moreover, the doping profile remains steep around the center point $x = 0.5$, which numerically confirms the result obtained from the analysis of the optimality system.

An analogous effect happens with the potential V, which is the actual design variable in our approach, and with the electron and hole densities, i.e., the shape changes strongly compared to the initial one for $\epsilon = 10^{-2}$, and if we decrease the value of ϵ, the resulting current flow can be forced to be closer to the desired flow only by a change in magnitude. The resulting potentials for both values compared to the initial ones are shown in Figure 5.2. The electron and hole densities are shown in Figure 5.3 for the value of $\epsilon = 10^{-3}$.

Finally, we illustrate the behavior of the objective functional, observation, and penalizing energy term in the left-hand plot of Figure 5.4, and the change in the current-voltage characteristic in the right-hand plot (both for $\epsilon = 10^{-3}$). Not surprisingly, the objective functional and observation are reduced in a few iterations, while the energy initially increases since the doping is pulled away from the initial one. In the later stage of the iteration, the observation part remains almost constant, and a decrease in the objective functional is obtained only due to a (slow) decrease in the energy. An inspection of the current-voltage plot shows that the optimized doping profile also yields a characteristic whose absolute value increases exponentially for positive applied voltages, but which lies above the initial one for all applied voltages.

5.2. An npn-diode. Our second numerical example is an npn-diode, with the same parameter settings as in the example of section 5.1, and with the same choice of objective. The initial doping profile is a piecewise constant function taking the values

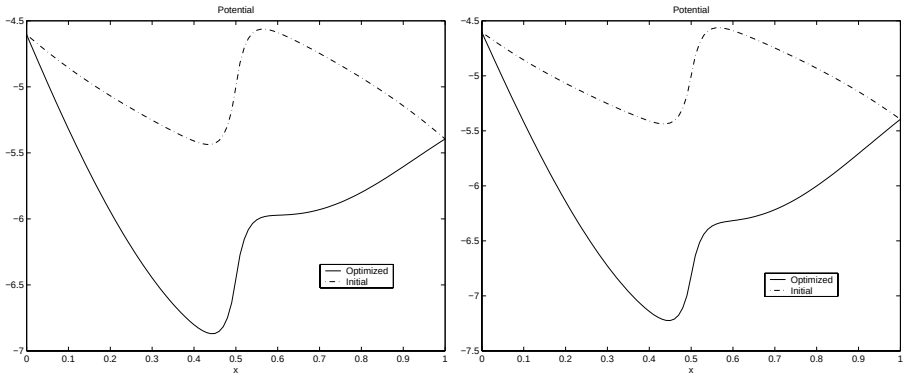


FIG. 5.2. *Initial (dash-dotted) and optimized (solid) potential in the example of section* 5.1*, for* $\epsilon = 10^{-2}$ *(left) and* $\epsilon = 10^{-3}$ *(right).*

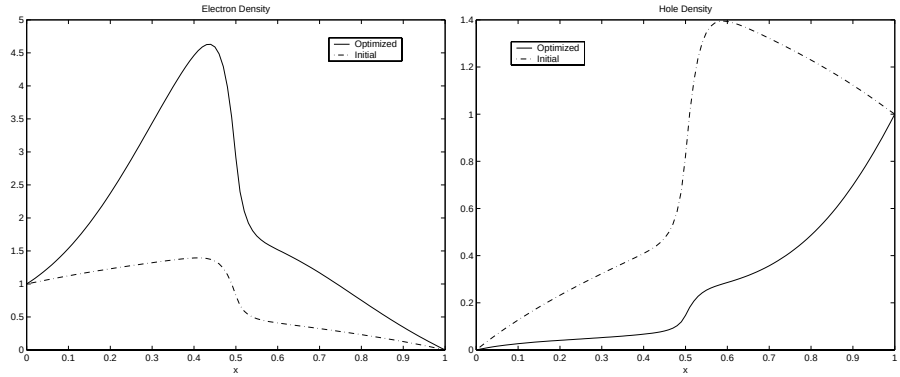


FIG. 5.3. *Initial (dash-dotted) and optimized (solid) electron (left) and hole density (right) in the example of section* 5.1*, for* $\epsilon = 10^{-3}$*.*

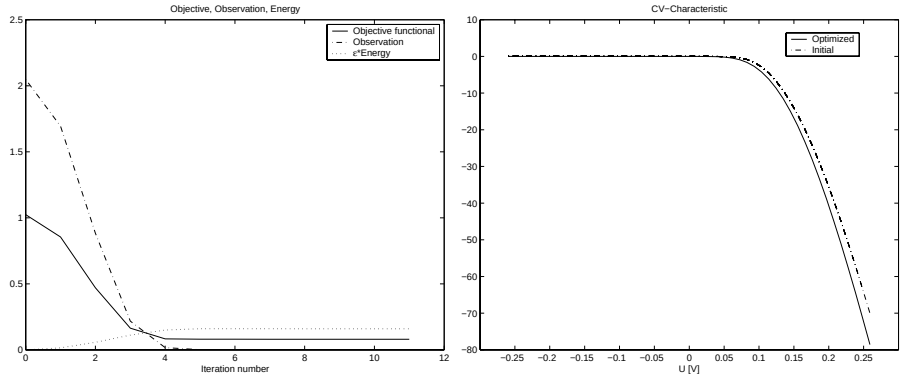


FIG. 5.4. *Evolution of the objective functional (solid), observation (dash-dotted), and energy (dotted) in the example of section* 5.1 *(left), and the current-voltage characteristic for* $\epsilon = 10^{-3}$ *(right, scaled current plotted vs. U[V]).*

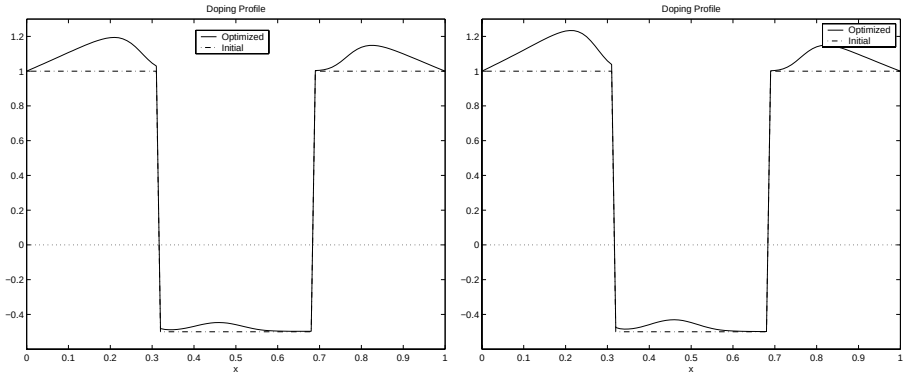


FIG. 5.5. *Initial (dash-dotted) and optimized (solid) doping profile in the example of section* 5.2, *for* $\epsilon = 10^{-6}$ *(left) and* $\epsilon = 10^{-8}$ *(right).*

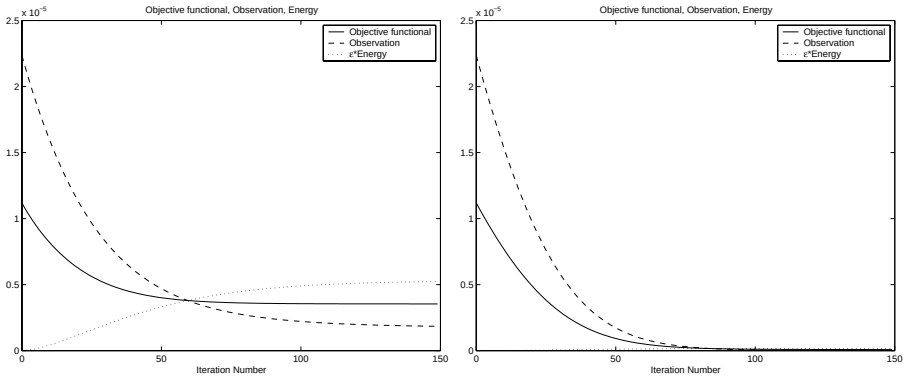


FIG. 5.6. *Evolution of the objective functional (solid), observation (dashed), and energy (dotted) in the example of section* 5.2, *for* $\epsilon = 10^{-6}$ *(left) and* $\epsilon = 10^{-8}$ *(right).*

one and $-\frac{1}{2}$; it is shown as the dash-dotted function in Figure 5.5. The values of the parameter ϵ that lead to useful results are now between 10^{-6} and 10^{-8}, which is due to the lower absolute values of the current obtained in this example. The objective value obtained in the first case is around 0.4, while the objective is reduced almost to zero for the the second. The evolution of the objective functional, the observation, and the energy term is plotted in Figure 5.6, showing a similar behavior as in the example of section 5.1. For $\epsilon = 10^{-8}$, the energy term is already negligible, and the iteration is driven by the reduction of the observation error.

The optimized and initial values for the doping profile are plotted in Figure 5.5 and for the potential in Figure 5.7. In this case, the change in the potential is quite small, and the doping profile is increased slightly both in the n- and p-regions. A more significant change happens in the electron density, shown for the case of $\epsilon = 10^{-8}$ in Figure 5.8. As one might expect, the electron density is changed mainly in the n-regions, while the hole density is changed in the p-region.

Finally, we plot the negative CV-characteristics (i.e., the map $U \mapsto J$) of the initial and optimized devices for both values of ϵ in Figure 5.9. The shape of these current-voltage curves remains similar for all values of ϵ, but obviously changes in magnitude as ϵ tends to zero.

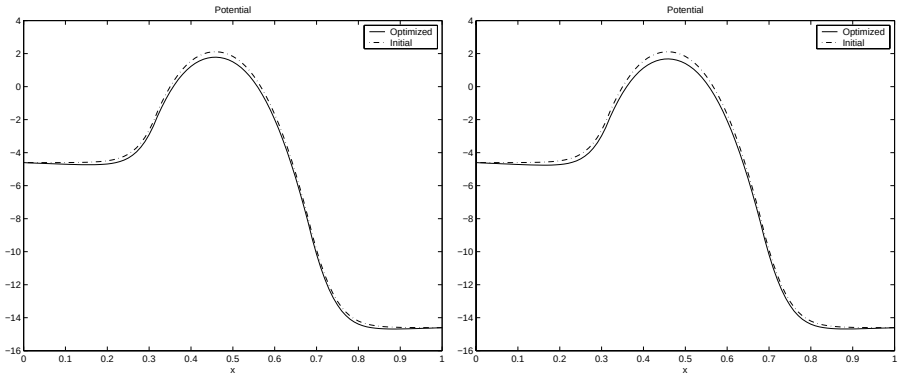


FIG. 5.7. *Initial (dash-dotted) and optimized (solid) potential in the example of section* 5.2, *for* $\epsilon = 10^{-6}$ *(left) and* $\epsilon = 10^{-8}$ *(right).*

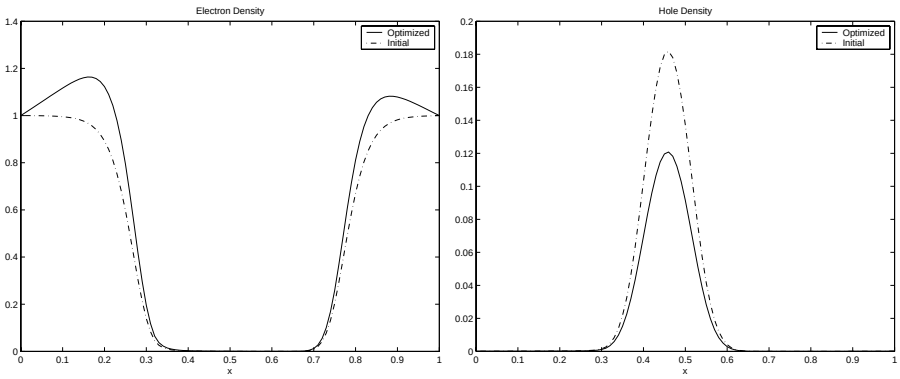


FIG. 5.8. *Initial (dash-dotted) and optimized (solid) electron density (left) and hole density (right) in the example of section*5.2, *for* $\epsilon = 10^{-8}$.

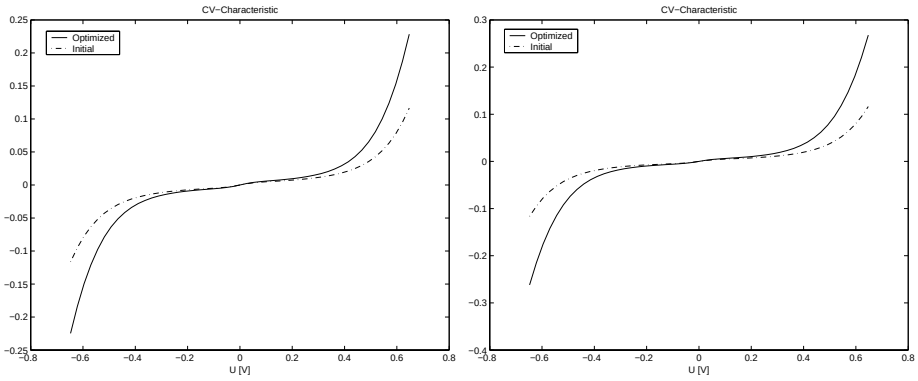


FIG. 5.9. *Negative current-voltage characteristic obtained with the initial (dash-dotted) and the optimized (solid) doping profiles in the example of section* 5.2, *for* $\epsilon = 10^{-6}$ *(left) and* $\epsilon = 10^{-8}$ *(right).*

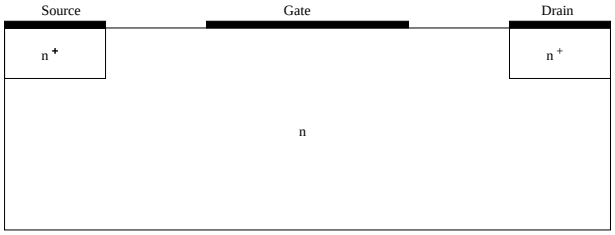


FIG. 5.10. *Device geometry in the example of section* 5.3.

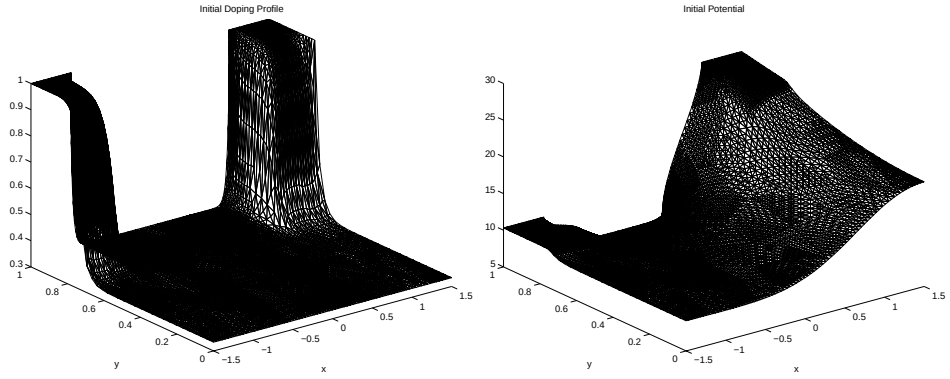


FIG. 5.11. *Initial value of the doping profile and the potential in the example of section* 5.3.

5.3. A MESFET device. As our final example, we consider the optimal design of a MESFET in two spatial dimensions. We use a device geometry and initial doping profile as in an example considered in [12], with a length of 6μm and a width of 2μm. The geometry and the position of the contacts is shown in Figure 5.10. The scaled initial doping profile (by the value $C_s = 10^{14}\text{cm}^{-3}$) is shown in Figure 5.11; in order to improve the visibility, we plot the initial values and subsequently the results with different scaling of the x- and the y-axes. A MESFET can be modeled as a unipolar device, which is also reflected by the positivity of the doping profile in the whole device region. Thus, we have $p = v = 0$ in Ω, and the equation for v as well as the adjoint equation determining the Lagrangian variable $\mu_2 \equiv 0$ can be eliminated, which reduces the computational effort. The boundary data are specified by $n = 0.5(C + \sqrt{C^2 + 4\delta^2})$, a temperature of $T = 300^\circ$ on each contact, and

- at the source, $V = V_{bi} - 0.1[V] = 0.1670[V]$;
- at the drain, $V = V_{bi} + 0.4[V] = 0.6670[V]$;
- at the gate, $V = V_{bi} = 0.2385[V]$.

Our objective is to increase the current flow over the drain by 50%, and consequently we choose

$$I^* = 1.5 \cdot \int_\Gamma J_0 \cdot d\nu, \tag{5.2}$$

where Γ is the drain contact and J_0 is the current flow density obtained with the initial doping.

For the finite element discretization of the original problem we used an adaptive solver (with a standard error estimate for the Poisson equation), with a resulting mesh

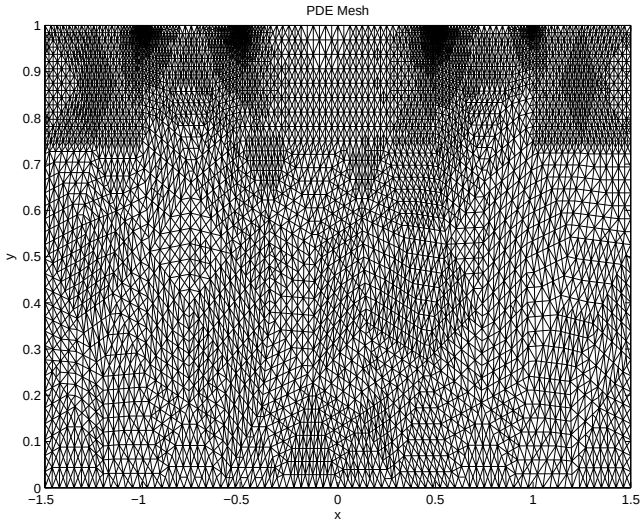


FIG. 5.12. *Mesh used for the optimization in the example of section* 5.3.

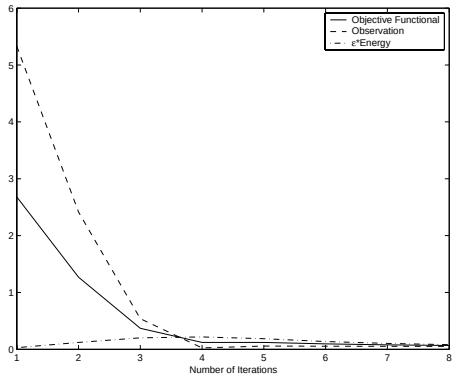


FIG. 5.13. *Evolution of the objective functional (solid), observation error (dashed), and energy term (dash-dotted) in the example of section* 5.3.

(shown in Figure 5.12) consisting of $n_t = 15434$ triangular elements. This mesh is used subsequently also for the optimization algorithm. The refinement was started from a relatively fine mesh in the bulk. Probably one could use a coarser mesh for all computations, but due to the efficiency of our optimization algorithm we are able to solve the optimal design problem with reasonable effort even for this fine triangulation. This strategy of using grid adaption only for the initial problem is motivated by the above observation that steep junctions will remain at the same location during the optimization process.

In this case, it turns out that a suitable choice of ϵ is 10^{-3}, and here we show the results for this value. Changes of ϵ lead to a behavior similar to that in the previous examples. The resulting increase in the current flow over the drain is around 45% for this parameter value. The evolution of the objective function, the observer error, and the energy term are shown in Figure 5.13. The optimized doping profile and the optimized potential are shown in Figure 5.14. A comparison with the initial value

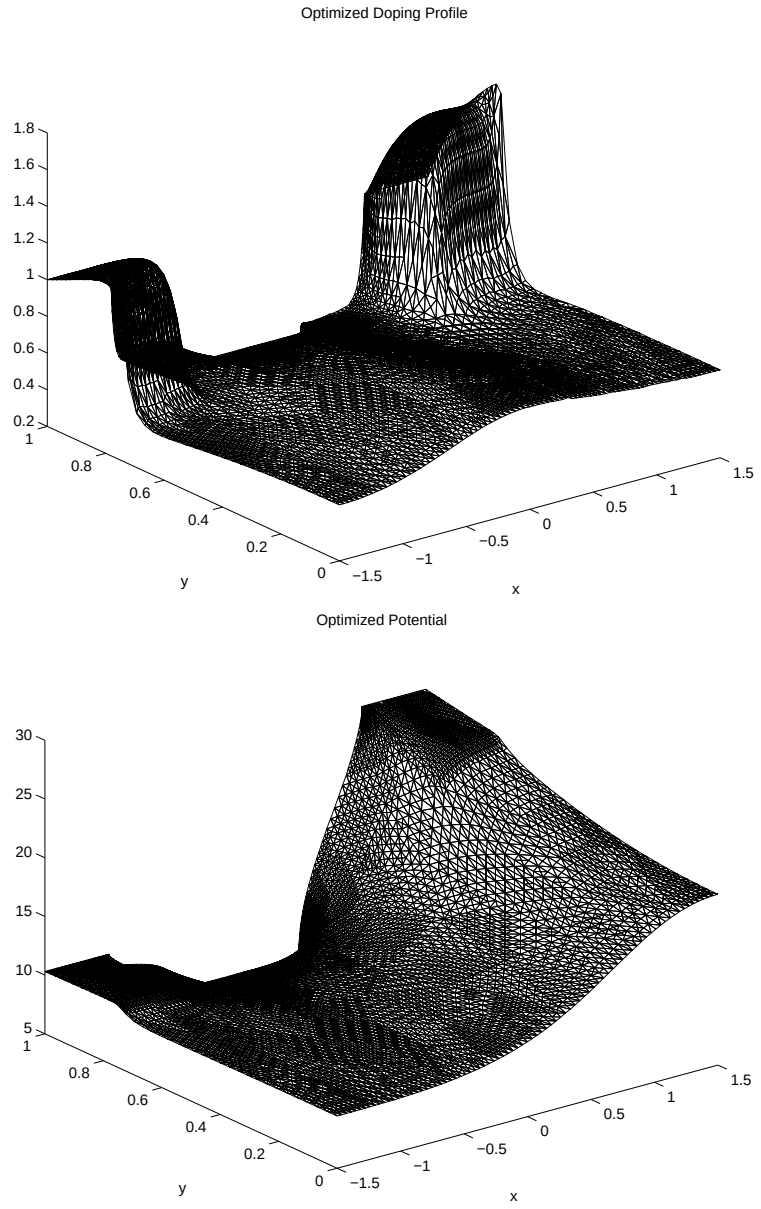


FIG. 5.14. *Optimized doping profile and potential in the example of section* 5.3, $\epsilon = 10^{-3}$.

shows that the doping profile is increased mainly close to the drain, while it is slightly decreased close to the device corner opposite to the drain. The change in the potential is less significant, due to the fact that the main shape of the potential is determined by the rather high difference in the boundary values. Changes in the doping profile (and in the electron density $n = e^V u$) are mainly caused by the Laplacian of $V - V^*$, which is still of considerable magnitude.

An inspection of the evolution of the objective demonstrates again the efficiency of our approach, since a minimum is obtained with only a few iterations. Since, in

each iteration, we have to solve only four scalar elliptic partial differential equations, the numerical effort per iteration is similar to two Gummel-type iteration steps for the (unipolar) drift-diffusion system. Since in general the number of iterations in a Gummel-type method for the drift-diffusion system is of similar size to the number of iterations needed for the optimal design problem, the overall numerical effort for optimization is around the effort for two forward solves of the nonlinear drift-diffusion system, which is a surprising result.

6. Conclusions. We have presented a new, fast approach to the optimal design of semiconductor devices, which can be applied if a performance optimization of the device at a single fixed applied voltage is desired. The numerical experiments illustrate reasonable convergence properties of the simple algorithm we have proposed to solve the optimal design problem, and clearly demonstrate its efficiency. In particular, we have obtained an optimization procedure with a numerical effort of similar magnitude to one with few forward solves.

We finally would also like to mention the natural limitations or possible generalizations of the approach presented in this paper. These limitations arise if the optimization goal involves current flows for several applied voltages and consequently several different potentials, since we can interpret only one of the potentials as the design variable in this case. This statement applies, in particular, in the context of identifying unknown doping profiles, which is usually done by minimizing a least-squares functional involving a large number of different voltages (cf. [2]). In many typical optimal design situations, however, the aim is to control at most the currents for two different voltages, namely, for an *on-state voltage* and an *off-state voltage* (close to equilibrium). The usual aim in such a situation is to maximize the on-state current flow on a contact by keeping the off-state current flow below some threshold value (cf. [21, 23]). In optimal design problems of this type it seems natural to eliminate the Poisson equation for the on-state potential and keep that for the off-state potential, since solutions of the drift-diffusion system close to equilibrium are rather cheap. A numerical investigation of such optimal design situations shall be left to future research.

Acknowledgments. The authors thank Prof. Peter Markowich (University of Vienna) for interesting discussions and, in particular, for stimulating this collaboration.

REFERENCES

[1] F. Brezzi, L. D. Marini, and P. Pietra, *Two-dimensional exponential fitting and applications to drift-diffusion models*, SIAM J. Numer. Anal., 26 (1989), pp. 1342–1355.

[2] M. Burger, H. W. Engl, P. Markowich, and P. Pietra, *Identification of doping profiles in semiconductor devices*, Inverse Problems, 17 (2001), pp. 1765–1795.

[3] M. Burger, H. W. Engl, and P. Markowich, *Inverse doping problems for semiconductor devices*, in Recent Progress in Computational and Applied PDEs, T. F. Chan, Y. Huang, T. Tang, J. A. Xu, and L. A. Ying, eds., Kluwer Academic Publishers, Boston, Dordrecht, London, 2002, pp. 39–54.

[4] M. Burger and R. Pinnau, *Exponential Fitting for Adjoint Equations in Semiconductor Design*, in preparation, 2002.

[5] L. Ciampolini, *Scanning Capacitance Microscope Imaging and Modelling*, Ph.D. thesis, ETH Zürich, Hartung-Gorre Verlag, Konstanz, 2001.

[6] A. C. Diebold, M. R. Kump, J. J. Kopanski, and D. G. Seiler, *Characterization of two-dimensional dopant profiles: Status and review*, J. Vac. Sci. Technol. B, 14 (1996), pp. 196–201.

[7] W. Fang and E. Cumberbatch, *Inverse problems for metal oxide semiconductor field-effect transistor contact resistivity*, SIAM J. Appl. Math., 52 (1992), pp. 699–709.
[8] W. Fang and K. Ito, *Identifiability of semiconductor defects from LBIC images*, SIAM J. Appl. Math., 52 (1992), pp. 1611–1626.
[9] W. Fang and K. Ito, *Reconstruction of semiconductor doping profile from laser-beam-induced current image*, SIAM J. Appl. Math., 54 (1994), pp. 1067–1082.
[10] M. Hinze and R. Pinnau, *Optimal control of the drift-diffusion model for semiconductor devices*, in Optimal Control of Complex Structures, Internat. Ser. Numer. Math. 139, K.-H. Hoffmann, I. Lasiecka, G. Leugering, and J. Sprekels, eds., Birkhäuser Boston, Cambridge, MA, 2001, pp. 95–106.
[11] M. Hinze and R. Pinnau, *An optimal control approach to semiconductor design*, Math. Models Methods Appl. Sci., 12 (2002), pp. 89–107.
[12] S. Holst, A. Jüngel, and P. Pietra, *A mixed finite-element discretization of the energy-transport model for semiconductors*, SIAM J. Sci. Comput., 24 (2003), pp. 2058–2075.
[13] N. Khalil, *ULSI Characterization with Technology Computer-Aided Design*, Ph.D. thesis, Technical University Vienna, Vienna, Austria, 1995.
[14] N. Khalil, J. Faricelli, D. Bell, and S. Selberherr, *The extraction of two-dimensional MOS transistor doping via inverse modeling*, IEEE Electron Device Lett., 16 (1995), pp. 17–19.
[15] P. A. Markowich, *The Stationary Semiconductor Device Equations*, Springer, Vienna, New York, 1986.
[16] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer, Vienna, New York, 1990.
[17] M. S. Mock, *Analysis of Mathematical Models of Semiconductor Devices*, 1st ed., Boole Press, Dublin, 1983.
[18] R. Plasun, M. Stockinger, R. Strasser, S. Selberherr, *Simulation based optimization environment and its application to semiconductor devices*, in Proceedings of the IASTED International Conference on Applied Modelling and Simulation, Honolulu, HI, Acta Press, Calgary, AB, 1998, pp. 313–316.
[19] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer, Vienna, New York, 1984.
[20] R. E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, AMS, Providence, RI, 1997.
[21] M. Stockinger, *Optimization of Ultra-Low-Power CMOS Transistors*, Ph.D. thesis, Technical University Vienna, Vienna, Austria, 2000.
[22] M. Stockinger, R. Strasser, R. Plasun, A. Wild, and S. Selberherr, *A qualitative study on optimized MOSFET doping profiles*, in Proceedings of the Simulation of Semiconductor Processes and Devices (SISPAD 98) Conference, Leuven, Belgium, 1998, Springer-Verlag, New York, 1998, pp. 77–80.
[23] M. Stockinger, R. Strasser, R. Plasun, A. Wild, and S. Selberherr, *Closed-loop MOSFET doping profile optimization for portable systems*, in Proceedings of the International Conference on Modeling and Simulation of Microsystems, Semiconductors, Sensors, and Actuators, San Juan, 1999, Applied Computational Research Society, Cambridge, MA, 1999, pp. 411–414.
[24] W. R. Van Roosbroeck, *Theory of flow of electrons and holes in germanium and other semiconductors*, Bell Syst. Tech. J., 29 (1950), pp. 560–607.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 127–151

ASYMPTOTIC SHAPE OF THE ERLANG CAPACITY REGION OF A MULTISERVICE SHARED RESOURCE*

JOHN A. MORRISON† AND DEBASIS MITRA†

Abstract. We consider a loss model of an unbuffered resource having C channels, which are shared by several different types of service connections. Connections of each type arrive in a Poisson stream and request a number of channels, which depends on the type. An arriving connection is blocked and lost if there are not enough free channels. Otherwise, the channels are held for the duration of the connection, and the holding period is generally distributed. It is assumed that C and the traffic intensities are proportionately large. The admission control problem is considered for specified upper bounds on the blocking probabilities, and the boundary of the admissible set is investigated asymptotically. The results are derived by investigating the local behavior with respect to the tangent hyperplane at a point on the boundary of the admissible set. The lowest order results that hold in the asymptotic limit $C \to \infty$ are given first. Importantly, the boundary is linear for the critically loaded and overloaded regimes and weakly convex for the underloaded regime. Next, refined results that hold for $C \gg 1$ are given, which indicate that the boundary is nonconvex, although only slightly so, for the overloaded and underloaded regimes. The critically loaded regime requires further investigation, which is carried out in [J. A. Morrison, *SIAM J. Appl. Math.*, 64 (2003), pp. 1–17].

Key words. admissible set, asymptotics, Erlang capacity, network design, network economics

AMS subject classifications. 60K30, 90B12

DOI. 10.1137/S0036139901399417

1. Introduction.

We consider an unbuffered resource having C channels, which are shared by J different types of connections. Connections of type j arrive in a Poisson stream with mean rate λ_j, and they require d_j channels. An arriving connection is blocked and lost if there are fewer than d_j free channels. Otherwise, d_j channels are held for the duration of the connection, and the holding period is generally distributed with mean $1/\mu_j$ and is independent of earlier arrival and holding times. The traffic intensity of type j connections is $\rho_j = \lambda_j/\mu_j$, and the product form and the insensitivity property hold [4], [5], [7]; i.e., the joint stationary distribution of the number of active connections of each type depends on the distributions only through ρ_i, $i = 1, \ldots, J$. The blocking probabilities L_j for type j connections satisfy $L_j > 0$ for $\rho_j > 0$ and, assuming that $C \geq \max_i d_i$, $L_j \to 0+$ only if $\rho_i \to 0+$, $i = 1, \ldots, J$. The admissible set in $\mathbb{R}^J$ contains all combinations of ρ_j, $j = 1, \ldots, J$, such that the blocking probability for each connection type satisfies specified bounds, i.e., $L_j \leq \ell_j$, $j = 1, \ldots, J$, where ℓ_j is a prescribed function of C.

Characterization of the admissible set is extremely useful, not only for connection-level admission control, which is the context in which this topic has typically been considered in the past, but also for higher level objectives, such as network economics and network design and operations. The asymptotic view of the admissible set is particularly appropriate for the latter, where the fine details are not as important as the qualitative properties of the shape of the set and tractability of the numerical calculations for large systems.

*Received by the editors December 10, 2001; accepted for publication (in revised form) March 11, 2003; published electronically November 19, 2003. The main results of this paper were stated without proof in J. A. Morrison and D. Mitra, *Asymptotic shape of the Erlang capacity region of a multi-service shared resource*, Perform. Eval., 49 (2002), pp. 273–281.

http://www.siam.org/journals/siap/64-1/39941.html

†Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974 (johnmorrison@lucent.com, mitra@lucent.com).

Special importance is attached to admissible regions with linear boundaries; the solution space is determined by its vertices, which are relatively easy to compute. Optimizations within such spaces are also much easier computationally. Network economics applications are given in [6]. In one such example, the objective function is the profit of a service provider giving several quality of service (QoS) levels at prices that are the solution to the corresponding optimization problem. The admissible region of solutions is defined by a collection of inequalities imposed by available capacity, one for each QoS level. For further details see [6], and for other such applications see [2], [8], [19] and references therein. See [1], [11] for applications to routing and control. For recent work on loss models of optical networking see [17] and [18].

Convexity is also important, since the tangent hyperplane at points on the boundary of the admissible set intersects the positive axes. Hence, if the boundary of the admissible set is convex, then the region in the positive orthant bounded by the tangent hyperplane at a point on the boundary is admissible and may be used in an approximate optimization.

Mitra and Morrison [10] considered our model here (as well as the finite-sources version), and they investigated the case when C and the traffic intensities $\rho_j = \alpha_j C$, $j = 1, \ldots, J$, are proportionately large, so that $\alpha_j = O(1)$ is bounded away from zero. They derived uniform asymptotic approximations to the blocking probabilities L_j, $j = 1, \ldots, J$, for type j connections. The results for Poisson arrivals are obtained from the finite-sources version as a limiting case. They presented numerical results for $J = 2$ and $J = 3$ types for the finite-sources model. These results constitute a numerical procedure but do not provide a characterization of the admissible set, nor do they resolve specific questions on the linearity and convexity of the boundary.

In this paper we investigate the boundary of the admissible set in the case of Poisson arrivals. The admissible set $\mathcal{A}$ is given by

$$\mathcal{A} = \{\alpha_1, \ldots, \alpha_J | L_j(\alpha_1, \ldots, \alpha_J; C) \leq \ell_j(C),\ j = 1, \ldots, J\}. \tag{1.1}$$

The uniform asymptotic approximations to the blocking probabilities are specialized [10] to three regimes in which their behavior is markedly different, namely, the overloaded, the critically loaded, and the underloaded regimes, corresponding to $\sum_{j=1}^{J} d_j \alpha_j > 1$, $\sum_{j=1}^{J} d_j \alpha_j - 1 = O(1/\sqrt{C})$, and $\sum_{j=1}^{J} d_j \alpha_j < 1$, respectively. The corresponding blocking probabilities L_j are $O(1)$, $O(1/\sqrt{C})$, and exponentially small in C, respectively. The shape of the admissible set is investigated separately for each of the three regimes, and it is assumed that $\min_j \ell_j$ is $O(1)$, $O(1/\sqrt{C})$, and exponentially small in C, respectively.

The lowest order results are stated in section 2. In the asymptotic limit $C \to \infty$, with $\rho_j = \alpha_j C$, $j = 1, \ldots, J$, the boundary of the admissible set lies in a hyperplane if the resource is critically loaded, which is the regime of greatest interest, or if it is overloaded. If the resource is underloaded, the boundary of the admissible set, in the limit $C \to \infty$, is convex, but not strictly so, except when $J = 2$.

The refined results, which pertain to $C \gg 1$, are stated in section 3. In general, the boundary of the admissible set is not convex. If $J = 2$, then the boundary is slightly convex if the resource is critically loaded, but slightly concave if the resource is overloaded. For $J = 2$, the convexity is maintained for an underloaded resource. Unfortunately, for $J \geq 3$, the boundary of the admissible set is *not* convex, although only slightly so, whether the resource is overloaded or underloaded. The case of a critically loaded resource requires further investigation, which is to appear in a forthcoming paper [12], and shows that the boundary of the admissible set is *not* convex, although only very slightly so.

The above results follow from consideration of the local behavior with respect to the tangent hyperplane at a point on the boundary of the admissible set. For instance, in the underloaded regime, the tangent hyperplane is given by $\xi_J = 0$. Locally, for $J \geq 3$, the boundary has the asymptotic form

$$\xi_J \sim A\xi_{J-1}^2 - \frac{B}{C}\xi_{J-2}^2 \,, \tag{1.2}$$

where ξ_{J-1} and ξ_{J-2} are linear in α_i, $i = 1, \ldots, J-1$, and A and B are $O(1)$ and positive. To this order, for $J > 3$, quadratic terms involving ξ_i, $i = 1, \ldots J-3$, are absent. Since A and B are positive, the boundary is nonconvex, but only slightly so because of the $O(1/C)$ negative coefficient of ξ_{J-2}^2. In the limit $C \to \infty$ the boundary is not strictly convex, because of the degeneracy in the quadratic form.

Although our results are negative, in that they show that the boundary of the admissible set is not convex in general, it is important that the practitioner be aware of this. Moreover, since the boundary is only slightly nonconvex, a slightly smaller admissible region with a convex boundary may be used by the practitioner.

The lowest order analysis of the boundary of the admissible set for an underloaded resource is given in section 4. Both the lowest order and refined approximations for an overloaded resource are derived in section 5. A critically loaded resource is analyzed in section 6, and a refined approximation to the boundary of the admissible set is derived in section 7. A refined approximation for an underloaded resource is derived in section 8, and a check on a negative coefficient, which gives rise to the nonconvexity of the boundary of the admissible set, is carried out in section 9.

2. Lowest order results. Throughout the paper, we assume that

$$C \gg 1; \ \rho_j = \alpha_j C, \quad j = 1, \ldots, J, \tag{2.1}$$

where C is an integer and $\alpha_j > 0$ is $O(1)$ and bounded away from zero. We also assume that d_j, $j = 1, \ldots, J$, are distinct positive integers, not large relative to C, and that the greatest common divisor of $d_1, \ldots, d_J$ is 1. The admissible set corresponds to

$$L_j\left(\alpha_1, \ldots, \alpha_J; \ C\right) \leq \ell_j, \quad j = 1, \ldots, J, \tag{2.2}$$

where the function L_j gives the blocking probability for type j connections. It is shown, in all three regimes, that asymptotically

$$\frac{\partial L_j}{\partial \alpha_k} > 0, \quad j, k = 1, \ldots, J, \tag{2.3}$$

and the boundary of the admissible set is expressed in the form

$$\alpha_J = \alpha_J(\boldsymbol{\alpha}; C), \quad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{J-1}) \,. \tag{2.4}$$

We first give the lowest order asymptotic results, which are obtained from the finite-sources version [10] as a limiting case. We have the following proposition.

PROPOSITION 2.1. (A) *If ℓ_j, $j = 1, \ldots, J$, is asymptotically bounded away from zero, then on the boundary of the admissible set the resource is overloaded. Let*

$$a_j = (1 - \ell_j)^{1/d_j} \,, \quad j = 1, \ldots, J, \tag{2.5}$$

and, by choice of notation,

$$a_J = \max_j a_j. \tag{2.6}$$

Then the boundary of the admissible set satisfies

$$\sum_{j=1}^{J} d_j a_J^{d_j} \alpha_j = 1 + O\left(\frac{1}{C}\right), \tag{2.7}$$

which in the asymptotic limit $C \to \infty$ lies in a hyperplane.

(B) *If*

$$\sqrt{C} \min_j \ell_j = O(1) \tag{2.8}$$

is bounded below by a positive constant, then on the boundary of the admissible set the resource is critically loaded. The boundary satisfies

$$\sum_{j=1}^{J} d_j \alpha_j = 1 + O\left(1/\sqrt{C}\right), \tag{2.9}$$

which in the asymptotic limit $C \to \infty$ lies in a hyperplane. Note, by comparison with (2.7)*, that for a common $C \gg 1$ the admissible set here is smaller and the order of magnitude of the error is larger. An asymptotic approximation to the error term in* (2.9) *is derived in section 6.*

(C) *If at least one ℓ_j is exponentially small, the resource is underloaded. Let*

$$\ell_j = \frac{e^{-C\omega}}{\sqrt{2\pi C}} \beta_j, \quad j = 1, \ldots, J; \quad \min_j \beta_j = 1, \quad \omega > 0. \tag{2.10}$$

Note that some of the ℓ_j may not be exponentially small. Also, let $\tau > 1$ be the unique solution of

$$\sum_{j=1}^{J} \alpha_j \left(d_j \tau^{d_j} \log \tau + 1 - \tau^{d_j}\right) = \omega. \tag{2.11}$$

Then the boundary of the admissible set satisfies

$$\sum_{j=1}^{J} d_j \alpha_j \tau^{d_j} = 1 + O\left(\frac{1}{C}\right), \tag{2.12}$$

which is nonlinear in view of (2.11)*.* ∎

We investigate further the nature of the boundary of the admissible set given by (2.11) and (2.12). We have the following proposition.

PROPOSITION 2.2. *Suppose that* (2.10) *holds so that the resource is underloaded. Let $(\boldsymbol{\alpha}^{(0)}, \alpha_J^{(0)})$, where, corresponding to* (2.4)*, $\alpha_J^{(0)} = \alpha_J(\boldsymbol{\alpha}^{(0)}; C)$, be a point on the boundary of the admissible set. Define*

$$a_i(\tau) = d_J \tau^{d_J} \left(\tau^{d_i} - 1\right) - d_i \tau^{d_i} \left(\tau^{d_J} - 1\right), \tag{2.13}$$

and let $\tau_0 > 1$ be the solution of (2.11) *corresponding to $\alpha_j = \alpha_j^{(0)}$, $j = 1, \ldots, J$. Also, let*

$$\xi_J = \left(\tau_0^{d_J} - 1\right) \left[\alpha_J - \alpha_J^{(0)} - \sum_{i=1}^{J-1} \frac{\partial \alpha_J}{\partial \alpha_i} \left(\boldsymbol{\alpha}^{(0)}\right) \left(\alpha_i - \alpha_i^{(0)}\right)\right] \tag{2.14}$$

and

$$\xi_{J-1} = \sum_{i=1}^{J-1} a_i(\tau_0)\left(\alpha_i - \alpha_i^{(0)}\right). \tag{2.15}$$

The tangent hyperplane to the boundary of the admissible set at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(0)}$ *is given by* $\xi_J = 0$. *From* (2.11) *and* (2.12), *it follows that*

$$\left(\tau^{d_J} - 1\right)\frac{\partial \alpha_J}{\partial \alpha_i} = -\left(\tau^{d_i} - 1\right) + O\left(\frac{1}{C}\right), \quad i = 1, \ldots, J-1, \tag{2.16}$$

and hence, from (2.14), *that*

$$\xi_J = \sum_{j=1}^{J}\left(\tau_0^{d_j} - 1\right)\left(\alpha_j - \alpha_j^{(0)}\right) + O\left(\frac{1}{C}\right). \tag{2.17}$$

Let

$$v_0 = \sum_{j=1}^{J} d_j^2 \alpha_j^{(0)} \tau_0^{d_j}. \tag{2.18}$$

If $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$, *so that* $\xi_{J-1} = O(\epsilon)$, *then*

$$\xi_J = \xi_{J-1}^2 \left[\frac{1}{2v_0\left(\tau_0^{d_J} - 1\right)^2} + O(\epsilon)\right] + O\left(\frac{\epsilon^2}{C}\right). \tag{2.19}$$

Hence in the asymptotic limit $C \to \infty$, *the boundary of the admissible set is convex, but not strictly so, except when* $J = 2$. □

We remark that, in the limit $C \to \infty$, from (2.11) and (2.12), $\tau \equiv \tau_0$ corresponds to the intersection of two hyperplanes of dimension $J-1$, i.e., to a hyperplane of dimension $J-2$. In particular, if $J = 3$, $\tau \equiv \tau_0$ corresponds to a straight line, and the boundary of the admissible set is a ruled surface.

Proposition 2.2 is established in section 4. Results for the overloaded regime, the subject of Proposition 2.1(A), are derived in section 5. The results for the critically loaded regime in (B) are proven in section 6.

3. Refined results. We now consider refined approximations to the boundary of the admissible set. Corresponding to Proposition 2.1(A), which is for the overloaded regime, we have the following proposition.

PROPOSITION 3.1. *If* ℓ_j, $j = 1, \ldots, J$, *is asymptotically bounded away from zero, we consider the case when*

$$C(a_J - a_i) \gg 1, \quad i = 1, \ldots, J-1, \tag{3.1}$$

where a_j *is given by* (2.5). *With this assumption the boundary of the admissible set is given by* $L_J = \ell_J$. *Then, we have the following.*

(i) *If $J = 2$, then $0 > d\alpha_2/d\alpha_1 = O(1)$ and $0 > d^2\alpha_2/d\alpha_1^2 = O(1/C)$, so that the boundary of the admissible set is* concave*, although only slightly.*

(ii) *If $J \geq 3$, the linear transformation of variables*

$$\eta_J = d_J a_J^{d_J} \left[\alpha_J - \alpha_J^{(0)} - \sum_{i=1}^{J-1} \frac{\partial \alpha_J}{\partial \alpha_i} \left(\boldsymbol{\alpha}^{(0)} \right) \left(\alpha_i - \alpha_i^{(0)} \right) \right], \tag{3.2}$$

$$\eta_{J-1} = \sum_{i=1}^{J-1} d_i (d_i - d_J) a_J^{d_i} \left(\alpha_i - \alpha_i^{(0)} \right), \tag{3.3}$$

$$\eta_{J-2} = \sum_{i=1}^{J-1} d_i \left(d_i^2 - d_J^2 \right) a_J^{d_i} \left(\alpha_i - \alpha_i^{(0)} \right), \tag{3.4}$$

and

$$\eta_i = \alpha_i - \alpha_i^{(0)}, \quad i = 1, \ldots, J-3 \ (J \geq 4), \tag{3.5}$$

is nonsingular, and $\eta_J = 0$ corresponds to the tangent hyperplane to the boundary of the admissible set at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(0)}$. From (2.7) *and* (3.2) *we have*

$$\eta_J = \sum_{j=1}^{J} d_j a_J^{d_j} \left(\alpha_j - \alpha_j^{(0)} \right) + O\left(\frac{1}{C} \right). \tag{3.6}$$

If $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$, then

$$\eta_J = \frac{1}{2C \left(\displaystyle\sum_{j=1}^{J} \alpha_j^{(0)} d_j^2 a_J^{d_j} \right)^2} \left[\frac{\displaystyle\sum_{j=1}^{J} \alpha_j^{(0)} d_j^3 a_J^{d_j}}{\displaystyle\sum_{j=1}^{J} \alpha_j^{(0)} d_j^2 a_J^{d_j}} \eta_{J-2}^2 - \eta_{J-1} \eta_{J-2} \right] + O\left(\frac{\epsilon^2}{C^2} \right) + O\left(\frac{\epsilon^3}{C} \right). \tag{3.7}$$

Hence, asymptotically, the boundary of the admissible set is non*convex, although only slightly so.* ∎

Next, corresponding to Proposition 2.1(B), we have the following proposition.

PROPOSITION 3.2. *If* (2.8) *holds, then we have the following:*

(i) *If $J = 2$, then $0 > d\alpha_2/d\alpha_1 = O(1)$ and $0 < d^2\alpha_2/d\alpha_1^2 = O(1/\sqrt{C})$, so that the boundary of the admissible set is convex, although only slightly so.*

(ii) *We consider the case when*

$$\frac{d_J}{\sqrt{C}\,\ell_J} - \frac{d_i}{\sqrt{C}\,\ell_i} \gg \frac{1}{\sqrt{C}}, \quad i = 1, \ldots, J-1. \tag{3.8}$$

If $J \geq 3$, the linear transformation of variables

$$\zeta_J = d_J \left[\alpha_J - \alpha_J^{(0)} - \sum_{i=1}^{J-1} \frac{\partial \alpha_J}{\partial \alpha_i} \left(\boldsymbol{\alpha}^{(0)} \right) \left(\alpha_i - \alpha_i^{(0)} \right) \right], \tag{3.9}$$

$$\zeta_{J-1} = \sum_{i=1}^{J-1} d_i (d_i - d_J) \left(\alpha_i - \alpha_i^{(0)} \right), \tag{3.10}$$

$$\zeta_{J-2} = \sum_{i=1}^{J-1} d_i \left(d_i^2 - d_J^2 \right) \left(\alpha_i - \alpha_i^{(0)} \right), \tag{3.11}$$

and

$$\zeta_i = \alpha_i - \alpha_i^{(0)}, \quad i = 1, \ldots, J-3 \ (J \geq 4), \tag{3.12}$$

is nonsingular, and $\zeta_J = 0$ corresponds to the tangent hyperplane to the boundary of the admissible set at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(0)}$. From (2.9) *and* (3.9) *we have*

$$\zeta_J = \sum_{j=1}^{J} d_j \left(\alpha_j - \alpha_j^{(0)} \right) + O\left(\frac{1}{\sqrt{C}} \right). \tag{3.13}$$

Let

$$\sigma_0^2 = 2 \sum_{j=1}^{J} d_j^2 \alpha_j^{(0)}, \quad \sigma_0 > 0. \tag{3.14}$$

If $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$, then

$$\begin{aligned} \zeta_J = {} & \frac{\zeta_{J-1}^2}{2\sqrt{C}} \left[P_2(\sigma_0) + O\left(\frac{1}{\sqrt{C}} \right) + O\left(\zeta_{J-1} \right) \right] \\ & + \frac{\zeta_{J-1} \zeta_{J-2}}{C} \left[R_2(\sigma_0) + O(\zeta_{J-1}) \right] + O\left(\frac{\epsilon^2}{C\sqrt{C}} \right), \end{aligned} \tag{3.15}$$

and $P_2(\sigma_0) > 0$. □

If $\zeta_{J-1} = O(\epsilon/\sqrt{C})$, then the leading terms in (3.15) are all $O(\epsilon^2/C\sqrt{C})$, and the term involving $\zeta_{J-2}^2/C\sqrt{C}$, in particular, is needed to ascertain whether or not the boundary of the admissible set is convex. Hence, the next order term in the asymptotic expansion in powers of $1/\sqrt{C}$ is required, and this is derived in [12].

Finally, corresponding to Proposition 2.1(C), we have the following proposition.

PROPOSITION 3.3. *If* (2.10) *holds, we let $\tau_0 > 1$ be the solution of* (2.11) *corresponding to $\alpha_j = \alpha_j^{(0)}$, $j = 1, \ldots, J$, and define*

$$B(\tau) = \min_j \left[\frac{\beta_j (\tau - 1)}{(\tau^{d_j} - 1)} \right], \quad \tau > 1. \tag{3.16}$$

In general, $B(\tau)$ is only piecewise differentiable. We consider the case when $B(\tau)$ is differentiable at $\tau = \tau_0$. For $J \geq 3$, with $a_i(\tau)$ given by (2.13)*, the linear transformation of variables* (2.14), (2.15),

$$\xi_{J-2} = \sum_{i=1}^{J-1} a_i'(\tau_0) \left(\alpha_i - \alpha_i^{(0)}\right), \tag{3.17}$$

where the prime denotes derivative, and

$$\xi_i = \alpha_i - \alpha_i^{(0)}, \quad i = 1, \ldots, J-3 \ (J \geq 4), \tag{3.18}$$

is nonsingular. If $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$, then

$$\begin{aligned} \xi_J = \xi_{J-1}^2 &\left[\frac{1}{2v_0\left(\tau_0^{d_J} - 1\right)^2} + O(\epsilon) + O\left(\frac{1}{C}\right)\right] \\ &- \frac{\tau_0^2 \xi_{J-2}^2}{4Cv_0^2\left(\tau_0^{d_J} - 1\right)^2} + O\left(\frac{\epsilon}{C}\xi_{J-1}\right) + O\left(\frac{\epsilon^2}{C^2}\right). \end{aligned} \tag{3.19}$$

Hence, asymptotically, the boundary of the admissible set is non*convex, although only slightly so.* ∎

4. Underloaded regime. We establish Proposition 2.2. We consider an underloaded resource, so that

$$\sum_{j=1}^{J} d_j \alpha_j < 1. \tag{4.1}$$

Let z^* be the unique positive solution of

$$\sum_{j=1}^{J} d_j \alpha_j (z^*)^{d_j} = 1. \tag{4.2}$$

Then, from (4.1), $z^* > 1$. We define

$$f(z) = \sum_{j=1}^{J} \alpha_j (z^{d_j} - 1) - \log z \tag{4.3}$$

and

$$v(z^*) = \sum_{j=1}^{J} \alpha_j d_j^2 (z^*)^{d_j}. \tag{4.4}$$

Then (see [10]), the blocking probabilities satisfy

$$L_j = \frac{e^{Cf(z^*)}}{\sqrt{2\pi C v(z^*)}} \left\{\frac{[(z^*)^{d_j} - 1]}{(z^* - 1)} + O\left(\frac{1}{C}\right)\right\}. \tag{4.5}$$

Since, from (4.2) and (4.3), $f'(z^*) = 0$, it follows that

$$\frac{\partial f(z^*)}{\partial \alpha_k} = (z^*)^{d_k} - 1 > 0, \tag{4.6}$$

and (4.5) implies (2.3).

Let

$$\psi(z) = \sum_{j=1}^{J} \alpha_j \left(d_j z^{d_j} \log z + 1 - z^{d_j} \right). \tag{4.7}$$

Then, from (4.2) and (4.3), $\psi(z^*) = -f(z^*)$, and

$$\psi'(z) = \log z \sum_{j=1}^{J} \alpha_j d_j^2 z^{d_j - 1} > 0, \quad z > 1. \tag{4.8}$$

From (2.2), (2.10), and (4.5), it follows that $\psi(z^*) \geq \omega + O(1/C)$. Hence $z^* \geq \tau + O(1/C)$, where $\psi(\tau) = \omega$, so that (2.11) holds, and the boundary of the admissible set satisfies (2.12). We consider α_J, and hence τ, as functions of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{J-1})$ and C. From (2.11) and (2.12) we obtain

$$\sum_{j=1}^{J} \alpha_j \left(1 - \tau^{d_j}\right) + \log \tau = \omega + O\left(\frac{1}{C}\right), \tag{4.9}$$

and hence (2.16).

Now, from (2.12), for $k = 1, \ldots, J-1$,

$$d_k \tau^{d_k} + d_J \tau^{d_J} \frac{\partial \alpha_J}{\partial \alpha_k} + \sum_{j=1}^{J} \alpha_j d_j^2 \tau^{d_j - 1} \frac{\partial \tau}{\partial \alpha_k} = O\left(\frac{1}{C}\right). \tag{4.10}$$

We let

$$v = \sum_{j=1}^{J} \alpha_j d_j^2 \tau^{d_j}. \tag{4.11}$$

Then, from (2.13), (2.16), and (4.10), we have

$$\frac{\partial \tau}{\partial \alpha_k} = \frac{\tau a_k(\tau)}{v(\tau^{d_J} - 1)} + O\left(\frac{1}{C}\right), \quad k = 1, \ldots, J-1. \tag{4.12}$$

Also, from (2.13),

$$\tau \frac{d}{d\tau} \left[\frac{(\tau^{d_i} - 1)}{(\tau^{d_J} - 1)} \right] = -\frac{a_i(\tau)}{(\tau^{d_J} - 1)^2}. \tag{4.13}$$

Hence, from (2.16), (4.12), and (4.13), we obtain

$$\frac{\partial^2 \alpha_J}{\partial \alpha_i \partial \alpha_k} = \frac{a_i(\tau) a_k(\tau)}{v(\tau^{d_J} - 1)^3} + O\left(\frac{1}{C}\right), \quad i, k = 1, \ldots, J-1. \tag{4.14}$$

From (2.13), we have

$$\frac{d}{dt}\left[\frac{a_i(t)}{t^{d_J+d_i}}\right] = \frac{d_i d_J}{t}\left(\frac{1}{t^{d_i}} - \frac{1}{t^{d_J}}\right), \tag{4.15}$$

which is nonzero for $t > 1$ and $i = 1, \ldots, J-1$, since d_j, $j = 1, \ldots, J$, are distinct. Hence $a_i(\tau) \neq 0$, since $\tau > 1$ and $a_i(1) = 0$. With ξ_{J-1} defined by (2.15), it follows, by induction, from (4.12) and (4.14), that, for $n \geq 2$, the leading term in

$$\sum_{i_1=1}^{J-1} \cdots \sum_{i_n=1}^{J-1} \frac{\partial^n \alpha_J}{\partial\alpha_{i_1} \ldots \partial\alpha_{i_n}}(\alpha_1^{(0)}, \ldots, \alpha_{J-1}^{(0)}) \prod_{m=1}^{n} \left(\alpha_{i_m} - \alpha_{i_m}^{(0)}\right) \tag{4.16}$$

contains the factor ξ_{J-1}^2. If we use the Taylor series expansion

$$\begin{aligned} \alpha_J = \alpha_J^{(0)} &+ \sum_{i=1}^{J-1} \frac{\partial \alpha_J}{\partial \alpha_i}\left(\boldsymbol{\alpha}^{(0)}\right)\left(\alpha_i - \alpha_i^{(0)}\right) \\ &+ \frac{1}{2}\sum_{i=1}^{J-1}\sum_{k=1}^{J-1} \frac{\partial^2 \alpha_J}{\partial\alpha_i \partial\alpha_k}\left(\boldsymbol{\alpha}^{(0)}\right)\left(\alpha_i - \alpha_i^{(0)}\right)\left(\alpha_k - \alpha_k^{(0)}\right) + \cdots \end{aligned} \tag{4.17}$$

and let $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$, then, from (2.14), (2.18), and (4.14), we obtain (2.19), where $\tau_0 > 1$ is the solution of (2.11) corresponding to $\alpha_j = \alpha_j^{(0)}$, $j = 1, \ldots, J$. This establishes Proposition 2.2.

5. Overloaded regime. We now consider an overloaded resource, so that

$$\sum_{j=1}^{J} d_j \alpha_j > 1. \tag{5.1}$$

Then, from (4.2), $0 < z^* < 1$, and [10], the blocking probabilities satisfy

$$L_j = 1 - (z^*)^{d_j} + O(1/C). \tag{5.2}$$

However, from (4.2), $\partial z^*/\partial\alpha_k < 0$, $k = 1, \ldots, J$. It follows, from (5.2), that (2.3) holds. Moreover, from (2.2), (2.5), and (2.6), the boundary of the admissible set satisfies $z^* = a_J + O(1/C)$, and (4.2) implies (2.7). Consequently, $\partial^2\alpha_J/\partial\alpha_i\partial\alpha_k = O(1/C)$, and it is necessary to consider the first order correction term in (5.2).

In Appendix A we establish the following proposition.

PROPOSITION 5.1. *Let*

$$t(z^*) = \sum_{j=1}^{J} \alpha_j d_j^2 (d_j - 3)(z^*)^{d_j}. \tag{5.3}$$

Then, with $v(z^*)$ given by (4.4),

$$L_j = 1 - (z^*)^{d_j} + \frac{d_j (z^*)^{d_j}}{2Cv(z^*)}\left[d_j - 5 + \frac{2}{(1-z^*)} - \frac{t(z^*)}{v(z^*)}\right] + O\left(\frac{1}{C^2}\right). \tag{5.4}$$

∎

It follows from (2.2), (2.5), and (5.4) that

$$z^* \geq a_j \left\{ 1 + \frac{1}{2Cv(a_j)} \left[d_j - 5 + \frac{2}{(1-a_j)} - \frac{t(a_j)}{v(a_j)} \right] + O\left(\frac{1}{C^2}\right) \right\}, \tag{5.5}$$

$j = 1, \ldots, J$. Under the assumption (3.1), the maximum in (5.5) is obtained for $j = J$ and, from (4.2), the boundary of the admissible set satisfies

$$\sum_{j=1}^{J} \alpha_j d_j a_J^{d_j} \left\{ 1 + \frac{d_j}{2Cv(a_J)} \left[d_J - 5 + \frac{2}{(1-a_J)} - \frac{t(a_J)}{v(a_J)} \right] + O\left(\frac{1}{C^2}\right) \right\} = 1. \tag{5.6}$$

In view of (4.4), this may be written in the form

$$\sum_{j=1}^{J} \alpha_j d_j a_J^{d_j} + \frac{1}{2C} \left[d_J - 5 + \frac{2}{(1-a_J)} - \frac{t(a_J)}{v(a_J)} \right] + O\left(\frac{1}{C^2}\right) = 1. \tag{5.7}$$

We consider α_J as a function of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{J-1})$ and C, as in (2.4). Then, from (4.4), for $i = 1, \ldots, J-1$,

$$\frac{\partial v(a_J)}{\partial \alpha_i} = d_i^2 a_J^{d_i} + d_J^2 a_J^{d_J} \frac{\partial \alpha_J}{\partial \alpha_i}, \tag{5.8}$$

and, from (5.3),

$$\frac{\partial t(a_J)}{\partial \alpha_i} = d_i^2 (d_i - 3) a_J^{d_i} + d_J^2 (d_J - 3) a_J^{d_J} \frac{\partial \alpha_J}{\partial \alpha_i}. \tag{5.9}$$

Hence, from (2.7),

$$\frac{\partial v(a_J)}{\partial \alpha_i} = d_i (d_i - d_J) a_J^{d_i} + O\left(\frac{1}{C}\right), \tag{5.10}$$

and

$$\frac{\partial t(a_J)}{\partial \alpha_i} = d_i (d_i - d_J)(d_i + d_J - 3) a_J^{d_i} + O\left(\frac{1}{C}\right). \tag{5.11}$$

It follows from (5.7) that

$$\begin{aligned} & d_i a_J^{d_i} + d_J a_J^{d_J} \frac{\partial \alpha_J}{\partial \alpha_i} \\ & \quad = \frac{d_i (d_i - d_J)}{2Cv(a_J)} a_J^{d_i} \left[(d_i + d_J - 3) - \frac{t(a_J)}{v(a_J)} \right] + O\left(\frac{1}{C^2}\right). \end{aligned} \tag{5.12}$$

Finally, from (5.10)–(5.12), for $i, k = 1, \ldots, J-1$, we obtain

$$\begin{aligned} d_J a_J^{d_J} \frac{\partial^2 \alpha_J}{\partial \alpha_i \partial \alpha_k} = {} & \frac{d_i d_k (d_i - d_J)(d_k - d_J)}{2C[v(a_J)]^3} a_J^{d_i + d_k} \\ & \cdot \left[2t(a_J) + (6 - d_i - d_k - 2d_J) v(a_J) \right] + O\left(\frac{1}{C^2}\right). \end{aligned} \tag{5.13}$$

If $J = 2$, then

$$t(a_2) + (3 - d_1 - d_2)v(a_2) = -d_1 d_2 \left(\alpha_1 d_1 a_2^{d_1} + \alpha_2 d_2 a_2^{d_2} \right) < 0, \tag{5.14}$$

and Proposition 3.1(i) follows from (2.7) and (5.13). If $J \geq 3$, we use the Taylor series expansion (4.17). We note, from (4.4) and (5.3), that

$$t(a_J) + 3v(a_J) = \sum_{j=1}^{J} \alpha_j d_j^3 a_J^{d_j}. \tag{5.15}$$

From (5.13), (5.15), and the transformation of variables (3.2)–(3.5), which is nonsingular since $d_1, \ldots, d_J$ are distinct, we obtain (3.7) and hence Proposition 3.1(ii).

6. Critically loaded regime. We now consider a critically loaded resource, so that

$$\sum_{j=1}^{J} d_j \alpha_j = 1 - \frac{\delta}{\sqrt{C}}, \tag{6.1}$$

where $\delta = O(1)$ may have either sign. Let

$$\sigma^2 = 2 \sum_{j=1}^{J} d_j^2 \alpha_j, \quad \sigma > 0, \tag{6.2}$$

and

$$\beta = \frac{2e^{-(\delta/\sigma)^2}}{\sigma \sqrt{\pi}\ \mathrm{Erfc}\,(-\delta/\sigma)}, \tag{6.3}$$

where the complementary error function is given by

$$\mathrm{Erfc}\,(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-\xi^2} d_\xi. \tag{6.4}$$

Then (see [3], [10], [16]), the blocking probabilities satisfy

$$L_j = \frac{d_j \beta}{\sqrt{C}} \left[1 + O\left(\frac{1}{\sqrt{C}} \right) \right]. \tag{6.5}$$

It follows, from (6.1)–(6.5), that

$$\frac{\partial L_j}{\partial \alpha_k} = \frac{4 d_j d_k e^{-(\delta/\sigma)^2}}{\sigma^2 \sqrt{\pi}\ \mathrm{Erfc}\,(-\delta/\sigma)} \left[\frac{\delta}{\sigma} + \frac{e^{-(\delta/\sigma)^2}}{\sqrt{\pi}\ \mathrm{Erfc}\,(-\delta/\sigma)} \right] + O\left(\frac{1}{\sqrt{C}} \right). \tag{6.6}$$

The quantity in square brackets in (6.6) was shown [14] to be positive, and hence (2.3) holds.

As in [10], we define

$$\kappa = \max_j \frac{d_j}{\sqrt{C} \ell_j}, \tag{6.7}$$

which is $O(1)$, in view of (2.8), and

$$\phi(y) = \frac{\sqrt{\pi}}{2} e^{y^2} \text{ Erfc}\,(-y). \tag{6.8}$$

Then the admissible set satisfies

$$\sigma\phi(\delta/\sigma) \geq \kappa + O(1/\sqrt{C}). \tag{6.9}$$

However, from (6.4) and (6.8),

$$\phi(y) = e^{y^2} \int_{-y}^{\infty} e^{-\xi^2} d\xi = \int_0^{\infty} e^{2yu} e^{-u^2} du. \tag{6.10}$$

Hence $\phi'(y) > 0$, $-\infty < y < \infty$, so that $\phi(y)$ has a unique inverse, and (6.9) may be written, as in [10], in the form $\delta \geq \sigma\phi^{-1}(\kappa/\sigma) + O(1/\sqrt{C})$. We define

$$\chi(\sigma) = \phi^{-1}(\kappa/\sigma). \tag{6.11}$$

Then, from (6.1), the boundary of the admissible set satisfies

$$\sum_{j=1}^{J} d_j \alpha_j = 1 - \frac{\sigma}{\sqrt{C}}\, \chi(\sigma) + O\left(\frac{1}{C}\right). \tag{6.12}$$

This establishes Proposition 2.1(B) and provides an asymptotic approximation to the error term in (2.9).

We now consider α_J as a function of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{J-1})$ and C, as in (2.4). We define

$$\theta^2 = 2\left[d_J + \sum_{i=1}^{J-1} d_i(d_i - d_J)\alpha_i\right], \quad \theta > 0. \tag{6.13}$$

Then, from (6.2) and (6.12), $\sigma = \theta + O(1/\sqrt{C})$, and

$$d_J \alpha_J = 1 - \sum_{i=1}^{J-1} d_i \alpha_i - \frac{\theta}{\sqrt{C}}\, \chi(\theta) + O\left(\frac{1}{C}\right). \tag{6.14}$$

It follows, from (6.13) and (6.14), that

$$\theta \frac{\partial \theta}{\partial \alpha_i} = d_i(d_i - d_J), \quad i = 1, \ldots, J-1, \tag{6.15}$$

and

$$\frac{\partial \alpha_J}{\partial \alpha_i} = -\frac{d_i}{d_J}\left\{1 + \frac{(d_i - d_J)}{\sqrt{C}}\left[\chi'(\theta) + \frac{\chi(\theta)}{\theta}\right]\right\} + O\left(\frac{1}{C}\right) < 0, \tag{6.16}$$

where the prime denotes derivative. Also, for $i = 1, \ldots, J-1$ and $k = 1, \ldots, J-1$,

$$d_J \frac{\partial^2 \alpha_J}{\partial \alpha_i \partial \alpha_k} = d_i d_k (d_i - d_J)(d_k - d_J)\frac{P_2(\theta)}{\sqrt{C}} + O\left(\frac{1}{C}\right), \tag{6.17}$$

where

$$P_2(\theta) = \frac{1}{\theta^3}\left[\chi(\theta) - \theta\chi'(\theta) - \theta^2\chi''(\theta)\right]. \tag{6.18}$$

It is shown in Appendix B that $P_2(\theta) > 0$ for $\theta > 0$. Hence, if $J = 2$, we obtain Proposition 3.2(i).

From (6.15) and (6.17) it follows, by induction, that the higher order partial derivatives have the form

$$d_J \frac{\partial^n \alpha_J}{\partial\alpha_{i_1}\cdots\partial\alpha_{i_n}} = \frac{1}{\sqrt{C}}\prod_{m=1}^{n} d_{i_m}(d_{i_m} - d_J)P_n(\theta) + O\left(\frac{1}{C}\right) \tag{6.19}$$

for $n \geq 2$, where $P_{n+1}(\theta) = P_n'(\theta)/\theta$. If $J \geq 3$, we use the Taylor series expansion (4.17) and introduce the linear transformation of variables (3.9)–(3.12). The transformation is nonsingular, since $d_1, \ldots, d_J$ are distinct. If $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$, then, from (6.19), we obtain

$$\zeta_J = \frac{\zeta_{J-1}^2}{2\sqrt{C}}\left[P_2(\theta_0) + O(\zeta_{J-1})\right] + O\left(\frac{\epsilon^2}{C}\right), \tag{6.20}$$

where $\theta_0 = \theta(\boldsymbol{\alpha}^{(0)})$, and we note that $\theta_0 = \sigma_0 + O(1/\sqrt{C})$, where $\sigma_0 = \sigma(\boldsymbol{\alpha}^{(0)})$ is given by (3.14). Hence, it is necessary to consider a refined asymptotic approximation to ascertain the nature of the boundary of the admissible set when ζ_{J-1} is close to zero.

7. Refined critically loaded approximation. Let

$$\eta = \sum_{j=1}^{J} d_j^3 \alpha_j. \tag{7.1}$$

Then, a refined asymptotic approximation to the blocking probabilities in the critically loaded regime is (see [14])

$$L_j = \frac{d_j\beta}{\sqrt{C}}\left\{1 + \frac{\delta}{\sigma^2\sqrt{C}}\left[\frac{2\eta}{\sigma^2}\left(\frac{2\delta^2}{3\sigma^2} - 1\right) + d_j - 1\right]\right. \tag{7.2}$$
$$\left. - \frac{\beta}{2\sqrt{C}}\left[1 + \frac{2\eta}{3\sigma^2}\left(1 - \frac{2\delta^2}{\sigma^2}\right)\right] + O\left(\frac{1}{C}\right)\right\}.$$

We consider the case when (3.8) holds, so that, from (6.7),

$$\kappa = \frac{d_J}{\sqrt{C}\ell_J}, \tag{7.3}$$

which is $O(1)$. Then, asymptotically, the boundary of the admissible set satisfies

$$1 = \kappa\beta\left\{1 + \frac{\delta}{\sigma^2\sqrt{C}}\left[\frac{2\eta}{\sigma^2}\left(\frac{2\delta^2}{3\sigma^2} - 1\right) + d_J - 1\right]\right. \tag{7.4}$$
$$\left. - \frac{\beta}{2\sqrt{C}}\left[1 + \frac{2\eta}{3\sigma^2}\left(1 - \frac{2\delta^2}{\sigma^2}\right)\right] + O\left(\frac{1}{C}\right)\right\}.$$

Now, from (6.1), (6.2), (6.13), and (6.14),

$$\sigma = \theta + O(1/\sqrt{C}) \tag{7.5}$$

and

$$\delta = \theta\chi(\theta) + O(1/\sqrt{C}). \tag{7.6}$$

We define

$$\psi = d_J^2 + \sum_{i=1}^{J-1} d_i \left(d_i^2 - d_J^2\right) \alpha_i. \tag{7.7}$$

Then, from (6.1) and (7.1),

$$\eta = \psi + O(1/\sqrt{C}). \tag{7.8}$$

From (6.3) and (6.8), we have

$$\frac{1}{\beta} = \sigma\phi\left(\frac{\delta}{\sigma}\right). \tag{7.9}$$

Hence, from (7.4)–(7.6), (7.8), and (7.9), it follows that $\beta = 1/\kappa + O(1/\sqrt{C})$, and

$$\begin{aligned}\sigma\phi\left(\frac{\delta}{\sigma}\right) = \kappa\Bigg\{1 + \frac{\chi(\theta)}{\theta\sqrt{C}}\left[\frac{2\psi}{\theta^2}\left\{\frac{2}{3}[\chi(\theta)]^2 - 1\right\} + d_J - 1\right] \\ - \frac{1}{2\kappa\sqrt{C}}\left[1 + \frac{2\psi}{3\theta^2}\left\{1 - 2[\chi(\theta)]^2\right\}\right] + O\left(\frac{1}{C}\right)\Bigg\}.\end{aligned} \tag{7.10}$$

From (6.2) and (6.14), it follows that

$$\sigma^2 = \theta^2 - \frac{2}{\sqrt{C}} d_J \theta\chi(\theta) + O\left(\frac{1}{C}\right), \tag{7.11}$$

so that

$$\sigma = \theta - \frac{d_J}{\sqrt{C}}\chi(\theta) + O\left(\frac{1}{C}\right) \tag{7.12}$$

and

$$\sigma\chi(\sigma) = \theta\chi(\theta) - \frac{d_J}{\sqrt{C}}\chi(\theta)\left[\chi(\theta) + \theta\chi'(\theta)\right] + O\left(\frac{1}{C}\right). \tag{7.13}$$

Hence, from (6.11) and (7.10), it is found that

$$\begin{aligned}\delta = \theta\chi(\theta) - \frac{d_J}{\sqrt{C}}\chi(\theta)\left[\chi(\theta) + \theta\chi'(\theta)\right] \\ + \frac{\chi'(\theta)}{\sqrt{C}}\left\{\frac{1}{2\kappa}\left[\theta^2 + \frac{2}{3}\psi\left\{1 - 2\left[\chi(\theta)\right]^2\right\}\right]\right. \\ \left. - \theta\chi(\theta)\left[\frac{2\psi}{\theta^2}\left\{\frac{2}{3}[\chi(\theta)]^2 - 1\right\} + d_J - 1\right]\right\} + O\left(\frac{1}{C}\right).\end{aligned} \tag{7.14}$$

We define

$$F(\theta) = \theta\chi(\theta)\chi'(\theta) + \frac{\theta^2}{2\kappa}\chi'(\theta) - d_J\chi(\theta)[\chi(\theta) + 2\theta\chi'(\theta)] \tag{7.15}$$

and

$$G(\theta) = \chi'(\theta)\left[\frac{2\chi(\theta)}{\theta}\left\{1 - \frac{2}{3}[\chi(\theta)]^2\right\} + \frac{1}{3\kappa}\{1 - 2[\chi(\theta)]^2\}\right]. \tag{7.16}$$

Then, from (6.1) and (7.14)–(7.16), we obtain

$$d_J\alpha_J = 1 - \sum_{i=1}^{J-1} d_i\alpha_i - \frac{\theta}{\sqrt{C}}\chi(\theta) - \frac{1}{C}[F(\theta) + \psi G(\theta)] + O\left(\frac{1}{C^{3/2}}\right). \tag{7.17}$$

However, from (7.7),

$$\frac{\partial\psi}{\partial\alpha_i} = d_i\left(d_i^2 - d_J^2\right), \quad i = 1, \ldots, J-1. \tag{7.18}$$

Hence, from (6.15), (7.17), and (7.18),

$$\begin{aligned}\frac{\partial\alpha_J}{\partial\alpha_i} = &-\frac{d_i}{d_J}\left\{1 + \frac{(d_i - d_J)}{\sqrt{C}}\left[\chi'(\theta) + \frac{\chi(\theta)}{\theta}\right]\right.\\ &\left. + \ \frac{(d_i - d_J)}{C\theta}[F'(\theta) + \psi G'(\theta)] + \left(d_i^2 - d_J^2\right)\frac{G(\theta)}{C}\right\} + O\left(\frac{1}{C^{3/2}}\right).\end{aligned} \tag{7.19}$$

We now define

$$Q_2(\theta) = \frac{F'(\theta)}{\theta^3} - \frac{F''(\theta)}{\theta^2}, \qquad R_2(\theta) = -\frac{G'(\theta)}{\theta}. \tag{7.20}$$

Then, with $P_2(\theta)$ as in (6.18), for $i = 1, \ldots, J-1$ and $k = 1, \ldots, J-1$, we obtain

$$\begin{aligned}d_J\frac{\partial^2\alpha_J}{\partial\alpha_i\partial\alpha_k} = &\ d_id_k(d_i - d_J)(d_k - d_J)\left\{\frac{P_2(\theta)}{\sqrt{C}} + \frac{1}{C}\left[Q_2(\theta) + \frac{\psi}{\theta}R_2'(\theta)\right]\right\}\\ &+ \ d_id_k[(d_i - d_J)(d_k^2 - d_J^2) + (d_k - d_J)(d_i^2 - d_J^2)]\frac{R_2(\theta)}{C} + O\left(\frac{1}{C^{3/2}}\right).\end{aligned} \tag{7.21}$$

From (6.15), (7.18), and (7.21) it follows, by induction, that the higher order partial derivatives have the form

$$\begin{aligned}d_J\frac{\partial^n\alpha_J}{\partial\alpha_{i_1}\cdots\partial\alpha_{i_n}} = &\prod_{m=1}^{n} d_{i_m}(d_{i_m} - d_J)\left\{\frac{P_n(\theta)}{\sqrt{C}} + \frac{1}{C}\left[Q_n(\theta) + \frac{\psi}{\theta}R_n'(\theta)\right]\right\}\\ &+ \ \sum_{r=1}^{n}\prod_{\substack{m=1\\ m\neq r}}^{n} d_{i_m}(d_{i_m} - d_J)d_{i_r}(d_{i_r}^2 - d_J^2)\frac{R_n(\theta)}{C} + O\left(\frac{1}{C^{3/2}}\right)\end{aligned} \tag{7.22}$$

for $n \geq 2$, where

$$P_{n+1}(\theta) = P_n'(\theta)/\theta, \quad Q_{n+1}(\theta) = Q_n'(\theta)/\theta, \quad R_{n+1}(\theta) = R_n'(\theta)/\theta. \tag{7.23}$$

Hence, from (3.10) and (3.11), for $n \geq 2$,

$$d_J \sum_{i_1=1}^{J-1} \cdots \sum_{i_n=1}^{J-1} \frac{\partial^n \alpha_J}{\partial \alpha_{i_1} \cdots \partial \alpha_{i_n}}(\boldsymbol{\alpha}^{(0)}) \prod_{m=1}^{n} (\alpha_{i_m} - \alpha_{i_m}^{(0)}) \tag{7.24}$$

$$= \left\{ \frac{P_n(\theta_0)}{\sqrt{C}} + \frac{1}{C}\left[Q_n(\theta_0) + \frac{\psi_0}{\theta_0} R_n'(\theta_0) \right] \right\} \zeta_{J-1}^n + \frac{n}{C} R_n(\theta_0) \zeta_{J-1}^{n-1} \zeta_{J-2} + O\left(\frac{1}{C^{3/2}} \right),$$

where $\theta_0 = \theta(\boldsymbol{\alpha}^{(0)})$ and $\psi_0 = \psi(\boldsymbol{\alpha}^{(0)})$.

For $J \geq 3$, we use the Taylor series expansion (4.17) and introduce the linear transformation of variables (3.9)–(3.12). If $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$, then it follows from (7.24) that

$$\zeta_J = \frac{\zeta_{J-1}^2}{2\sqrt{C}} \left[P_2(\theta_0) + O\left(\frac{1}{\sqrt{C}} \right) + O(\zeta_{J-1}) \right] \tag{7.25}$$

$$+ \frac{1}{C} \zeta_{J-1} \zeta_{J-2} \left[R_2(\theta_0) + O(\zeta_{J-1}) \right] + O\left(\frac{\epsilon^2}{C\sqrt{C}} \right).$$

Since $\theta_0 = \sigma_0 + O(1/\sqrt{C})$, where $\sigma_0 = \sigma(\boldsymbol{\alpha}^{(0)})$ is given by (3.14), we obtain (3.15) and hence Proposition 3.2(ii).

8. Refined underloaded approximation. We now return to an underloaded resource. From (2.2), (2.10), and (4.5), since $\psi(z^*) = -f(z^*)$, it follows that

$$\psi(z^*) \geq \omega - \frac{1}{C} \log \left\{ \frac{\beta_j (z^* - 1)}{[(z^*)^{d_j} - 1]} \right\} - \frac{\log v(z^*)}{2C} + O\left(\frac{1}{C^2} \right), \tag{8.1}$$

$j = 1, \ldots, J$. Hence, $z^* \geq \tau + \rho/C + O(1/C^2)$, where, as in [10],

$$\rho \psi'(\tau) = -\log B(\tau) - \frac{1}{2} \log v, \tag{8.2}$$

where $B(\tau)$ and v are given by (3.16) and (4.11). However, from (4.8),

$$\sum_{j=1}^{J} \alpha_j d_j \left[\tau + \frac{\rho}{C} + O\left(\frac{1}{C^2} \right) \right]^{d_j} = \sum_{j=1}^{J} \alpha_j d_j \tau^{d_j} + \frac{\rho \psi'(\tau)}{C \log \tau} + O\left(\frac{1}{C^2} \right). \tag{8.3}$$

Hence, the boundary of the admissible set satisfies

$$\sum_{j=1}^{J} \alpha_j d_j \tau^{d_j} - \frac{1}{C \log \tau} \left[\log B(\tau) + \frac{1}{2} \log v \right] + O\left(\frac{1}{C^2} \right) = 1. \tag{8.4}$$

From (2.11) and (8.4), we obtain

$$\sum_{j=1}^{J} \alpha_j (1 - \tau^{d_j}) + \log \tau + \frac{1}{C} \left[\log B(\tau) + \frac{1}{2} \log v \right] + O\left(\frac{1}{C^2} \right) = \omega. \tag{8.5}$$

Hence, for $i = 1, \ldots, J-1$,

$$1 - \tau^{d_i} + (1 - \tau^{d_J})\frac{\partial \alpha_J}{\partial \alpha_i} + \frac{1}{2Cv}\,\frac{\partial v}{\partial \alpha_i} + \frac{1}{\tau}\,\frac{\partial \tau}{\partial \alpha_i}\left[1 - \sum_{j=1}^{J} \alpha_j d_j \tau^{d_j} + \frac{\tau B'(\tau)}{CB(\tau)}\right] = O\left(\frac{1}{C^2}\right). \tag{8.6}$$

However, from (4.11),

$$\frac{\partial v}{\partial \alpha_i} = d_i^2 \tau^{d_i} + d_J^2 \tau^{d_J}\frac{\partial \alpha_J}{\partial \alpha_i} + \sum_{j=1}^{J} \alpha_j d_j^3 \tau^{d_j - 1}\frac{\partial \tau}{\partial \alpha_i}. \tag{8.7}$$

Also, from (2.13),

$$\tau a_i'(\tau) = d_J^2 \tau^{d_J}(\tau^{d_i} - 1) - d_i^2 \tau^{d_i}(\tau^{d_J} - 1). \tag{8.8}$$

Hence, from (2.16) and (4.12),

$$\frac{\partial v}{\partial \alpha_i} = \frac{a_i(\tau)}{v(\tau^{d_J} - 1)} \sum_{j=1}^{J} \alpha_j d_j^3 \tau^{d_j} - \frac{\tau a_i'(\tau)}{(\tau^{d_J} - 1)} + O\left(\frac{1}{C}\right). \tag{8.9}$$

We let

$$A = \frac{1}{v(\tau^{d_J} - 1)^2}\left\{\frac{[\log B(\tau) + \frac{1}{2}\log v]}{\log \tau} - \frac{\tau B'(\tau)}{B(\tau)}\right\} \tag{8.10}$$

and

$$D = \frac{\tau}{2v(\tau^{d_J} - 1)^2}, \qquad E = -\frac{1}{2v^2(\tau^{d_J} - 1)^2}. \tag{8.11}$$

Then, from (4.12), (8.4), (8.6), and (8.9), we obtain

$$\frac{\partial \alpha_J}{\partial \alpha_i} = -\frac{(\tau^{d_i} - 1)}{(\tau^{d_J} - 1)} - \frac{1}{C}\left[\left(A + E\sum_{j=1}^{J} \alpha_j d_j^3 \tau^{d_j}\right) a_i(\tau) + D a_i'(\tau)\right] + O\left(\frac{1}{C^2}\right) \tag{8.12}$$

for $i = 1, \ldots, J-1$. However, from (2.11),

$$d_i \tau^{d_i}\log \tau + 1 - \tau^{d_i} + \frac{\partial \alpha_J}{\partial \alpha_i}\left(d_J \tau^{d_J}\log \tau + 1 - \tau^{d_J}\right) + \sum_{j=1}^{J} \alpha_j d_j^2 \tau^{d_j - 1}\log \tau \frac{\partial \tau}{\partial \alpha_i} = 0, \quad i = 1, \ldots, J-1. \tag{8.13}$$

It follows, from (2.13), (4.11), (8.12), and (8.13), that

$$\frac{v}{\tau}\,\frac{\partial \tau}{\partial \alpha_i} = \frac{a_i(\tau)}{(\tau^{d_J - 1})} + \frac{\left(d_J \tau^{d_J}\log \tau + 1 - \tau^{d_J}\right)}{C\log \tau} \cdot \left[\left(A + E\sum_{j=1}^{J} \alpha_j d_j^3 \tau^{d_j}\right) a_i(\tau) + D a_i'(\tau)\right] + O\left(\frac{1}{C^2}\right) \tag{8.14}$$

for $i = 1, \ldots, J-1$.

From (2.16) and (4.12), we have

$$(\tau^{d_J}-1)\frac{\partial}{\partial\alpha_k}\left(\sum_{j=1}^{J}\alpha_j d_j^3\tau^{d_j}\right) \tag{8.15}$$

$$= d_k^3\tau^{d_k}(\tau^{d_J}-1) - d_J^3\tau^{d_J}(\tau^{d_k}-1) + \frac{a_k(\tau)}{v}\sum_{j=1}^{J}\alpha_j d_j^4\tau^{d_j} + O\left(\frac{1}{C}\right).$$

However, from (8.8), it follows that

$$\tau[\tau a_k'(\tau)]' = d_J^3\tau^{d_J}(\tau^{d_k}-1) - d_k^3\tau^{d_k}(\tau^{d_J}-1) + d_k d_J(d_J-d_k)\tau^{d_J+d_k}. \tag{8.16}$$

Hence, from (2.13) and (8.8), we obtain

$$d_J^3\tau^{d_J}(\tau^{d_k}-1) - d_k^3\tau^{d_k}(\tau^{d_J}-1) \tag{8.17}$$

$$= \tau^2 a_k''(\tau) + \tau a_k'(\tau) - \frac{d_J\tau^{d_J}}{(\tau^{d_J}-1)}[\tau a_k'(\tau) - d_J a_k(\tau)].$$

We define

$$H = \frac{1}{v(\tau^{d_J}-1)^2}\left[\frac{d_J\tau^{d_J}}{(\tau^{d_J}-1)} - \frac{1}{2} - \frac{\log v}{2\log\tau} + \frac{\tau B'(\tau)}{B(\tau)} - \frac{\log B(\tau)}{\log\tau} + \frac{\tau}{v}\,\frac{\partial v}{\partial\tau}\right]. \tag{8.18}$$

Then, from (4.11), (8.10), and (8.11), it is found that

$$A + E\sum_{j=1}^{J}\alpha_j d_j^3\tau^{d_j} + \frac{\partial D}{\partial\tau} + \frac{\partial D}{\partial v}\sum_{j=1}^{J}\alpha_j d_j^3\tau^{d_j-1} = -H \tag{8.19}$$

and

$$\frac{D(d_J\tau^{d_J}\log\tau + 1 - \tau^{d_J})}{\tau(\tau^{d_J}-1)\log\tau} \tag{8.20}$$

$$+ v\left\{\frac{\partial A}{\partial v} + \frac{\partial E}{\partial v}\sum_{j=1}^{J}\alpha_j d_j^3\tau^{d_j} + E\left[1 - \frac{d_J\tau^{d_J}}{(\tau^{d_J}-1)}\right]\right\} = H.$$

Finally, from (4.13), (8.9), (8.11), (8.12), (8.14), (8.15), (8.17), (8.19), and (8.20), for $i,k = 1,\ldots,J-1$, we obtain

$$\frac{\partial^2\alpha_J}{\partial\alpha_i\partial\alpha_k} = \left[\frac{1}{v(\tau^{d_J}-1)^3} + O\left(\frac{1}{C}\right)\right]a_i(\tau)a_k(\tau) \tag{8.21}$$

$$- \frac{\tau^2[a_i''(\tau)a_k(\tau) + a_i'(\tau)a_k'(\tau) + a_i(\tau)a_k''(\tau)]}{2Cv^2(\tau^{d_J}-1)^3}$$

$$+ \frac{\tau H}{Cv(\tau^{d_J}-1)}[a_i'(\tau)a_k(\tau) + a_i(\tau)a_k'(\tau)] + O\left(\frac{1}{C^2}\right).$$

We assume now that $J \geq 3$. From (2.13) and (8.8), it is found that

$$\tau[a_i'(\tau)a_k(\tau) - a_i(\tau)a_k'(\tau)] = \tau^{d_J+d_i+d_k}(\tau^{d_J}-1)m_{ik}(\tau), \tag{8.22}$$

where

$$m_{ik}(t) = d_i d_k (d_i - d_k)\left(1 - \frac{1}{t^{d_J}}\right) + d_J d_i (d_J - d_i)\left(1 - \frac{1}{t^{d_k}}\right) - d_J d_k (d_J - d_k)\left(1 - \frac{1}{t^{d_i}}\right). \tag{8.23}$$

Hence,

$$t^{d_J+1} m'_{ik}(t) = d_J d_i d_k n_{ik}(t), \tag{8.24}$$

where

$$n_{ik}(t) = d_i - d_k + (d_J - d_i)t^{d_J - d_k} + (d_k - d_J)t^{d_J - d_i}. \tag{8.25}$$

Thus,

$$t n'_{ik}(t) = (d_J - d_i)(d_J - d_k)t^{d_J}\left(\frac{1}{t^{d_k}} - \frac{1}{t^{d_i}}\right). \tag{8.26}$$

Since d_j, $j = 1, \ldots, J$, are distinct, it follows, for $i \neq k$, $i, k = 1, \ldots, J-1$, and for $t > 1$, that $n'_{ik}(t) \neq 0$, which implies $n_{ik}(t) \neq 0$, since $n_{ik}(1) = 0$, i.e., $m'_{ik}(t) \neq 0$, which implies $m_{ik}(t) \neq 0$, since $m_{ik}(1) = 0$.

In particular, since $\tau > 1$, $m_{J-2,J-1}(\tau) \neq 0$. It follows from (8.22) that the linear transformation of variables (2.14), (2.15), (3.17), and (3.18) is nonsingular. We recall that, for $n \geq 2$, the leading term in (4.16) contains the factor ξ_{J-1}^2. Then, if we use the Taylor series expansion (4.17), and let $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$, we obtain (3.19) from (8.21), where v_0 is given by (2.18) and $\tau_0 > 1$ is the solution of (2.11) corresponding to $\alpha_j = \alpha_j^{(0)}$, $j = 1, \ldots, J$. This establishes Proposition 3.3.

9. Check on negative coefficient. As a check on the negative coefficient in (3.19), which pertains to $J \geq 3$, we now consider points on the boundary of the admissible set corresponding to $\tau(\boldsymbol{\alpha}) \equiv \tau(\boldsymbol{\alpha}^{(0)}) = \tau_0$, where $\boldsymbol{\alpha}$ is given by (2.4). From (2.11) and (2.12), in the limit $C \to \infty$, this corresponds to the intersection of two hyperplanes of dimension $J-1$, i.e., to a hyperplane of dimension $J-2$, and hence $\xi_J = 0$. For $C \gg 1$, from (2.11)–(2.13), with $\tau = \tau_0$, if we eliminate α_J to lowest order, we obtain

$$\sum_{i=1}^{J-1} a_i(\tau_0)\alpha_i = d_J \tau_0^{d_J} \log \tau_0 + 1 - \tau_0^{d_J} - \omega d_J \tau_0^{d_J} + O\left(\frac{1}{C}\right). \tag{9.1}$$

Also, from (8.8), we have

$$\left(\tau_0^{d_J} - 1\right)\sum_{j=1}^{J} d_j^2 \tau_0^{d_j}\alpha_j + \tau_0 \sum_{i=1}^{J-1} a'_i(\tau_0)\alpha_i = d_J^2 \tau_0^{d_J}\sum_{j=1}^{J}(\tau_0^{d_j} - 1)\alpha_j. \tag{9.2}$$

We now let $\alpha_i - \alpha_i^{(0)} = O(\epsilon)$, $0 < \epsilon \ll 1$, $i = 1, \ldots, J-1$. Then, from (2.4), $\alpha_J - \alpha_J^{(0)} = O(\epsilon)$. Hence, from (8.5), with $\tau = \tau_0$,

$$\sum_{j=1}^{J}(\tau_0^{d_j} - 1)(\alpha_j - \alpha_j^{(0)}) = \frac{1}{2C}\log\left(\frac{v}{v_0}\right) + O\left(\frac{\epsilon}{C^2}\right), \tag{9.3}$$

where, from (4.11),

$$v = \sum_{j=1}^{J} d_j^2 \tau_0^{d_j} \alpha_j \tag{9.4}$$

and v_0 is given by (2.18). It follows, from (3.17) and (9.2)–(9.4), that

$$v - v_0 = -\frac{\tau_0 \xi_{J-2}}{(\tau_0^{d_J} - 1)} + O\left(\frac{\epsilon}{C}\right). \tag{9.5}$$

From (2.14), (2.15), (3.17), (8.12), and (9.3), we obtain

$$\xi_J = \frac{1}{2C} \log\left(\frac{v}{v_0}\right) + \frac{(\tau_0^{d_J} - 1)}{C} \left[\left(A_0 + E_0 \sum_{j=1}^{J} d_j^3 \tau_0^{d_j} \alpha_j^{(0)}\right) \xi_{J-1} + D_0 \xi_{J-2}\right] + O\left(\frac{\epsilon}{C^2}\right). \tag{9.6}$$

However, from (9.5),

$$\log\left(\frac{v}{v_0}\right) = -\frac{\tau_0 \xi_{J-2}}{v_0(\tau_0^{d_J} - 1)} - \frac{\tau_0^2 \xi_{J-2}^2}{2v_0^2(\tau_0^{d_J} - 1)^2} + O\left(\frac{\epsilon}{C}\right). \tag{9.7}$$

Also, from (2.15) and (9.1), $\xi_{J-1} = O(\epsilon/C)$. Hence, from (8.11), (9.6), and (9.7),

$$\xi_J = -\frac{\tau_0^2 \xi_{J-2}^2}{4Cv_0^2(\tau_0^{d_J} - 1)^2} + O\left(\frac{\epsilon^2}{C^2}\right), \tag{9.8}$$

since, from (2.14) and (4.17), the error term contains the factor ϵ^2. This establishes the correctness of the negative coefficient in (3.19).

Appendix A. It was shown [15] that the blocking probabilities are given by

$$L_j = 1 - \frac{G(C - d_j)}{G(C)}, \tag{A.1}$$

where, for integer values of n,

$$G(C - n) = \frac{1}{2\pi i} \int_{|z|<1} k_n(z) e^{Cf(z)} dz, \tag{A.2}$$

with $f(z)$ as in (4.3),

$$k_n(z) = \frac{z^{n-1}}{(1 - z)}, \tag{A.3}$$

and the integral is taken in a counterclockwise direction around a circle of radius less than 1. Moreover, asymptotically, since $0 < z^* < 1$ in the case under consideration,

$$G(C - n) = \frac{e^{Cf(z^*)}}{2\sqrt{\pi C}} \cdot \left\{\beta_1 k_n(z^*) + \frac{1}{2C}\left[3\beta_3 k_n(z^*) + 3\beta_1\beta_2 k_n'(z^*) - \frac{1}{2}\beta_1^3 k_n''(z^*)\right] + O\left(\frac{1}{C^2}\right)\right\}, \tag{A.4}$$

where

$$\frac{\beta_1}{z^*} = \frac{\sqrt{2}}{\sqrt{v(z^*)}}, \qquad \frac{\beta_2}{z^*} = \frac{t(z^*)}{3[v(z^*)]^2}, \tag{A.5}$$

and z^*, $v(z^*)$, and $t(z^*)$ are given by (4.2), (4.4), and (5.3).

The definition of β_3 is not needed here since, from (A.3) and (A.4),

$$\frac{G(C-n)}{G(C)} = (z^*)^n \left\{1 + \frac{1}{2C}\left(3\beta_2\left[\frac{k_n'(z^*)}{k_n(z^*)} - \frac{k_0'(z^*)}{k_0(z^*)}\right] - \frac{1}{2}\beta_1^2\left[\frac{k_n''(z^*)}{k_n(z^*)} - \frac{k_0''(z^*)}{k_0(z^*)}\right]\right) + O\left(\frac{1}{C^2}\right)\right\}. \tag{A.6}$$

However, from (A.3),

$$\log k_n(z) = (n-1)\log z - \log(1-z), \tag{A.7}$$

and it follows that

$$\begin{aligned} \frac{k_n'(z)}{k_n(z)} &= \frac{1}{(1-z)} + \frac{(n-1)}{z}, \\ \frac{k_n''(z)}{k_n(z)} &= \frac{2}{(1-z)^2} + \frac{2(n-1)}{z(1-z)} + \frac{(n-1)(n-2)}{z^2}. \end{aligned} \tag{A.8}$$

Hence,

$$\frac{k_n'(z)}{k_n(z)} - \frac{k_0'(z)}{k_0(z)} = \frac{n}{z}, \qquad \frac{k_n''(z)}{k_n(z)} - \frac{k_0''(z)}{k_0(z)} = \frac{2n}{z(1-z)} + \frac{n(n-3)}{z^2}. \tag{A.9}$$

Then, from (A.5) and (A.6), we obtain

$$\frac{G(C-n)}{G(C)} = (z^*)^n \left\{1 - \frac{n}{2Cv(z^*)}\left[n - 5 + \frac{2}{(1-z^*)} - \frac{t(z^*)}{v(z^*)}\right] + O\left(\frac{1}{C^2}\right)\right\}. \tag{A.10}$$

The expression for L_j in (5.4) follows from (A.1).

Appendix B. We show here that $P_2(\theta) > 0$ for $\theta > 0$, where $P_2(\theta)$ is given by (6.18). From (6.11), we have

$$\theta\phi[\chi(\theta)] = \kappa, \tag{B.1}$$

where $\phi(y)$ is given by (6.10). We recall that $\phi'(y) > 0$ for $-\infty < y < \infty$, so that $\phi(y)$ has a unique inverse. Since $\phi(-\infty) = 0$ and $\phi(\infty) = \infty$, it follows that $-\infty < \chi < \infty$ for $\theta > 0$. From (B.1), we obtain

$$\theta\chi'(\theta) = -\frac{\phi(\chi)}{\phi'(\chi)}, \qquad \theta\chi'(\theta) + \theta^2\chi''(\theta) = \frac{\phi(\chi)}{\phi'(\chi)}\left\{1 - \frac{\phi(\chi)\phi''(\chi)}{[\phi'(\chi)]^2}\right\}. \tag{B.2}$$

Hence,

$$\chi(\theta) - \theta\chi'(\theta) - \theta^2\chi''(\theta) = \frac{K(\chi)}{[\phi'(\chi)]^3}, \tag{B.3}$$

where

$$K(\chi) = \chi[\phi'(\chi)]^3 - \phi(\chi)[\phi'(\chi)]^2 + [\phi(\chi)]^2\phi''(\chi). \tag{B.4}$$

From (6.18), since $\phi'(\chi) > 0$, it suffices to show that $K(\chi) > 0$ for $-\infty < \chi < \infty$.

Now,

$$K'(\chi) = 3\chi[\phi'(\chi)]^2\phi''(\chi) + [\phi(\chi)]^2\phi'''(\chi). \tag{B.5}$$

However, from (6.10),

$$\phi'(\chi) = 2\int_0^\infty ue^{2\chi u}e^{-u^2}du > 0, \tag{B.6}$$

$$\phi''(\chi) = 4\int_0^\infty u^2e^{2\chi u}e^{-u^2}du > 0, \tag{B.7}$$

and

$$\phi'''(\chi) = 8\int_0^\infty u^3e^{2\chi u}e^{-u^2}du > 0. \tag{B.8}$$

Hence, $K'(\chi) > 0$ for $0 \le \chi < \infty$. However (see [9]), from (6.10), $\phi(0) = \sqrt{\pi}/2$, $\phi'(0) = 1$, and $\phi''(0) = \sqrt{\pi}$, so that $K(0) > 0$. It follows that $K(\chi) > 0$ for $0 \le \chi < \infty$.

We now consider $-\infty < \chi < 0$. If we integrate by parts in (B.6), we obtain

$$\phi'(\chi) = 1 + 2\chi\phi(\chi). \tag{B.9}$$

Hence,

$$\phi''(\chi) = 2[\phi(\chi) + \chi\phi'(\chi)] = 2[\chi + (2\chi^2 + 1)\phi(\chi)]. \tag{B.10}$$

However, from (6.10),

$$\phi(\chi) = e^{\chi^2}I(\chi), \qquad I(\chi) = \int_{-\chi}^\infty e^{-\xi^2}d\xi. \tag{B.11}$$

It follows, from (B.4) and (B.9)–(B.11), that

$$H(\chi) \triangleq e^{-3\chi^2}K(\chi) \tag{B.12}$$

$$= (8\chi^4 + 2)I^3 + (12\chi^3 - 2\chi)e^{-\chi^2}I^2 + (6\chi^2 - 1)e^{-2\chi^2}I + \chi e^{-3\chi^2}.$$

Since $I'(\chi) = \exp(-\chi^2)$, we obtain

$$H'(\chi) = -4\chi^3 I\, M(\chi), \tag{B.13}$$

where

$$M(\chi) = -8I^2 - \left(\frac{10}{\chi} + \frac{1}{\chi^3}\right)e^{-\chi^2}I - \frac{3}{\chi^2}e^{-2\chi^2}. \tag{B.14}$$

Next, we find that

$$M'(\chi) = \left(4 + \frac{12}{\chi^2} + \frac{3}{\chi^4}\right)e^{-\chi^2}N(\chi), \tag{B.15}$$

where

$$N(\chi) = I + \frac{(5\chi + 2\chi^3)}{(3 + 12\chi^2 + 4\chi^4)} e^{-\chi^2}. \tag{B.16}$$

Hence,

$$e^{\chi^2} N'(\chi) = 1 + \frac{(5 - 4\chi^2 - 4\chi^4)}{(3 + 12\chi^2 + 4\chi^4)} - \frac{8\chi^2(3 + 2\chi^2)(5 + 2\chi^2)}{(3 + 12\chi^2 + 4\chi^4)^2}. \tag{B.17}$$

After some algebraic simplification, we obtain

$$N'(\chi) = \frac{24 e^{-\chi^2}}{(3 + 12\chi^2 + 4\chi^4)^2} > 0. \tag{B.18}$$

Now, from (B.11), $I(-\infty) = 0$ and hence, from (B.16), $N(-\infty) = 0$. It follows, from (B.18), that $N(\chi) > 0$ for $\chi > -\infty$. Hence, from (B.15), $M'(\chi) > 0$ for $-\infty < \chi < 0$. Since $M(-\infty) = 0$, from (B.14), it follows that $M(\chi) > 0$ for $-\infty < \chi < 0$, and hence, from (B.13), that $H'(\chi) > 0$ for $-\infty < \chi < 0$. However, from (B.11), $I(\chi)$ is exponentially small for $-\chi \gg 1$. It follows, from (B.12), that $H(\chi) > 0$, and hence $K(\chi) > 0$ for $-\infty < \chi < 0$. This completes the proof.

Acknowledgments. The authors are grateful to the referees for their helpful suggestions for improving the presentation.

REFERENCES

[1] E. Bouillet, D. Mitra, and K. G. Ramakrishnan, *The structure and management of service level agreements in networks*, IEEE J. Selected Areas Communications, 20 (2002), pp. 691–699.

[2] C. A. Courcoubetis, A. Dimakis, and M. I. Reiman, *Providing bandwidth guarantees over a best-effort network: Call admission and pricing*, in Proceedings of the Twentieth Annual Joint Conference of the IEEE Computer and Communication Societies, Vol. 1, IEEE Press, Piscataway, NJ, 2001, pp. 459–467.

[3] S. P. Evans, *Optimal bandwidth management and capacity provision in a broadband network using virtual paths*, Perform. Eval., 13 (1991), pp. 27–43.

[4] J. S. Kaufman, *Blocking in a shared resource environment*, IEEE Trans. Commun., 29 (1981), pp. 1474–1481.

[5] F. P. Kelly, *Reversibility and Stochastic Networks*, John Wiley, New York, 1980.

[6] K. Kumaran, M. Mandjes, D. Mitra, and I. Saniee, *Resource usage and charging in a multi-service multi-QoS packet network*, in Proceedings of the MIT/Tufts Workshop on Internet Service Quality Economics, 1999, to be published by MIT Press.

[7] S. S. Lam, *Queueing networks with population size constraints*, IBM J. Res. Develop., 21 (1977), pp. 370–378.

[8] S. Lanning, W. A. Massey, B. Rider, and Q. Wang, *Optimal pricing in queueing systems with quality of service constraints*, in Teletraffic Engineering in a Competitive World, Vol. 3B, Elsevier, New York, 1999, pp. 747–756.

[9] W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[10] D. Mitra and J. A. Morrison, *Erlang capacity and uniform approximations for shared unbuffered resources*, IEEE/ACM Trans. Networking, 2 (1994), pp. 558–570.

[11] M. Montgomery and G. de Veciana, *Hierarchical source routing through clouds*, in Proceedings of the Seventeenth Annual Joint Conference of the IEEE Computer and Communication Societies, Vol. 2, IEEE Press, Piscataway, NJ, 1998, pp. 685–692.

[12] J. A. Morrison, *Asymptotic shape of the Erlang capacity region of a critically loaded multi-service shared resource*, SIAM J. Appl. Math., 64 (2003), pp. 1–17.

[13] J. A. Morrison and D. Mitra, *Asymptotic shape of the Erlang capacity region of a multi-service shared resource*, Perform. Eval., 49 (2002), pp. 273–281.

[14] J. A. Morrison and K. G. Ramakrishnan, *Asymptotic solution to an inverse problem for a shared unbuffered resource*, SIAM J. Appl. Math., 63 (2002), pp. 222–240.

[15] J. A. Morrison, K. G. Ramakrishnan, and D. Mitra, *Refined asymptotic approximations to loss probabilities and their sensitivities in shared unbuffered resources*, SIAM J. Appl. Math., 59 (1998), pp. 494–513.

[16] M. I. Reiman, *A critically loaded multiclass Erlang loss system*, Queueing Systems Theory Appl., 9 (1991), pp. 65–81.

[17] S. Subramaniam, M. Azizoglu, and A. K. Somani, *All-optical networks with sparse wavelength conversion*, IEEE/ACM Trans. Networking, 4 (1996), pp. 544–557.

[18] T. Tripathi and K. N. Sivarajan, *Computing approximate blocking probabilities in wavelength routed all-optical networks with limited-range wavelength conversion*, in Proceedings of the Eighteenth Annual Joint Conference of the IEEE Computer and Communication Societies, Vol. 1, IEEE Press, Piscataway, NJ, 1999, pp. 329–336.

[19] Q. Wang, J. M. Peha, and M. A. Sirbu, *Optimal pricing for integrated services networks*, in Internet Economics, L. W. McKnight and J. P. Bailey, eds., MIT Press, Cambridge, MA, 1997, pp. 352–376.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 152–169

A MEMBRANE IN ADHESIVE CONTACT*

KEVIN T. ANDREWS†, L. CHAPMAN†, J. R. FERNÁNDEZ‡, M. FISACKERLY†, M. SHILLOR†, L. VANERIAN†, AND T. VANHOUTEN†

Abstract. A model for the process of quasi-static evolution of an elastic membrane in adhesive contact with a rigid obstacle is developed, analyzed, and numerically simulated. The model consists of an elliptic variational inequality for the membrane displacements and a nonlinear ordinary differential equation for the evolution of the adhesion field. By using regularity results from the theory of elliptic variational inequalities and a fixed point argument, the system is shown to have a unique weak solution. A fully discrete algorithm is described and shown to converge, and its error estimates are derived. In this process we make critical use of the regularity properties of the solution. Finally, the results of numerical simulations, based on the fully discrete algorithm, are presented.

Key words. contact, obstacle, membrane, free boundary, adhesion, existence and uniqueness, subdifferential, elliptic variational inequality, error estimates, numerical solutions

AMS subject classifications. 74M15, 35J85, 74K15, 74S05

DOI. 10.1137/S0036139902406206

1. Introduction.

This work deals with a new version of the classical contact problem between a stretched membrane and a rigid obstacle which lies beneath it. The novelty consists of allowing for adhesion between the membrane and the obstacle. Indeed, we assume that the obstacle, or a part of it, is covered with an adhesive which binds the membrane. As a result of the forces acting in the system, its state evolves in time; in particular, the bonds may break and consequently the bonding may deteriorate. The adhesion process is modeled by the introduction of the bonding field which measures the fractional density of the active bonds. Assuming a quasi-static process, the model consists of an elliptic variational inequality for the displacements coupled with a nonlinear ordinary differential equation for the bonding field. The problem is a free boundary problem, since the contact zone is unknown, and its determination is a part of the solution.

Adhesion processes are of considerable interest in industry because nonmetallic parts and components cannot be joined by welding and an adhesive needs to be utilized. The modeling and simulation of an adhesive contact problem applied to laminate materials, as well as further examples of industrial applications, can be found in [18].

Existence, uniqueness, and other results related to the classical contact problem for the membrane can be found in [6, 16, 19, 21] and references therein. Models for adhesive contact problems are very recent and can be found in [8, 5, 7, 9, 10, 12, 13, 15, 17, 18, 20] and references therein. The existence of a weak solution for the problem of dynamic contact of a membrane has been established in [1].

Following this introduction the rest of the paper is organized as follows. The model is described in section 2 and involves, in addition to the displacement field, the bonding field β, which measures the fractional density of the adhesive. In section 3 the

*Received by the editors April 23, 2002; accepted for publication (in revised form) May 1, 2003; published electronically November 19, 2003.
http://www.siam.org/journals/siap/64-1/40620.html
†Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309 (andrews@oakland.edu, shillor@oakland.edu).
‡Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, 15706 Santiago de Compostela, Spain (jramon@usc.es).

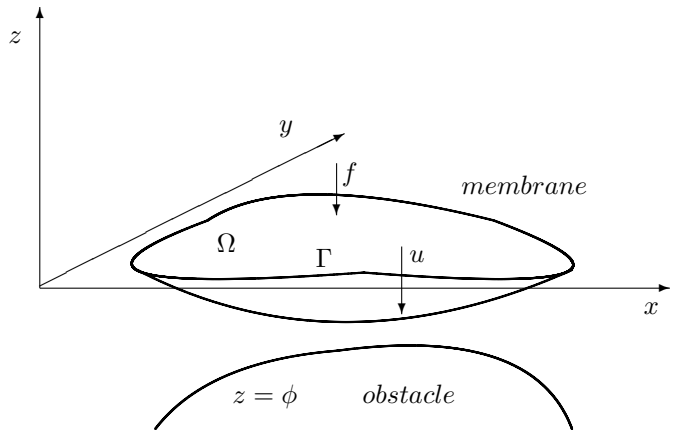


FIG. 1. *The membrane above the obstacle.*

model is set as an elliptic variational inequality for the displacements coupled with an ordinary differential equation for the bonding field β. Using regularity results from the theory of elliptic variational inequalities and a fixed point argument, we prove the existence of a unique weak solution for the problem in section 4. A fully discrete scheme for numerical simulations of the problem is developed in section 5, where it is shown to converge. This work uses the regularity properties of the solution. Also, an error estimate on the approximate numerical solutions is obtained. Then, in section 6, we present numerical simulations obtained by using a computer code based on the fully explicit scheme. We conclude the paper in section 7.

2. The model. We construct a model for the quasi-static process of adhesive contact between a membrane and a rigid obstacle which lies below it. The membrane is attached to a rigid rim, its displacements are restricted to lying on or above the rigid obstacle, and the membrane is in adhesive contact with the obstacle.

Let Ω denote the projection of the membrane on the xy plane, let $\Gamma = \partial\Omega$ be its boundary, let $z = \phi(x, y)$ represent the location of the rigid obstacle, and let $\Omega_T = \Omega \times (0, T)$. We assume that the membrane is being acted upon by an external vertical force f and that contact between the membrane and the obstacle involves adhesion. The setting is depicted in Figure 1.

We let $u = u(x, y, t)$ represent the vertical displacement of the membrane at location (x, y) and time t and take the upward direction as positive. We denote by $\xi(x, y, t)$ the reaction force of the obstacle, which is also positive when directed upward, and let $\eta(x, y, t)$ represent the tensile adhesive force, which will be described shortly.

The process is assumed to be quasi-static, so we neglect the inertial term in the equation of motion. Thus, the evolution of the state of the membrane is governed by

$$-\alpha\Delta u = f + \xi + \eta \quad \text{in } \Omega_T. \tag{2.1}$$

Here $\alpha = \hat{T}/\rho$, where $\hat{T}$ is the tension in the membrane and ρ is its surface density, both assumed to be positive constants.

Since the membrane is restricted to always lie above the obstacle ϕ, we have

$$u \geq \phi \quad \text{in } \Omega_T. \tag{2.2}$$

When contact between the membrane and the obstacle takes place, the obstacle's reaction force ξ is directed upward and exactly cancels the applied force, and so

$$u = \phi \quad \text{implies} \quad \xi \geq 0 \quad \text{in } \Omega_T. \tag{2.3}$$

When there is no contact the reaction force vanishes; thus $u > \phi$ implies $\xi = 0$. We may combine these three statements into the following linear complementarity condition:

$$\phi \leq u, \quad 0 \leq \xi, \quad \xi(u - \phi) = 0. \tag{2.4}$$

The last condition prevents both inequalities from being strict at the same time, since when contact takes place $\phi = u$ and in the absence of contact $\xi = 0$.

We now describe the adhesion process, which in this work is assumed to be irreversible, i.e., once a bond is severed there is no rebonding (see [9, 10, 18]). Models which allow for rebonding can be found in the recent articles [5, 4, 7, 15]. Let $\beta = \beta(x, y, t)$ denote the *bonding field*, which measures the fractional density of active bonds between the membrane and the obstacle. When $\beta = 1$ the bonding is complete at a point; when $\beta = 0$ there is no bonding, since all the bonds have been broken. Partial bonding at a point occurs when $0 < \beta < 1$. Thus, the bonding field has to satisfy

$$0 \leq \beta \leq 1 \quad \text{in } \Omega_T. \tag{2.5}$$

We assume that the adhesive is spread over the whole of the obstacle. The modifications needed when the adhesive is spread only on a part of the obstacle are straightforward, and we comment on them at the end of section 4.

The adhesive restoring force $\eta = \eta(x, y, t)$ is directed downward, as it acts to prevent the separation of the membrane from the obstacle. It is assumed to be jointly proportional to the distance from the obstacle and to β^2 (cf. [4]); thus,

$$\eta = -\kappa(u - \phi)\beta^2 \quad \text{in } \Omega_T, \tag{2.6}$$

where $\kappa > 0$ is the bonding coefficient or interface stiffness and $\kappa\beta^2$ is the system's "spring constant." We note that (2.2) implies that $\eta \leq 0$ and that when the membrane is in contact with the obstacle there is no adhesive restoring force, i.e., $\eta = 0$.

Following [5, 9, 10], we assume that the process is irreversible and the evolution of the adhesion field is given by

$$\beta' = -\gamma\kappa(u - \phi)^2\beta \quad \text{in } \Omega_T. \tag{2.7}$$

Here $1/\gamma > 0$ is the adhesion rate constant, and here and below we use a prime to denote a partial time derivative. To accompany this equation we must prescribe an initial condition $\beta(x, y, 0) = \beta_0(x, y)$, where β_0 is a given adhesive intensity distribution. We note that both $-\eta$ and $\kappa(u - \phi)^2\beta$ are related to partial derivatives of the free energy; see, e.g., [9, 10]. Indeed, if the surface free energy $\Psi(u, \beta)$ is defined as

$$\Psi(u, \beta) = w(1 - \beta) + \frac{1}{2}\kappa\beta^2(u - \phi)^2$$

[10, p. 154], where w is the Dupré adhesion energy, then (2.6) is obtained as $-\eta = \partial\Psi/\partial u$, where the minus sign reflects the downward direction of the adhesive force. Equation (2.7) follows from $-(1/\gamma)\beta' = \partial\Psi/\partial\beta$, where we omit the constant w for

the sake of simplicity. We turn next to describing the obstacle reaction force ξ. To that end we introduce the indicator function of the interval $(-\infty, 0]$,

$$\chi_{(-\infty,0]}(r) = \begin{cases} 0, & r \le 0, \\ \infty, & r > 0, \end{cases}$$

and its subdifferential,

$$\partial\chi_{(-\infty,0]}(r) = \begin{cases} 0, & r < 0, \\ [0,\infty), & r = 0, \\ \emptyset, & r > 0. \end{cases} \tag{2.8}$$

Using (2.8), we may rewrite the impenetrability condition (2.3) in the form

$$\xi \in \partial\chi_{(-\infty,0]}(\phi - u).$$

Then, using (2.6), equation (2.1) can be written as

$$-\alpha\Delta u - f + \kappa(u-\phi)\beta^2 \in \partial\chi_{(-\infty,0]}(\phi - u). \tag{2.9}$$

To complete the model we specify the displacements $u = g$ on the boundary Γ for $0 \le t \le T$.

Collecting the equations and conditions above, we can give the following formulation of the problem of quasi-static adhesive contact between a membrane and a rigid obstacle.

Problem P. Find a pair of functions $\{u, \beta\}$ such that

$$-\alpha\Delta u - f + \kappa(u-\phi)\beta^2 \in \partial\chi_{(-\infty,0]}(\phi - u) \quad \text{in } \Omega_T, \tag{2.10}$$

$$u \ge \phi \quad \text{in } \Omega_T, \tag{2.11}$$

$$u = g \quad \text{on } \Gamma \times (0,T), \tag{2.12}$$

$$\beta' = -\gamma\kappa(u-\phi)^2\beta \quad \text{in } \Omega_T, \tag{2.13}$$

$$\beta(0) = \beta_0 \quad \text{in } \Omega. \tag{2.14}$$

The classical obstacle problem for the membrane is obtained when $\beta \equiv 0$. Clearly, taking adhesion into account adds an interesting new twist to the problem.

We remark that it is possible to relate the two conditions (2.5) and (2.7) on β to the single inclusion

$$\frac{1}{\gamma}\beta' + \kappa(u-\phi)^2\beta + \partial\chi_{(-\infty,0]}(\beta') + \partial\chi_{[0,1]}(\beta) \ni 0.$$

Here the first subdifferential $\partial\chi_{(-\infty,0]}$ enforces the condition $\beta' \le 0$, and the second $\partial\chi_{[0,1]}$ enforces $0 \le \beta \le 1$. However, the way the problem is set, $\beta' \le 0$ follows anyway, and we do not need to enforce $0 \le \beta \le 1$ separately for the following reasons. When the initial condition for the bonding field satisfies $\beta_0 \le 1$, then it follows from (2.7) that $\beta' \le 0$ and so $\beta \le 1$ in Ω_T. If $\beta = 0$ at a point, then (2.7) implies $\beta' = 0$ there and thus $\beta = 0$ for all subsequent times. Consequently, if we begin with $0 \le \beta_0 \le 1$, then we have $0 \le \beta \le 1$ at all subsequent times. The fact that the adhesion process is irreversible is reflected in (2.7) and thus is responsible for obtaining $0 \le \beta \le 1$ in Ω_T. If we allow for the possibility of rebonding, then we no longer have this implication, and the condition $0 \le \beta \le 1$ has to be enforced separately, as has been done in [1].

The problem, as has been mentioned above, is a free boundary problem. Indeed, if

$$\Lambda(t) = \{(x, y) \in \Omega : u(x, y, t) = \phi(x, y)\}$$

is the contact set, then its boundary $\Gamma^* = \Gamma^*(t) = \partial\Lambda(t)$ is the *free boundary* which separates the contact set from the set where the membrane is above the obstacle. The evolution of this free boundary is an aspect of the problem that is of independent interest and will be addressed in the future.

The steady states of the problem can be obtained by assuming that the forces and the dependent variables are time independent. Let $\overline{f} = \overline{f}(x, y)$, $\overline{u} = \overline{u}(x, y)$, and $\overline{\beta} = \overline{\beta}(x, y)$ be the steady forces, displacements, and bonding field, respectively. It follows from (2.13) that either $\overline{u} = \phi$ or $\overline{\beta} = 0$. We conclude that $\{\overline{u}, \overline{\beta}\}$ satisfies the problem

$$\begin{aligned} -\alpha\Delta\overline{u} - \overline{f} &\in \partial\chi_{(-\infty,0]}(\phi - \overline{u}) \quad \text{in } \Omega, \\ \overline{u} &\geq \phi \quad \text{in } \Omega, \\ \overline{u} &= g \quad \text{on } \Gamma. \end{aligned}$$

This is the classical contact problem for the membrane and it has a unique weak solution $\overline{u}$; see, e.g., [16, 19]. On the other hand, if $\Lambda = \{\overline{u} = \phi\}$ is the contact set, then $\overline{\beta} = 0$ outside of Λ, and on Λ we have $\overline{\beta} = \beta^*$, where β^* is an arbitrary function in $L^\infty(\Lambda)$ such that $0 \leq \beta^* \leq 1$.

It may be of interest to identify those steady state functions $\overline{\beta}$ which are limits, as $t \to \infty$, of solutions of the evolution problem (2.10)–(2.14). This topic is currently open.

Finally, for technical reasons, it is convenient to formulate the problem so that the boundary condition on Γ is homogeneous. For this purpose we assume that $g(t)$ is the restriction to Γ of some function in $H^2(\Omega)$, which we also denote by $g(t)$. Let $w = u - g$, and denote $\tilde{f} = f + \Delta g$. Then the above problem may be formulated as follows.

Problem P. Find a pair of functions $\{w, \beta\}$ such that

$$-\alpha\Delta w - \tilde{f} + \kappa(w + g - \phi)\beta^2 \in \partial\chi_{(-\infty,0]}(\phi - g - w) \quad \text{in } \Omega_T, \tag{2.15}$$

$$w \geq \phi - g \quad \text{in } \Omega_T, \tag{2.16}$$

$$w = 0 \quad \text{on } \Gamma \times (0, T), \tag{2.17}$$

$$\beta' = -\gamma\kappa(w + g - \phi)^2\beta \quad \text{in } \Omega_T, \tag{2.18}$$

$$\beta(0) = \beta_0 \quad \text{in } \Omega. \tag{2.19}$$

In the next section we derive a variational or weak formulation of this problem.

3. Variational formulation. We proceed to obtain a variational formulation of the conditions (2.15)–(2.17). Let

$$V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma\},$$

and let the set of admissible functions be given by

$$K = \{v \in V : v \geq \phi - g(t) \quad \text{in } \Omega\}.$$

We assume that $\phi \leq g(t)$ on Ω, so that $0 \in K \neq \emptyset$. We note that K is a closed convex subset of V.

Next, let $t \in [0,T]$ be fixed, and for the sake of simplicity we write $w(t)$ instead of $w(x,y,t)$. We multiply both sides of (2.15) by $(v - w(t))$, where $v \in K$ is a test function, and integrate over Ω. Using the divergence theorem and the boundary condition (2.17), we have that the left-hand side can be written as

$$\alpha \int_\Omega \nabla w(t) \cdot \nabla (v - w(t))\,\mathrm{d}x + \kappa \int_\Omega (w(t) + g(t) - \phi)\beta^2(t)(v - w(t))\,\mathrm{d}x \\ - \int_\Omega \tilde{f}(v - w(t))\,\mathrm{d}x,$$

where $\tilde{f} = f + \Delta g$. Note that the boundary term vanishes since $v = w = 0$ on Γ. For the right-hand side we have, for any $\xi(t) \in \partial\chi_{(-\infty,0]}(\phi - g(t) - w)$, that

$$\int_\Omega \xi(t)(v - w(t))\,\mathrm{d}x = \int_\Omega \xi(t)(v + g(t) - \phi)\,\mathrm{d}x + \int_\Omega \xi(t)(\phi - g(t) - w(t))\,\mathrm{d}x.$$

The second integral on the right-hand side vanishes by (2.4). In the first integral on the right-hand side we have that $0 \le \xi$ and, since $v \in K$, $0 \le v + g(t) - \phi$; thus the first integral is nonnegative. We conclude that

$$\alpha \int_\Omega \nabla w(t) \cdot \nabla (v - w(t))\,\mathrm{d}x + \kappa \int_\Omega (w(t) + g(t) - \phi)\beta^2(t)(v - w(t))\,\mathrm{d}x \\ \ge \int_\Omega \tilde{f}(v - w(t))\,\mathrm{d}x.$$

This is a variational inequality for $w = w(t)$ for each $0 \le t \le T$.

Finally, we collect all the conditions and obtain the following variational formulation of problem (2.15)–(2.19).

Problem P_V. Find a pair of functions $\{w, \beta\}$ such that

$$w \in L^2(0,T;V), \quad w(t) \in K \ \text{ a.e. in } (0,T), \quad \beta \in W^{1,\infty}(0,T;L^\infty(\Omega)), \tag{3.1}$$

and, for a.e. $t \in (0,T)$ and each $v \in K$,

$$\alpha \int_\Omega \nabla w(t) \cdot \nabla (v - w(t))\,\mathrm{d}x + \kappa \int_\Omega (w(t) + g - \phi)\beta^2(t)(v - w(t))\,\mathrm{d}x \\ \ge \int_\Omega \tilde{f}(v - w(t))\,\mathrm{d}x \tag{3.2}$$

and

$$\beta'(t) = -\gamma\kappa(w(t) + g - \phi)^2\beta(t) \quad \text{a.e. in } \Omega_T, \tag{3.3}$$

$$\beta(0) = \beta_0 \quad \text{a.e. in } \Omega, \ \ t = 0. \tag{3.4}$$

We reformulate (3.2) in operator form by first defining, for each $\beta \in L^\infty(\Omega)$, the bilinear form $a : V \times V \to \mathbb{R}$ given by

$$a(\beta; v, w) = \int_\Omega (\alpha \nabla v \cdot \nabla w + \kappa\beta^2 vw)\,\mathrm{d}x, \qquad v, w \in V,$$

and the force function $F_\beta \in V'$ given by

$$(F_\beta, v) = \int_\Omega (\kappa(\phi - g)\beta^2 + \tilde{f})v\,\mathrm{d}x.$$

Then, Problem P_V can be written as follows: For almost every $t \in [0,T]$, find $w(t) \in K$ such that

$$a(\beta(t); w(t), v - w(t)) \geq (F_\beta(t), v - w(t)) \tag{3.5}$$

for all $v \in K$, together with (3.3) and initial conditions (3.4).

We use this formulation in the next section and in the numerical discretization of the problem.

4. Existence and uniqueness. In this section we establish the solvability of the variational problem (3.1)–(3.4). We assume that the domain Ω is of type $C^{1,1}$ [19, 21] so as to be able to obtain increased regularity of the solution. We make the following specific assumptions on the problem data:

$$g \in W^{1,\infty}(0,T; H^2(\Omega)), \tag{4.1}$$

$$f \in W^{1,\infty}(0,T; L^2(\Omega)), \tag{4.2}$$

$$\gamma = \text{const} > 0, \tag{4.3}$$

$$\phi \in H^2(\overline{\Omega}), \qquad \phi \leq g \text{ on } \overline{\Omega}, \tag{4.4}$$

$$\kappa = \text{const} > 0, \tag{4.5}$$

$$\beta_0 \in L^\infty(\Omega), \qquad 0 \leq \beta_0 \leq 1 \quad \text{a.e. on } \Omega. \tag{4.6}$$

We remark that it is possible to obtain solutions under weaker regularity conditions than those listed above; however, such solutions need not possess the regularity needed for the convergence analysis of section 5. Also, for the sake of simplicity we have chosen γ and κ to be positive constants, but our results hold more generally for the case when $\gamma \in L^\infty(\Omega)$, $\gamma > 0$ a.e. on Ω, $\kappa \in L^\infty(\Omega)$, and $\kappa \geq 0$ a.e. on Ω. We note that (4.4) implies that $0 \in K$.

We have the following existence and uniqueness result.

THEOREM 4.1. *Under the assumptions* (4.1)–(4.6), *there exists a unique weak solution* $\{w, \beta\}$ *of* (3.1)–(3.4) *such that*

$$w \in W^{1,\infty}(0,T; V) \cap L^\infty(0,T; H^2(\Omega)), \quad \beta \in W^{1,\infty}(0,T; L^\infty(\Omega)). \tag{4.7}$$

We note that the free energies associated with adhesion contact problems are nonconvex and consequently the usual convexity arguments that lead to uniqueness of the solution cannot be applied here. The uniqueness here follows from the fixed point argument used in the proof of the theorem.

The proof of the theorem uses a fixed point argument and is divided into several parts. Throughout this section, C will represent a positive generic constant which is independent of t and β but whose value may change from line to line.

We start by defining a convenient inner product on the space V. To this end we observe that there exists a $C > 0$ such that

$$C|v|_{H^1(\Omega)} \leq |\nabla v|_H \qquad \forall v \in V. \tag{4.8}$$

We consider now the inner product on V given by

$$(u, v)_V = (\nabla u, \nabla v)_H \tag{4.9}$$

and let $|\cdot|_V$ be the associated norm. Here and throughout the rest of the paper $H = L^2(\Omega)$. By using (4.8) we find that $|\cdot|_{H^1(\Omega)}$ and $|\cdot|_V$ are equivalent norms on V and, therefore, $(V, (\cdot,\cdot)_V)$ is a real Hilbert space.

We now establish an existence result for fixed β.

LEMMA 4.2. *Let* (4.1)–(4.5) *hold. Given any* $\beta \in C(0,T;H)$ *with* $\beta(x,t) \in [0,1]$, *a.e.* $x \in \Omega$, *for all* $t \in [0,T]$, *there exists a unique* $w = w_\beta \in C(0,T;V)$ *which satisfies* $w(t) \in K$ *and which solves* (3.5) *for all* $t \in [0,T]$. *Moreover,*

$$|w(t)|_{H^2(\Omega)} \le C, \tag{4.10}$$

where C *is independent of* β *and* t. *If, additionally, we have* $\beta \in W^{1,\infty}(0,T;L^\infty(\Omega))$, *then* $w \in W^{1,\infty}(0,T;V)$.

Proof. Using (4.5) and (4.9), we find that the bilinear symmetric form a satisfies

$$|a(u,v)| \le C|u|_V|v|_V \qquad \forall u,v \in V, \tag{4.11}$$

$$a(v,v) \ge C|v|_V^2 \qquad \forall v \in V, \tag{4.12}$$

and hence a is a continuous and coercive form on V.

Since K is a nonempty closed convex set of V, it follows from the projection theorem (see, e.g., [16, 19, 21]) that, for each $t \in [0,T]$, there exists a unique element $w(t) \in K$ which solves (3.5). Choosing $v = 0$ in (3.5) and using (4.12), we obtain that $|w(t)|_V \le C$. Now let $A_\beta : V \to V'$ be defined by $(A_\beta(w), v) = a(\beta; v, w)$. Since the hypotheses on the data imply that $F_\beta \in W^{1,\infty}(0,T;H)$, we have by Proposition 5.2.2 of [19] that

$$|A_\beta(w(t))|_H \le C(|F_\beta(t)|_H + |A_\beta(\phi)|_H) \le C. \tag{4.13}$$

Consequently, standard regularity results [19, 21] in the linear elliptic theory give that $w(t) \in H^2(\Omega)$. More specifically, by Theorem 2.24 of [21] and (4.13) we have that

$$|w(t)|_{H^2(\Omega)} \le C(|A_\beta(w(t))|_H + |w(t)|_V) \le C, \tag{4.14}$$

which gives (4.10). We next show that the function $t \to w(t)$ is continuous. To that end let t_1, $t_2 \in [0,T]$ and, for the sake of simplicity, we let $w(t_i) = w_i$, $\beta(t_i) = \beta_i$, $F_{\beta_i}(t_i) = F_i$ and let a_i denote the bilinear form $a(\beta_i; \cdot, \cdot)$. Using (3.5) twice and algebraic manipulations, we obtain

$$a_1(w_1 - w_2, w_1 - w_2) + \kappa((\beta_1^2 - \beta_2^2)w_2, w_1 - w_2)_H \le (F_1 - F_2, w_1 - w_2)_H. \tag{4.15}$$

Now, from (4.10), (4.12), (4.15), and Cauchy's inequality with ϵ, we find

$$|w_1 - w_2|_V^2 \le C(|F_1 - F_2|_H^2 + |\beta_1 - \beta_2|_H^2). \tag{4.16}$$

Since $\beta \in C(0,T;H)$, we obtain from (4.1), (4.2), (4.4), and (4.16) that $w \in C(0,T;V)$, which concludes the proof of the main statement of the lemma. To obtain the additional regularity result, we introduce, for any nonzero real number h, the difference quotient in t,

$$\delta_h w(t) = (w(t+h) - w(t))/h, \tag{4.17}$$

and then note that (4.16) implies that, for all t satisfying $|h| \le t \le T - |h|$,

$$|\delta_h w(t)|_V^2 \le C(|\delta_h F_\beta(t)|_H^2 + |\delta_h \beta(t)|_H^2). \tag{4.18}$$

Since $\beta' \in L^\infty(0,T;H)$ and $F'_\beta \in L^\infty(0,T;H)$, this implies that $w' \in L^\infty(0,T;V)$, which completes the proof of the lemma. ∎

Let w_β denote the solution in Lemma 4.2 corresponding to β, and consider the initial-value problem

$$\theta' + \gamma\kappa(w_\beta + g - \phi)^2\theta = 0 \quad \text{in} \quad \Omega_T, \tag{4.19}$$

$$\theta(0) = \beta_0 \quad \text{a.e. on } \Omega. \tag{4.20}$$

Under the assumptions on the data given in (4.1)–(4.6), there exists a unique function $\theta = \theta(\beta) \in W^{1,\infty}(0,T;L^\infty(\Omega))$ which solves (4.19)–(4.20). Let Z denote the closed convex subset of $C(0,T;H)$ which is defined by

$$Z = \{\ \beta \in C(0,T;H) :\ \beta(x,t) \in [0,1], \quad \text{a.e. } x \in \Omega, \quad \text{for all } \ t \in [0,T]\ \}. \tag{4.21}$$

We have the following containment result.

LEMMA 4.3. *If $\beta \in Z$, then $\theta(\beta) \in Z$.*

Proof. The result follows from (4.19) and the assumption that $\beta_0(x) \in [0,1]$ for almost every $x \in \Omega$. Indeed, (4.19) implies that for almost every $x \in \Omega$, the function $t \longmapsto \theta(\beta)(x,t)$ is decreasing and hence $\theta(\beta)(x,t) \le 1$ a.e. on Ω_T. Moreover, $\theta'(x,t)$ vanishes when $\theta(\beta)(x,t) = 0$, implying that $\theta(\beta)(x,t) = 0$ a.e. on Ω_T. This completes the proof of the lemma. ∎

To complete the proof of Theorem 4.1, we need only show that the map $\beta \to \theta^n(\beta)$ is a contraction on Z for some n. To that end, suppose that $\beta_i,\ i = 1,2$, are two functions in Z and let $t \in [0,T]$. We need to compare the functions $w_1 = w_{\beta_1}$ and $w_2 = w_{\beta_2}$.

Since $\beta_1,\ \beta_2 \in Z$, we find, by using arguments similar to those used in the proof of (4.16), that

$$|w_1(t) - w_2(t)|_V \le C|\beta_1(t) - \beta_2(t)|_H.$$

This implies, by the continuity of the embedding of V into H, that

$$|w_1(t) - w_2(t)|_H \le C|\beta_1(t) - \beta_2(t)|_H. \tag{4.22}$$

Now (4.19), (4.20), (4.10), and the continuity of the embedding of V into H yield

$$\begin{aligned}
&|\theta(\beta_1)(t) - \theta(\beta_2)(t)|_H \\
&\qquad \le \int_0^t |\gamma\kappa(w_1 + g - \phi)^2(s)\theta(\beta_1)(s) - \gamma\kappa(w_2 + g - \phi)^2(s)\theta(\beta_2)(s)|_H ds \\
&\qquad \le C\int_0^t |w_1(s) - w_2(s)|_H ds + C\int_0^t |\theta(\beta_1)(s) - \theta(\beta_2)(s)|_H ds.
\end{aligned}$$

Using a Gronwall-type inequality, we obtain

$$|\theta(\beta_1)(t) - \theta(\beta_2)(t)|_H \le C\int_0^t |w_1(s) - w_2(s)|_H ds. \tag{4.23}$$

Thus, (4.22) and (4.23) yield

$$|\theta(\beta_1)(t) - \theta(\beta_2)(t)|_H \le C\int_0^t |\beta_1(s) - \beta_2(s)|_H ds. \tag{4.24}$$

Iterating this inequality n times, we deduce

$$|\theta^n(\beta_1) - \theta^n(\beta_2)|_{C(0,T;H)} \le \frac{C^n T^n}{n!} |\beta_1 - \beta_2|_{C(0,T;H)}.$$

Therefore, θ^n is a contraction mapping on Z for any n sufficiently large, and hence θ has a unique fixed point in Z which is the unique solution of Theorem 4.1.

Remark. In the problem above, the adhesive is assumed to cover the whole of the obstacle so that the adhesion field β is defined over all of Ω. If the adhesive does not cover the entire obstacle, but only the portion corresponding to a subset $\Omega^* \subset\subset \Omega$, then we need to introduce the following minor modifications. Equations (2.13) and (2.14) hold in Ω_T^*. On the left-hand side of (2.10) we have to replace the term $\kappa(u-\phi)\beta^2$ with $\kappa(u-\phi)\tilde{\beta}^2$, where $\tilde{\beta}^2$ is equal to β^2 on Ω_T^* and is zero otherwise. In the variational formulation (3.2) the second integral on the left-hand side is over Ω^* instead of Ω, and (3.3) and (3.4) hold in Ω^*. If Ω^* is of class $C^{1,1}$, all the results above hold true for this problem too.

5. Numerical approximations. In this section we consider a fully discrete approximation of Problem P_V. For the sake of simplicity we assume that $g = 0$ and thus $\tilde{f} = f$ and $\phi \leq 0$; otherwise, we need to use $\tilde{f}$ and shift the dependent variable to $w = u - g$, as above. We assume that the membrane $\Omega \subset \mathbb{R}^2$ is a polygonal domain and we define a regular triangulation $\{\mathcal{T}^h\}_{h>0}$ composed of triangles on it, where h is the maximal diameter of these triangles. We define the element spaces

$$V^h = \{v^h \in C(\bar{\Omega});\, v^h_{|\tau} \in P_1(\tau), \quad \tau \in \mathcal{T}^h,\, v^h = 0 \quad \text{on} \quad \Gamma\},$$
$$B^h = \{\gamma^h \in L^\infty(\Omega)\,;\, \gamma^h_{|\tau} \in P_0(\tau), \quad \tau \in \mathcal{T}^h\},$$

where $P_1(\tau)$ denotes the polynomial space in two variables of degree at most one. We will use $V^h \subset V$ and $B^h \subset L^\infty(\Omega)$ to approximate the spaces V and $L^\infty(\Omega)$, respectively.

Next, we define the piecewise averaging operator $P^h : L^1(\Omega) \to B^h$ (see [13]) by

$$P^h u_{|\tau} = \frac{1}{\text{meas}(\tau)} \int_\tau u dx, \qquad \tau \in \mathcal{T}^h, \quad u \in L^1(\Omega).$$

We next select an approximation $U^h \subset V^h$ of the convex set U of admissible displacements, as follows. We denote by ϕ^h the Lagrange interpolation function of the obstacle shape function ϕ, and we assume that $\phi^h \geq \phi$ in Ω. Therefore, one choice of U^h is

$$U^h = \{v^h \in V^h;\, v^h \geq \phi^h \ \text{ in } \ \Omega\},$$

and, thus, the following conformity condition holds true:

$$U^h \subset U. \tag{5.1}$$

We suppose below that (5.1) is satisfied. Finally, to simplify the manipulations and some of the expressions, but without loss of generality, we assume that $\phi = 0$.

To simplify the presentation we redefine the form a as

$$a(v, w) = \alpha \int_\Omega \nabla v \cdot \nabla w \, \mathrm{d}x$$

and the functional j as

$$j(\beta; v, w) = \int_\Omega \kappa \beta^2 v w \, \mathrm{d}x.$$

To discretize the time derivatives, we consider a nonuniform partition of the time interval $[0,T]$, denoted by $0 = t_0 < t_1 < \cdots < t_N = T$. We define $k_n = t_n - t_{n-1}$, $n = 1, \ldots, N$, and let $k = \max_n k_n$ be the largest time step. For a continuous function $w(t)$ we let $w(t_n) = w_n$ and define $\rho = \kappa\gamma$.

A fully discrete approximation of Problem P_V is the following.

Problem PV^{hk}. Find the discrete displacement field $u^{hk} = \{u_n^{hk}\}_{n=0}^N \subset U^h$ and the discrete adhesion field $\beta^{hk} = \{\beta_n^{hk}\}_{n=0}^N \subset B^h$ such that

$$\beta_0^{hk} = \beta_0^h, \tag{5.2}$$

$$a(u_n^{hk}, v^h - u_n^{hk}) + j(\beta_n^{hk}; u_n^{hk}, v^h - u_n^{hk}) \geq (f_n, v^h - u_n^{hk})_H, \quad n = 0, 1, \ldots, N, \tag{5.3}$$

$$\beta_n^{hk} = \beta_{n-1}^{hk} - \rho k_n P^h[(u_{n-1}^{hk})^2](\beta_{n-1}^{hk})_+ \text{ in } \Omega, \quad n = 1, \ldots, N, \tag{5.4}$$

where β_0^h is an approximation of the initial condition β_0.

Here we have used $(\beta_{n-1}^{hk})_+ = \max\{0, \beta_{n-1}^{hk}\}$, which is the positive part of β_{n-1}^{hk}, to ensure that numerically $0 \leq \beta_{n-1}^{hk}$. The inequality $\beta_{n-1}^{hk} \leq 1$ is guaranteed since $\beta_0 \leq 1$, and the second term on the right-hand side is nonpositive.

By using classical results on variational inequalities (see, e. g., [11]), we find that Problem PV^{hk} has a unique solution $(u^{hk}, \beta^{hk}) \subset U^h \times B^h$.

In what follows we denote by c a generic constant that may vary from line to line but in any case depends only on the data and is independent of h or k. Our aim now is to obtain an error estimate for the differences $u_n - u_n^{hk}$ and $\beta_n - \beta_n^{hk}$. To that end we integrate (3.3) between 0 and t_n and obtain

$$\beta_n = \beta_0 - \rho \int_0^{t_n} (u(s))^2 \beta(s)_+ ds, \tag{5.5}$$

and we rewrite (5.4) in the form

$$\beta_n^{hk} = \beta_0^h - \rho \sum_{j=1}^n k_j P^h[(u_{j-1}^{hk})^2](\beta_{j-1}^{hk})_+. \tag{5.6}$$

Subtracting (5.6) from (5.5) and performing some algebraic manipulations, we get

(5.7)

$$|\beta_n - \beta_n^{hk}|_H \leq |\beta_0 - \beta_0^h|_H + c \left(I_n + \sum_{j=1}^n k_j |(I - P^h)[(u_{j-1})^2]|_H + \sum_{j=1}^n k_j |(u_{j-1})^2(\beta_{j-1})_+ - (u_{j-1}^{hk})^2(\beta_{j-1}^{hk})_+|_H \right),$$

where

$$I_n = \left| -\rho \int_0^{t_n} (u(s))^2 \beta_+(s) ds + \rho \sum_{j=1}^n k_j (u_{j-1})^2 (\beta_j)_+ \right|_H$$

is the integration error. Now, since $0 \leq \beta \leq 1$ we obtain

$$\begin{aligned} |(u_{j-1})^2(\beta_{j-1})_+ - (u_{j-1})^2(\beta_{j-1}^{hk})_+|_H & \\ &\leq c(|(u_{j-1})^2 - (u_{j-1}^{hk})^2|_H + |\beta_{j-1} - \beta_{j-1}^{hk}|_H) \\ &\leq c(|u_{j-1} - u_{j-1}^{hk}|_V + |\beta_{j-1} - \beta_{j-1}^{hk}|_H), \end{aligned}$$

and so we obtain

$$|\beta_n - \beta_n^{hk}|_H \le |\beta_0 - \beta_0^h|_H + c\left(I_n + \sum_{j=1}^{n} |(I-P^h)[(u_{j-1}^{hk})^2]|_H + \sum_{j=1}^{n} k_j e_{j-1}\right),$$

where $e_n = |u_n - u_n^{hk}|_V + |\beta_n - \beta_n^{hk}|_H$. Next, we put $v = v^h$ at time $t = t_n$ in (3.2) and subtract (5.3) to obtain

$$\begin{aligned} a(u_n - u_n^{hk}, u_n - u_n^{hk}) + j(\beta_n; u_n, u_n - u_n^{hk}) - j(\beta_n^{hk}; u_n^{hk}, u_n - u_n^{hk}) \\ \le a(u_n, u_n - v^h) + a(u_n - u_n^{hk}, u_n - v^h) + j(\beta_n; u_n, u_n - v^h) \\ - j(\beta_n^{hk}; u_n^{hk}, u_n - v^h) + (f_n, u_n - v^h)_H \quad \forall v^h \in V. \end{aligned}$$

Since

$$\begin{aligned} & j(\beta_n; u_n, u_n - u_n^{hk}) - j(\beta_n^{hk}; u_n^{hk}, u_n - u_n^{hk}) \\ = {} & j(\beta_n; u_n, u_n - u_n^{hk}) - j(\beta_n^{hk}; u_n, u_n - u_n^{hk}) + j(\beta_n^{hk}; u_n - u_n^{hk}, u_n - u_n^{hk}), \end{aligned}$$

after some calculations we obtain

$$|u_n - u_n^{hk}|_V^2 \le c\left(|u_n - v^h|_V + |u_n - u_n^{hk}|_V |u_n - v^h|_V + |\beta_n - \beta_n^{hk}|_H |u_n - u_n^{hk}|_V\right).$$

Using the Cauchy inequality $ab \le \epsilon a^2 + (1/4\epsilon)b^2$ for $a, b, \epsilon \in \mathbb{R},\ \epsilon > 0$, we find that

$$|u_n - u_n^{hk}|_V \le c\left(|u_n - v^h|_V^{1/2} + |u_n - v^h|_V + |\beta_n - \beta_n^{hk}|_H\right). \tag{5.8}$$

Taking into account (5.7) and (5.8), we obtain the error estimate

(5.9)

$$\begin{aligned} |u_n - u_n^{hk}|_V + |\beta_n - \beta_n^{hk}|_H \le c \Bigg(& |u_n - v^h|_V^{1/2} + |u_n - v^h|_V \\ & + I_n + \sum_{j=1}^{n} k_j |(I-P^h)[(u_{j-1})^2]|_H + \sum_{j=1}^{n} k_j e_{j-1} \Bigg). \end{aligned}$$

Proceeding as in [14], we obtain

$$I_n \le ck\left(|u'|_{L^\infty(0,T;H)} + |\beta'|_{L^\infty(0,T;H)}\right). \tag{5.10}$$

Applying a discrete version of the Gronwall inequality, we derive from (5.9) and (5.10) the following result.

THEOREM 5.1. *Let the assumptions of Theorem* 4.1 *hold. Then we have the error estimate*

(5.11)

$$\begin{aligned} \max_{1\le n\le N}\{|u_n - u_n^{hk}|_V + |\beta_n - \beta_n^{hk}|_H\} \le c \Bigg(& \inf_{v^h \in U^h} [|u_n - v^h|_V^{1/2} + |u_n - v^h|_V] \\ & + k\left(|u'|_{L^\infty(0,T;H)} + |\beta'|_{L^\infty(0,T;H)}\right) + \sum_{j=1}^{n} k_j |(I-P^h)[(u_{j-1})^2]|_H \Bigg). \end{aligned}$$

Estimate (5.11) is a basis for the convergence order analysis. Indeed, we have the following result.

COROLLARY 5.2. *Let the assumptions of Theorem* 4.1 *hold. If we assume, in addition, that the initial condition* β_0^h *is chosen so that* $|\beta_0 - \beta_0^h|_H \le ch$ *and that the operator* P^h *satisfies*

$$|(I - P^h)v|_H \le ch \quad \text{for } v \in H^2(\Omega),$$

then

(5.12)

$$\max_{1\le n\le N}\{|u_n - u_n^{hk}|_V + |\beta_n - \beta_n^{hk}|_H\} \le c\Big(h^{1/2}|u|_{L^\infty(0,T;H^2(\Omega))} + k(|u'|_{L^\infty(0,T;H)} + |\beta'|_{L^\infty(0,T;H)})\Big).$$

We remark that it is likely that the error estimate (5.12) holds for solutions of weaker regularity. However, in the present setting we have, as a consequence of (2.1) and the regularity of our solutions, that the reaction force satisfies

$$\xi \in L^\infty(0,T;H). \tag{5.13}$$

Consequently, it follows that

$$|a(u_n, v^h - u_n) + j(\beta_n; , u_n, v^h - u_n) - (f_n, v^h - u_n)_H| \le |\xi|_{L^\infty(0,T;H)}|u_n - v^h|_H.$$

Thus we obtain the following improved error estimate.

THEOREM 5.3. *Let the hypotheses of Corollary* 5.2 *hold. Then*

(5.14)

$$\max_{1\le n\le N}\{|u_n - u_n^{hk}|_V + |\beta_n - \beta_n^{hk}|_H\} \le c\Big(h(|\xi|_{L^\infty(0,T;H)} + |u|_{L^\infty(0,T;H^2(\Omega))}) + k(|u'|_{L^\infty(0,T;H)} + |\beta'|_{L^\infty(0,T;H)})\Big).$$

6. Numerical simulations. In order to verify the performance of the numerical algorithm described in the previous section and to gain insight into the behavior of the solutions, we performed several numerical experiments. In this section, we present three of the results obtained in these numerical simulations. Our main purpose is to show that the algorithm performed well. But also there is some interest in the results, which depict the evolution of the bonding field.

6.1. The numerical scheme. In the numerical simulations presented below, we use the fully discretized scheme PV^{hk} analyzed in the previous section. In all cases, we assume that V^h is composed of continuous and piecewise affine functions and B^h consists of piecewise constant functions. Therefore, Problem PV^{hk} is solved as follows.

Given $\beta_0^{hk} \in B^h$, *find* u_0^{hk} *satisfying*

$$a(u_0^{hk}, v^h - u_0^{hk}) + j(\beta_0^{hk}; u_0^{hk}, v^h - u_0^{hk}) \ge (f_0, v^h - u_0^{hk})_H \quad \forall v^h \in U^h. \tag{6.1}$$

Then, for $n = 1, \dots, N$, *find* (β_n^{hk}, u_n^{hk}) *satisfying*

$$\beta_n^{hk} = \beta_{n-1}^{hk} - \rho k_n P^h[(u_{n-1}^{hk})^2](\beta_{n-1}^{hk})_+ \text{ in } \Omega, \tag{6.2}$$

$$a(u_n^{hk}, v^h - u_n^{hk}) + j(\beta_n^{hk}; u_n^{hk}, v^h - u_n^{hk}) \ge (f_n, v^h - u_n^{hk})_H \quad \forall v^h \in U^h. \tag{6.3}$$

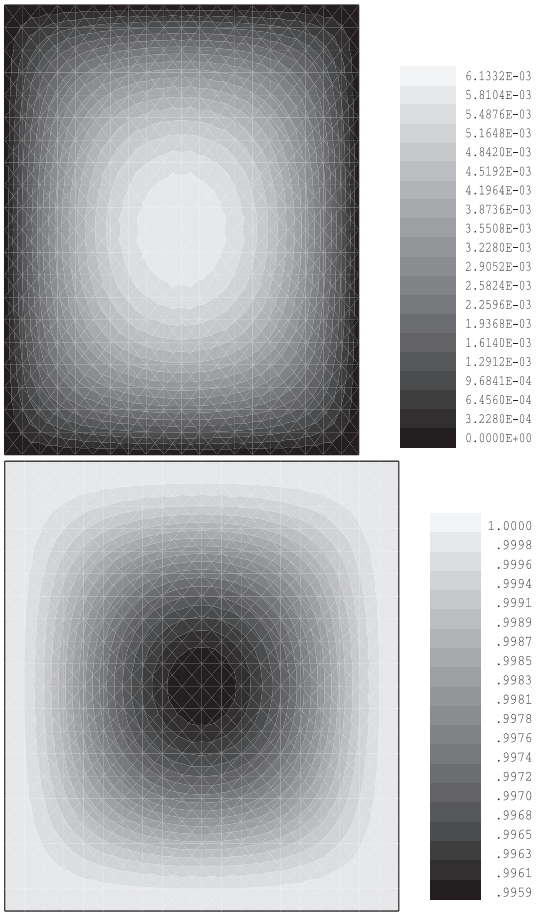


FIG. 2. *Test Problem* 1. *The displacement and adhesion fields at final time.*

The variational inequalities (6.1) and (6.3) have been solved by using a penalty-duality algorithm (see, e.g., [2, 3]). The numerical method was implemented on a IBM RISC6000 computer, and a typical run took about 15 seconds of CPU time.

6.2. Numerical solutions. We present now three test problems for the algorithm. In all examples, we choose $\Omega = (0,1) \times (0,1)$ as the membrane domain and we assume that the displacement field vanishes on $\Gamma = \partial\Omega$. Moreover, in the first two we assume that the obstacle shape function $\phi = 0$, while in the third example it is piecewise linear. We study the evolution of the displacement and adhesion fields during the time period of one second (i.e., $T = 1$ sec).

In the first example, we used the following data:

$$\kappa = 1\,N/m^3, \quad \gamma = 1\,m/N\,sec, \quad \alpha = 1\,N/m,$$
$$f(x,y,t) = 100\,N/m^2, \quad \phi = 0\,m,$$
$$\beta_0 = 1 \quad \text{in} \quad \Omega.$$

In Figure 2 we plot the displacement and the adhesion fields at the final time. The discretization parameter $k = 0.01$ was used.

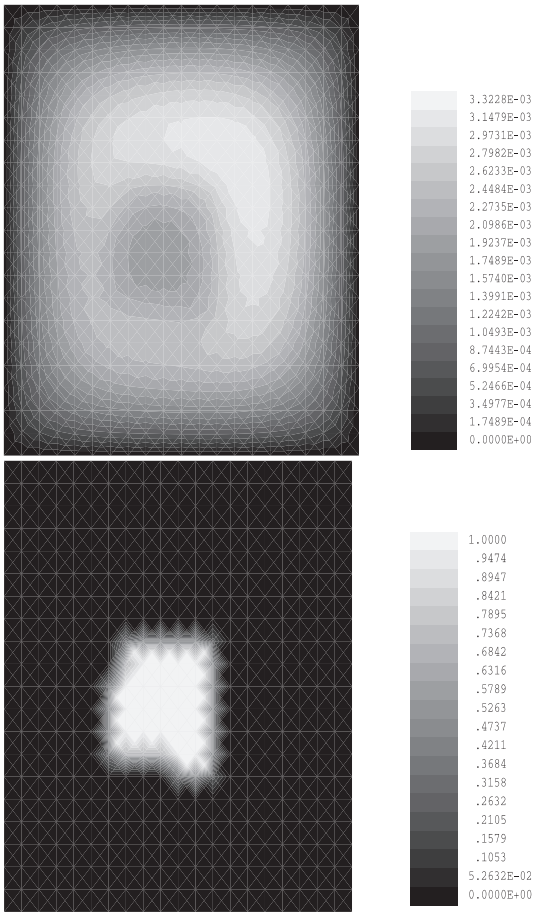


FIG. 3. *Test Problem* 2. *The displacement and adhesion fields at final time.*

In the second example, we used the following numerical data:

$$\kappa = 10^4\, N/m^3, \quad \gamma = 10^{-2}\, m/N\, sec, \quad \alpha = 1\, N/m,$$
$$f(x,y,t) = 100\, N/m^2, \quad \phi = 0\, m,$$
$$\beta_0(x,y,) = \begin{cases} 1 & \text{if} \quad 0.3 \le x \le 0.6,\ 0.3 \le y \le 0.6, \\ 0 & \text{elsewhere.} \end{cases}$$

We note that initially there is glue only on a part of the obstacle, and since the process is irreversible, the rest of the obstacle remains clear of glue.

Figure 3 depicts the displacement and adhesion fields at the final time for the value $k = 0.001$.

Finally, in the third example we used a nonflat obstacle of the form

$$\phi = \phi(x,y) = \begin{cases} 4x - 2 & \text{if } \ 0 \le x \le 0.25, \\ -1 & \text{if } \ 0.25 \le x \le 0.75, \\ 4x - 4 & \text{if } \ 0.75 \le x \le 1. \end{cases}$$

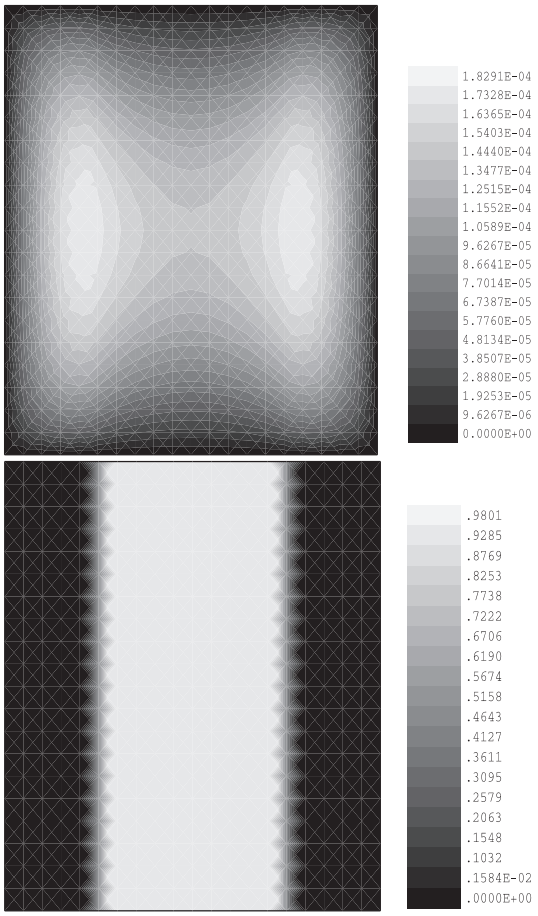


FIG. 4. *Test Problem* 3. *The displacement and adhesion fields for* $f = 10$.

We used the same values of the constants as in the two examples above, but with four different forces $f = 1, 10, 50, 100$ and initial adhesive

$$\beta_0(x, y) = \begin{cases} 1 & \text{if } 0.25 \le x \le 0.75, \\ 0 & \text{elsewhere.} \end{cases}$$

In Figure 4 we show the displacement and adhesion fields at $t = 1$ for $f = 10$.

In Figure 5 we show the displacement and adhesion fields at final time for the four different forces.

The algorithm was found to behave well, and the results are close to what one would expect. These simulations depict the behavior of the solutions. The settings and the conditions were chosen so that the evolution of the bonding field would be clear, and since the model is irreversible, the debonding is monotone.

The algorithm and the code could and will be used for thorough investigations, in conjunction with experimental results in real adhesion problems, for the purpose of parameter identification and model validation. In this way, they can be used to predict the behavior of real systems.

7. Conclusion. A model for the adhesive contact between a membrane and a rigid obstacle has been derived, its variational form obtained, and the existence of the

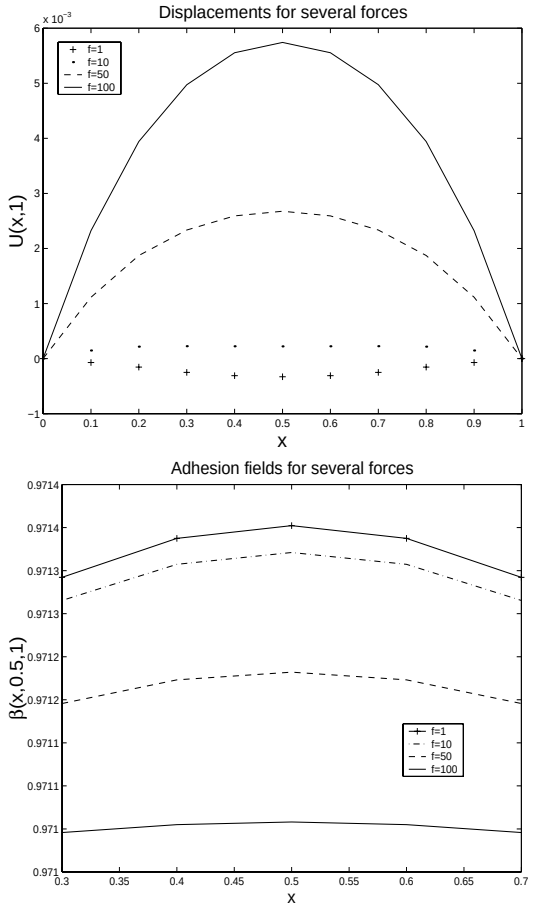


FIG. 5. *Test Problem* 3. *The displacement and adhesion fields for four forces.*

unique weak solution established. The model is in the form of an elliptic variational inequality for the membrane displacements coupled with an ordinary differential equation for the bonding field. The problem is a new and interesting version of the classical contact problem for the membrane and is a free boundary problem as well.

The existence and uniqueness of the weak solution have been proven by using regularity results in the theory of elliptic variational inequalities and a fixed point argument. A fully discrete algorithm for the problem has been derived and its convergence established. Error estimates were obtained together with the convergence order. The algorithm has been implemented and numerical simulations presented.

An open problem warranting further investigation is the evolution of the free boundary. It would be of interest to estimate and simulate the evolution of the contact set and the free boundary Γ^*. Another open question of interest deals with the asymptotic behavior of the solutions. In particular, it would be interesting to characterize the subset of functions in $L^\infty(\Omega)$ that are limits, as $t \to \infty$, of bonding fields β that are solutions to membrane problems.

Finally, the validation of the model by comparison to experimental results, including the recovery of parameters from experimental observations, are obvious applied next steps.

REFERENCES

[1] K. T. Andrews and M. Shillor, *Dynamic Adhesive Contact of a Membrane*, Adv. Math. Sci. Appl, 13 (2003), pp. 343–356.

[2] A. Bermúdez and C. Moreno, *Duality methods for solving variational inequalities*, Comput. Math. Appl., 7 (1981), pp. 43–58.

[3] M. Burguera and J. M. Viaño, *Numerical solving of frictionless contact problems in perfect plastic bodies*, Comput. Methods Appl. Mech. Engrg., 120 (1995), pp. 303–322.

[4] O. Chau, J. R. Fernández, M. Shillor, and M. Sofonea, *Variational and numerical analysis of a quasistatic viscoelastic contact problem with adhesion*, J. Comput. Appl. Math., to appear.

[5] O. Chau, M. Shillor, and M. Sofonea, *Dynamic frictionless contact with adhesion*, Z. Angew. Math. Phys., to appear.

[6] G. Duvaut and J. L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, 1976.

[7] J. R. Fernández, M. Shillor, and M. Sofonea, *Analysis and numerical simulations of a dynamic contact problem with adhesion*, Math. Comput. Modelling, 37 (2003), pp. 1317–1333.

[8] M. Frémond, *Adhérence des solides*, J. Méc. Théor. Appl., 6 (1987), pp. 383–407.

[9] M. Frémond, *Contact with adhesion*, in Topics in Nonsmooth Mechanics, J. J. Moreau, P. D. Panagiotopoulos, and G. Strang, eds., Birkhauser, Basel, 1988, pp. 157–186.

[10] M. Fremond, E. Sacco, N. Point, and J. M. Tien, *Contact with adhesion*, in ESDA Proceedings of the 1996 Engineering Systems Design and Analysis Conference, A. Lagarde and M. Raous, eds., ASME, New York, 1996, pp. 151–156.

[11] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.

[12] W. Han, K. L. Kuttler, M. Shillor, and M. Sofonea, *A beam in adhesive contact*, in Contact Mechanics, Proceedings of the Third Contact Mechanics International Symposium, J. A. C. Martins and M. D. P. Monteiro Marques, eds., Kluwer, Dordrecht, The Netherlands, 2002, pp. 277–284.

[13] W. Han, K. L. Kuttler, M. Shillor, and M. Sofonea, *Elastic beam in adhesive contact*, Internat. J. Solids Structures, 39 (2002), pp. 1145–1164.

[14] W. Han and M. Sofonea, *Numerical analysis of a frictionless contact problem for elastic-viscoplastic materials*, Comput. Methods Appl. Mech. Engrg., 190 (2000), pp. 179–191.

[15] L. Jianu, M. Shillor, and M. Sofonea, *A viscoelastic frictionless contact problem with adhesion*, Appl. Anal., 80 (2001), pp. 233–255.

[16] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Classics Appl. Math 31, SIAM, Philadelphia, 2000.

[17] N. Point, *Unilateral contact with adhesion*, Math. Methods Appl. Sci., 10 (1988), pp. 367–381.

[18] M. Raous, L. Cangémi, and M. Cocu, *A consistent model coupling adhesion, friction and unilateral contact*, Comput. Methods Appl. Mech. Engrg., 177 (1999), pp. 383–399.

[19] J.-F. Rodrigues, *Obstacle Problems in Mathematical Physics*, North–Holland, Amsterdam, 1987.

[20] J. Rojek and J. J. Telega, *Contact problems with friction, adhesion and wear in orthopaedic biomechanics*. I: *General developments*, J. Theoret. Appl. Mech., 39 (2001), pp. 655–677.

[21] G. Trioianello, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 170–195

HOMOGENIZATION OF THE MAXWELL EQUATIONS AT FIXED FREQUENCY*

NIKLAS WELLANDER† AND GERHARD KRISTENSSON‡

Abstract. The homogenization of the Maxwell equations at fixed frequency is addressed in this paper. The bulk (homogenized) electric and magnetic properties of a material with a periodic microstructure are found from the solution of a local problem on the unit cell by suitable averages. The material can be anisotropic and satisfies a coercivity condition. The exciting field is generated by an incident field from sources outside the material under investigation. A suitable sesquilinear form is defined for the interior problem, and the exterior Calderón operator is used to solve the exterior radiating fields. The concept of two-scale convergence is employed to solve the homogenization problem. A new a priori estimate is proved as well as a new result on the correctors.

Key words. Maxwell equations, homogenization, heterogeneous materials, periodic microstructure, effective properties, two-scale convergence, corrector results

AMS subject classifications. 35B27, 35Q60, 78A25, 78A40, 78A45, 78A48, 78M40

DOI. 10.1137/S0036139902403366

1. Introduction.

The concept of two-scale convergence is a well-established tool in the theory of homogenization of elliptic equations with rapidly oscillating coefficients; see, e.g., [2, 3, 7, 10, 12, 14, 17, 19, 21, 26, 27, 28]. The results apply to several types of partial differential equations that are used in the engineering sciences, such as heat conduction, elastic deformation, porous media, and acoustics. The situation is, however, different with the Maxwell equations, and the few results that exist adopt boundary conditions that are of less importance in applications. Specifically, the boundary conditions employed in the literature (see, e.g., [4, 6, 15, 18, 20, 26, 27, 28]) are those of perfectly conducting walls. This situation applies to the case of a resonator filled with a heterogeneous material, but for other situations these boundary conditions are less applicable. Moreover, there is a need for a better understanding of how a microscopic structure alters the macroscopic electric and magnetic behavior of the material if the sources of the electromagnetic fields are located outside the heterogeneous material. In fact, most applications in the engineering sciences use external excitations, and to find the homogenized parameters of a heterogeneous material, other boundary conditions, such as the penetrable boundary conditions, must be used.

The engineering literature is dominated by the simple mixture formulae, which are derived using physical arguments. For an excellent overview and history of the mixture formulae, see [22].

*Received by the editors October 28, 2002; accepted for publication May 13, 2003; published electronically November 19, 2003.

http://www.siam.org/journals/siap/64-1/40336.html

†Department of Mathematics, University of California, Santa Barbara, CA 93106. Current address: Swedish Defence Research Agency, FOI, P.O. Box 1165, SE-581 11 Linköping, Sweden (niklas.wellander@foi.se). The research of this author was supported by the Swedish Foundation for International Cooperation in Research and Higher Education (STINT), the Swedish Defence Research Agency (FOI), and by the Harald and Louise Ekman Foundation, whose support is gratefully acknowledged.

‡Department of Electroscience, Electromagnetic Theory, Lund Institute of Technology, P.O. Box 118, SE-221 00 Lund, Sweden (Gerhard.Kristensson@es.lth.se). The research of this author was supported by a grant from the Swedish Foundation for Strategic Research (SSF), whose support is gratefully acknowledged.

The object of this paper is to analyze thoroughly the homogenization of the Maxwell equations for a bounded object with penetrable boundary conditions. This homogenization problem seems not to have been published before in the literature. Moreover, the boundary condition implies that the excitation must be due to external sources. This situation is very important in many engineering applications, such as antenna applications. The two-scale convergence of the Maxwell equations depends on an a priori estimate of the fields. The external sources alter the traditional way of homogenization with two-scale convergence. In fact, in addition to the interior homogenization problem, there is an exterior scattering problem that couples via the boundary conditions to the interior problem. We solve this problem by introducing the Calderón operators, which map the tangential electric field to the tangential magnetic field on the bounding surface. In order to apply the boundary conditions and the Calderón operators, a new a priori estimate has been derived. The paper also includes new results on the correctors.

The paper is organized in the following way. Section 2 contains the prerequisites of the paper. The existence of solutions is proved in section 3, and the homogenization of the Maxwell equations is derived in section 4. We illustrate the exterior Calderón operator with two examples in section 5. The paper is concluded with a series of appendices that contain definitions of function spaces (Appendix A), and some important theorems (Appendix B). In Appendix C, the vector spherical waves used in section 5 are defined.

2. Formulation of the problem.

2.1. Domain and incident fields. Assume Ω is a bounded, open, simply connected set in $\mathbb{R}^3$ with $C^{1,1}$ boundary, $\partial\Omega$. The outward-pointing unit normal is $\hat{\boldsymbol{\nu}}$. The exterior of the volume Ω is denoted $\Omega_e = \mathbb{R}^3 \backslash \overline{\Omega}$, which is assumed vacuous. See Figure 2.1 for a typical geometry.

The incident fields, $\boldsymbol{E}_i$ and $\boldsymbol{H}_i$, are assumed to have their sources outside Ω in a bounded region Ω_i, i.e., $\Omega \cap \Omega_i = \emptyset$. It is assumed to be a fixed field throughout this paper. Outside this region the fields satisfy the time-harmonic Maxwell equations in vacuum time convention $e^{-i\omega t}$, i.e., they satisfy[1]

$$\begin{cases} \nabla \times \boldsymbol{E}_i(\boldsymbol{x}) = ik_0 \boldsymbol{H}_i(\boldsymbol{x}), \\ \nabla \times \boldsymbol{H}_i(\boldsymbol{x}) = -ik_0 \boldsymbol{E}_i(\boldsymbol{x}), \end{cases} \qquad \boldsymbol{x} \in \mathbb{R}^3.$$

The wave number in a vacuum is $k_0 = \omega/c_0$, where ω is the angular frequency of the fields, and c_0 is the speed of light in a vacuum. The incident fields $\boldsymbol{E}_i$ and $\boldsymbol{H}_i$ are assumed to have traces on $\partial\Omega$ belonging to $H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega)$, i.e., $(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_i, \hat{\boldsymbol{\nu}} \times \boldsymbol{H}_i) \in H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega) \times H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega)$; see Appendix A for definitions of the function spaces. Otherwise, the incident fields are arbitrary.

2.2. Interior problem. In Ω we assume there is a material modeled by the permittivity dyadic $\epsilon(\boldsymbol{x})$ and the permeability dyadic $\mu(\boldsymbol{x})$. The permittivity dyadic

[1] We use scaled electric and magnetic fields in this paper; i.e., the SI-unit fields $\boldsymbol{E}_{\mathrm{SI}}$ and $\boldsymbol{H}_{\mathrm{SI}}$ are related to the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ used in this paper by

$$\boldsymbol{E}_{\mathrm{SI}}(\boldsymbol{x}) = \frac{\boldsymbol{E}(\boldsymbol{x})}{\sqrt{\epsilon_0}}, \qquad \boldsymbol{H}_{\mathrm{SI}}(\boldsymbol{x}) = \frac{\boldsymbol{H}(\boldsymbol{x})}{\sqrt{\mu_0}},$$

where the permittivity and permeability of vacuum are denoted ϵ_0 and μ_0, respectively.

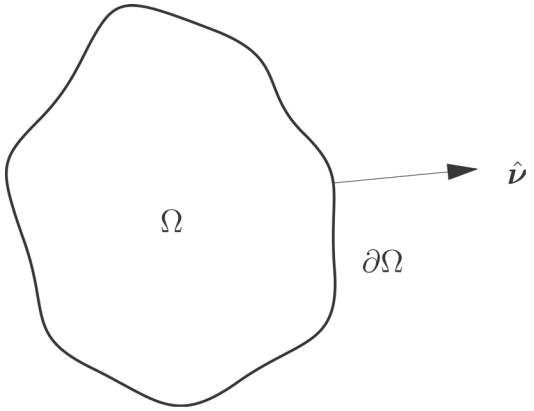


FIG. 2.1. *Typical geometry of the scattering problem in this paper.*

is assumed to satisfy

$$-ik_0\boldsymbol{\xi}\cdot\left(\epsilon(\boldsymbol{x})-\epsilon(\boldsymbol{x})^\dagger\right)\cdot\boldsymbol{\xi}^*\geq C_1|\boldsymbol{\xi}|^2 \qquad \text{for all } \boldsymbol{\xi}\in\mathbb{C}^3 \text{ and a.e. } \boldsymbol{x}\in\Omega \tag{2.1}$$

and

$$|\epsilon(\boldsymbol{x})\cdot\boldsymbol{\xi}|\leq C_2|\boldsymbol{\xi}| \qquad \text{for all } \boldsymbol{\xi}\in\mathbb{C}^3 \text{ and a.e. } \boldsymbol{x}\in\Omega, \tag{2.2}$$

where † denotes the Hermitian of the dyadic ϵ and $C_i>0$, $i=1,2$. The condition in (2.1) corresponds physically to a passive material, i.e., a material that show dissipation. The entries of $\epsilon(\boldsymbol{x})$ are assumed to belong to $L^\infty(\Omega)$, which implies (2.2). Similar assumptions hold for the permeability μ. We note that it follows that ϵ and μ are invertible and that the inverses have the same kind of properties [9, p. 22].

In Ω the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{H}$ satisfy the Maxwell equations

$$\begin{cases}\nabla\times\boldsymbol{E}(\boldsymbol{x})=ik_0\mu(\boldsymbol{x})\cdot\boldsymbol{H}(\boldsymbol{x}),\\ \nabla\times\boldsymbol{H}(\boldsymbol{x})=-ik_0\epsilon(\boldsymbol{x})\cdot\boldsymbol{E}(\boldsymbol{x}),\end{cases} \qquad \boldsymbol{x}\in\Omega. \tag{2.3}$$

We are looking for solutions $\boldsymbol{E}$ and $\boldsymbol{H}$ of these equations in the space $H(\mathrm{rot},\Omega)$. A weak formulation of the solution to this problem is found in section 3.2.1.

2.3. Exterior problem. The presence of the material in the domain Ω distorts the incident fields $\boldsymbol{E}_i$ and $\boldsymbol{H}_i$. This distortion is denoted by the scattered fields, $\boldsymbol{E}_s$ and $\boldsymbol{H}_s$. They belong to $H_{\mathrm{loc}}(\mathrm{rot},\overline{\Omega}_e)$ and satisfy

$$\begin{cases}\nabla\times\boldsymbol{E}_s(\boldsymbol{x})=ik_0\boldsymbol{H}_s(\boldsymbol{x}),\\ \nabla\times\boldsymbol{H}_s(\boldsymbol{x})=-ik_0\boldsymbol{E}_s(\boldsymbol{x}),\end{cases} \qquad \boldsymbol{x}\in\Omega_e. \tag{2.4}$$

Moreover, the scattered fields satisfy the Silver–Müller radiation condition at infinity, i.e., one of the following conditions (see [11]):

$$\begin{cases}\hat{\boldsymbol{x}}\times\boldsymbol{E}_s(\boldsymbol{x})-\boldsymbol{H}_s(\boldsymbol{x})=o(1/x),\\ \hat{\boldsymbol{x}}\times\boldsymbol{H}_s(\boldsymbol{x})+\boldsymbol{E}_s(\boldsymbol{x})=o(1/x)\end{cases} \qquad \text{as } x\to\infty \tag{2.5}$$

uniformly in all directions $\hat{\boldsymbol{x}}$.

In Ω_e the sum of the incident and the scattered fields is defined as the total field, i.e.,

$$\begin{cases} \boldsymbol{E}_t(\boldsymbol{x}) = \boldsymbol{E}_i(\boldsymbol{x}) + \boldsymbol{E}_s(\boldsymbol{x}), \\ \boldsymbol{H}_t(\boldsymbol{x}) = \boldsymbol{H}_i(\boldsymbol{x}) + \boldsymbol{H}_s(\boldsymbol{x}), \end{cases} \qquad \boldsymbol{x} \in \Omega_e.$$

The boundary conditions on $\partial\Omega$ are

$$\begin{cases} \hat{\boldsymbol{\nu}} \times \boldsymbol{E}_i|_{\partial\Omega} + \hat{\boldsymbol{\nu}} \times \boldsymbol{E}_s|_{\partial\Omega} = \hat{\boldsymbol{\nu}} \times \boldsymbol{E}|_{\partial\Omega}\,, \\ \hat{\boldsymbol{\nu}} \times \boldsymbol{H}_i|_{\partial\Omega} + \hat{\boldsymbol{\nu}} \times \boldsymbol{H}_s|_{\partial\Omega} = \hat{\boldsymbol{\nu}} \times \boldsymbol{H}|_{\partial\Omega}\,, \end{cases} \tag{2.6}$$

where the traces of the fields are taken from the outside (inside) in the left-hand (right-hand) side of the equations and belong to $H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega)$.

2.4. Calderón operators. The Calderón operator C^e utilizes the solution of a specific exterior problem. In fact, the following exterior problem, based upon (2.4) and (2.5) and given $\boldsymbol{m} \in H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega)$, is fundamental:

$$\begin{cases} (1) \quad (\boldsymbol{E}_s, \boldsymbol{H}_s) \in H_{\mathrm{loc}}(\mathrm{rot}, \overline{\Omega}_e) \times H_{\mathrm{loc}}(\mathrm{rot}, \overline{\Omega}_e), \\ (2) \quad \begin{cases} \nabla \times \boldsymbol{E}_s(\boldsymbol{x}) = ik_0 \boldsymbol{H}_s(\boldsymbol{x}), \\ \nabla \times \boldsymbol{H}_s(\boldsymbol{x}) = -ik_0 \boldsymbol{E}_s(\boldsymbol{x}), \end{cases} \quad \boldsymbol{x} \in \Omega_e, \\ (3) \quad \begin{cases} \hat{\boldsymbol{x}} \times \boldsymbol{E}_s(\boldsymbol{x}) - \boldsymbol{H}_s(\boldsymbol{x}) = o(1/x) \\ \text{or} \\ \hat{\boldsymbol{x}} \times \boldsymbol{H}_s(\boldsymbol{x}) + \boldsymbol{E}_s(\boldsymbol{x}) = o(1/x) \end{cases} \quad \text{as } x \to \infty, \\ (4) \quad \hat{\boldsymbol{\nu}} \times \boldsymbol{E}_s|_{\partial\Omega} = \boldsymbol{m} \in H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega). \end{cases} \tag{2.7}$$

(Problem (R))

This problem has a unique solution [4, 9]; see also section 3.1.

We have the following results proved in [9, p. 35].

THEOREM 2.1. *With the boundary $\partial\Omega$ of regularity $C^{1,1}$, the mapping*

$$\boldsymbol{\gamma}_\tau : \boldsymbol{u} \to \hat{\boldsymbol{\nu}} \times \boldsymbol{u}|_{\partial\Omega}$$

is a continuous mapping from $H_{loc}(\mathrm{rot}, \overline{\Omega}_e)$ onto $H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega)$.

The trace theorem is a local property of the field at the boundary, and the theorem shows that the field loses regularity on the boundary. We note that a similar result holds when the trace is taken from the inside of the boundary; see section 3.2.

The linear mapping of the electric field to the corresponding magnetic field on the boundary for a solution of the exterior problem is called the exterior Calderón operator. The following makes this definition precise.

DEFINITION 2.2. *The exterior Calderón operator C^e is defined as*

$$C^e : \boldsymbol{m} \to \hat{\boldsymbol{\nu}} \times \boldsymbol{H}_s|_{\partial\Omega}\,, \qquad H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega) \to H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega),$$

where $\boldsymbol{m} = \hat{\boldsymbol{\nu}} \times \boldsymbol{E}_s|_{\partial\Omega}$ and the fields $\boldsymbol{E}_s$ and $\boldsymbol{H}_s$ satisfy Problem (R) *in* (2.7).

Notice that the exterior Calderón operator C^e is uniquely defined for all $\boldsymbol{m} \in H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega)$, since Problem (R) has a unique solution. Two explicit examples of the exterior Calderón operator are given in section 5.

THEOREM 2.3. *The exterior Calderón operator defined in Definition* 2.2 *has the following properties:*

1. *The exterior Calderón operator satisfies the positivity condition*

$$\Re \iint_{\partial\Omega} C^e(\boldsymbol{m}) \cdot (\hat{\boldsymbol{\nu}} \times \boldsymbol{m}^*)\, dS \geq 0 \qquad \text{for all } \boldsymbol{m} \in H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega). \tag{2.8}$$

2. *The exterior Calderón operator satisfies*

$$(C^e)^2 = -\mathbf{I} \text{ on } H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega),$$

which implies that C^e is bounded on $H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega)$.

3. *The exterior Calderón operator is independent of the material properties inside Ω.*

Here dS denotes the surface measure of $\partial\Omega$.

Proof of Theorem 2.3. Property 1 is a simple consequence of the radiation condition and proved in, e.g., [9]. Specifically, the radiation conditions, (2.5), imply

$$\Re \iint_{\partial\Omega} \hat{\boldsymbol{\nu}} \cdot (\boldsymbol{E}_s \times \boldsymbol{H}_s^*)\, dS = \Re \iint_{|\boldsymbol{x}|=R} \hat{\boldsymbol{x}} \cdot (\boldsymbol{E}_s \times \boldsymbol{H}_s^*)\, dS = \iint_{|\boldsymbol{x}|=R} |\boldsymbol{E}_s|^2\, dS + o(1)$$

as $R \to \infty$, which implies (2.8), since $\hat{\boldsymbol{\nu}} \cdot (\boldsymbol{E}_s^* \times \boldsymbol{H}_s) = -C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_s) \cdot \boldsymbol{E}_s^*$.

Moreover, to prove property 2 we utilize the symmetry $\{\boldsymbol{E}_s, \boldsymbol{H}_s\} \to \{\boldsymbol{H}_s, -\boldsymbol{E}_s\}$ in (2.4) and the uniqueness of the exterior problem.

Property 3 is a consequence of the uniqueness of the exterior problem. ∎

An immediate consequence of the positivity property of C^e is that

$$-\Re \iint_{\partial\Omega} C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_s) \cdot \boldsymbol{E}_s^*\, dS \geq 0 \qquad \text{for all } \hat{\boldsymbol{\nu}} \times \boldsymbol{E}_s \in H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega). \tag{2.9}$$

3. Existence of solutions. The existence of exterior and interior solutions is addressed in this section.

3.1. Exterior problem. The system (2.4) with the radiation condition (2.5) supplied with the boundary condition

$$\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_s|_{\partial\Omega} = \boldsymbol{m} \quad \in H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega),$$

i.e., Problem (R) in (2.7), has a unique solution in $(\boldsymbol{E}_s, \boldsymbol{H}_s) \in H_{\mathrm{loc}}(\mathrm{rot}, \overline{\Omega}_e) \times H_{\mathrm{loc}}(\mathrm{rot}, \overline{\Omega}_e)$ for any $\boldsymbol{m} \in H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega)$ [9, p. 107].

3.2. Interior problem. We have the interior trace result, similar to Theorem 2.1.

THEOREM 3.1. *With the boundary $\partial\Omega$ of regularity $C^{1,1}$, the mapping*

$$\boldsymbol{\gamma}_\tau : \boldsymbol{u} \to \hat{\boldsymbol{\nu}} \times \boldsymbol{u}|_{\partial\Omega}$$

is a continuous mapping from $H(\mathrm{rot}, \Omega)$ onto $H^{-\frac{1}{2}}(\mathrm{div}, \partial\Omega)$.

3.2.1. Sesquilinear form and weak solutions. Using Theorem 3.1, we can now define the sesquilinear form (see [9])

$$\begin{aligned} a(\boldsymbol{u}, \boldsymbol{v}) = &- \iiint_\Omega \left\{ \frac{1}{ik_0} (\nabla \times \boldsymbol{v}^*) \cdot \mu^{-1} \cdot (\nabla \times \boldsymbol{u}) + ik_0 \boldsymbol{v}^* \cdot \epsilon \cdot \boldsymbol{u} \right\} dv \\ &- \iint_{\partial\Omega} C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{u}) \cdot \boldsymbol{v}^*\, dS \end{aligned}$$

for $\boldsymbol{u}$ and $\boldsymbol{v}$ in $H(\mathrm{rot}, \Omega)$. We denote the volume measure in $\mathbb{R}^3$ by dv in this paper.

A weak formulation of the original problem is then to find $\boldsymbol{E} \in H(\mathrm{rot}, \Omega)$ such that

$$a(\boldsymbol{E}, \boldsymbol{v}) = \iint_{\partial\Omega} (\hat{\boldsymbol{\nu}} \times \boldsymbol{H}_i - C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_i)) \cdot \boldsymbol{v}^* \, dS \quad \text{for all } \boldsymbol{v} \in H(\mathrm{rot}, \Omega). \tag{3.1}$$

This solution satisfies the boundary conditions, (2.6), and couples to the exterior solution in (2.4)–(2.5). The corresponding magnetic field $\boldsymbol{H}$ is then constructed as[2]

$$\begin{cases} \boldsymbol{H}(\boldsymbol{x}) = -\frac{i}{k_0}\mu^{-1}(\boldsymbol{x}) \cdot (\nabla \times \boldsymbol{E}(\boldsymbol{x})), \\ \nabla \times \boldsymbol{H}(\boldsymbol{x}) = -ik_0\epsilon(\boldsymbol{x}) \cdot \boldsymbol{E}(\boldsymbol{x}), \end{cases} \qquad \boldsymbol{x} \in \Omega.$$

To see this, let $\boldsymbol{E}$ be a sufficiently regular solution to the Maxwell equations, (2.3). Then (3.1) is equivalent to the Maxwell equations with a coupling to an exterior solution since

$$\begin{aligned} a(\boldsymbol{E}, \boldsymbol{v}) &= -\iiint_{\Omega} \{(\nabla \times \boldsymbol{v}^*) \cdot \boldsymbol{H} - \boldsymbol{v}^* \cdot (\nabla \times \boldsymbol{H})\} \, dv - \iint_{\partial\Omega} C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}) \cdot \boldsymbol{v}^* \, dS \\ &= \iint_{\partial\Omega} \{(\hat{\boldsymbol{\nu}} \times \boldsymbol{H}) \cdot \boldsymbol{v}^* - C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}) \cdot \boldsymbol{v}^*\} \, dS, \end{aligned}$$

which is identical to (3.1) by the use of the boundary conditions on $\partial\Omega$ and by the definition

$$\iint_{\partial\Omega} C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_s) \cdot \boldsymbol{v}^* \, dS = \iint_{\partial\Omega} (\hat{\boldsymbol{\nu}} \times \boldsymbol{H}_s) \cdot \boldsymbol{v}^* \, dS.$$

Moreover, the sesquilinear form a is coercive, i.e.,

$$\begin{aligned} \Re a(\boldsymbol{u}, \boldsymbol{u}) = &- \iiint_{\Omega} \frac{1}{ik_0} (\nabla \times \boldsymbol{u}^*) \cdot \left(\mu^{-1} - \mu^{-1^\dagger}\right) \cdot (\nabla \times \boldsymbol{u}) \, dv \\ &- \iiint_{\Omega} ik_0 \boldsymbol{u}^* \cdot \left(\epsilon - \epsilon^\dagger\right) \cdot \boldsymbol{u} \, dv \\ &- \Re \iint_{\partial\Omega} C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{u}) \cdot \boldsymbol{u}^* \, dS \geq C\|\boldsymbol{u}\|^2_{H(\mathrm{rot},\Omega)}, \end{aligned} \tag{3.2}$$

since from (2.1) we get[3]

$$\begin{cases} -ik_0\boldsymbol{\xi} \cdot \left(\epsilon(\boldsymbol{x}) - \epsilon(\boldsymbol{x})^\dagger\right) \cdot \boldsymbol{\xi}^* \geq C_1|\boldsymbol{\xi}|^2, \\ \frac{i}{k_0}\boldsymbol{\xi} \cdot \left(\mu^{-1}(\boldsymbol{x}) - \mu^{-1^\dagger}(\boldsymbol{x})\right) \cdot \boldsymbol{\xi}^* \geq C_2|\boldsymbol{\xi}|^2 \end{cases} \qquad \text{for all } \boldsymbol{\xi} \in \mathbb{C}^3 \text{ and a.e. } \boldsymbol{x} \in \Omega,$$

and we have also used (2.9).

---

[2]This construction is consistent since $-ik_0\epsilon(\boldsymbol{x}) \cdot \boldsymbol{E}(\boldsymbol{x})$ is the weak curl of $\boldsymbol{H}(\boldsymbol{x}) = -\frac{i}{k_0}\mu^{-1}(\boldsymbol{x}) \cdot (\nabla \times \boldsymbol{E}(\boldsymbol{x}))$. In fact, we have

$$(\boldsymbol{H}, \nabla \times \boldsymbol{\phi}) + ik_0 \, (\epsilon \cdot \boldsymbol{E}, \boldsymbol{\phi}) = 0 \qquad \text{for all } \boldsymbol{\phi} \in D(\Omega; \mathbb{C}^3)$$

since $a(\boldsymbol{E}, \boldsymbol{\phi}) = 0$ for all $\boldsymbol{\phi} \in D(\Omega; \mathbb{C}^3)$.

[3]With (2.1) we get

$$\frac{i}{k_0} \, (\mu^t \cdot \boldsymbol{\zeta}) \cdot \left(\mu^{-1} - \mu^{-1^\dagger}\right) \cdot (\mu^t \cdot \boldsymbol{\zeta})^* = \frac{i}{k_0}\boldsymbol{\zeta} \cdot \left(\mu^\dagger - \mu\right) \cdot \boldsymbol{\zeta}^* \geq \frac{C_2}{k_0^2}|\boldsymbol{\zeta}|^2.$$

Applying this result with $\boldsymbol{\zeta} = \mu^{t^{-1}} \cdot \boldsymbol{\xi}$, we get

$$\frac{i}{k_0}\boldsymbol{\xi} \cdot \left(\mu^{-1} - \mu^{-1^\dagger}\right) \cdot \boldsymbol{\xi}^* \geq \frac{C_2}{k_0^2}|\mu^{t^{-1}} \cdot \boldsymbol{\xi}|^2 \geq C|\boldsymbol{\xi}|^2$$

since μ is invertible.

3.2.2. Existence of a unique solution. Equation (3.1) has a unique solution in $H(\mathrm{rot},\Omega)$ due to the Lax–Milgram theorem (see Theorem A.1), since the sesquilinear form $a(\boldsymbol{u},\boldsymbol{v})$ is continuous, bounded, and coercive, and the right-hand side of (3.1) is continuous on $H(\mathrm{rot},\Omega)$. In fact,

$$\begin{aligned}
&\left| \iint_{\partial\Omega} (\hat{\boldsymbol{\nu}} \times \boldsymbol{H}_i - C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_i)) \cdot \boldsymbol{v}^* \, dS \right| \\
&\quad \leq \left(\|\hat{\boldsymbol{\nu}} \times \boldsymbol{H}_i\|_{H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)} + \|C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_i)\|_{H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)} \right) \|\boldsymbol{v}\|_{H^{-\frac{1}{2}}(\mathrm{rot},\partial\Omega)} \\
&\quad \leq C' \left(\|\hat{\boldsymbol{\nu}} \times \boldsymbol{H}_i\|_{H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)} + \|(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_i)\|_{H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)} \right) \|\boldsymbol{v}\|_{H(\mathrm{rot},\Omega)}
\end{aligned}$$

by Minkowski's inequality, duality [9, p. 38], and the continuous dependence of the trace norm on the norm of the corresponding function space.

4. Homogenization. So far we have considered a general heterogeneous scattering problem with a unique solution in $H(\mathrm{rot},\Omega)$ for a given incident electromagnetic field. But if the heterogeneous material in Ω has a typical spatial scale which is much smaller than the size of the domain, then one runs into severe numerical problems if one tries to apply some standard numerical code, e.g., a finite element method (FEM). The principal obstacle is that the fine scale requires a very fine numerical mesh which generates a far too large linear system of equations for any computer to solve. However, if the wavelength of the incident field is much larger than the fine scale, then the field cannot resolve the fine scale and the solution of the Maxwell equations can be approximated by the solution of a scattering problem with constant coefficients; i.e., the heterogeneous material in Ω has been replaced by a homogeneous material with the same effective material properties. The procedure for finding these effective properties of the heterogeneous material is called homogenization.

4.1. Heterogeneous problem. Let us begin with the definition of a Y-cell which is the open unit cube in $\mathbb{R}^3$, i.e., $Y =]0,1[^3$. Further, from now on we assume that ϵ and μ are Y-periodic, which is defined as $\epsilon(\boldsymbol{x}+\hat{\boldsymbol{e}}_k) = \epsilon(\boldsymbol{x})$ for every $k = 1,2,3$, where $\hat{\boldsymbol{e}}_k$, $k = 1,2,3$, is the canonical basis in $\mathbb{R}^3$.

In the following, we assume that the material in the domain Ω is periodic with period ε in the three Cartesian coordinate directions, i.e., it is the union of a collection of disjoint, open identical cubes[4] with side length ε (Y^ε-cells); see Figure 4.1. It is easily verified that the scaled permeability and permittivity, $\epsilon(\boldsymbol{x}/\varepsilon)$ and $\mu(\boldsymbol{x}/\varepsilon)$, are periodic with period ε.

In Ω the fields satisfy the source-free Maxwell equations[5]

$$\begin{cases}
\nabla \times \boldsymbol{E}^\varepsilon(\boldsymbol{x}) = ik_0 \boldsymbol{B}^\varepsilon(\boldsymbol{x}), \\
\nabla \times \boldsymbol{H}^\varepsilon(\boldsymbol{x}) = -ik_0 \boldsymbol{D}^\varepsilon(\boldsymbol{x}), \\
\nabla \cdot \boldsymbol{B}^\varepsilon(\boldsymbol{x}) = 0, \\
\nabla \cdot \boldsymbol{D}^\varepsilon(\boldsymbol{x}) = 0,
\end{cases} \qquad \boldsymbol{x} \in \Omega,$$

[4]More generally, $Y = (0,a_1) \times (0,a_2) \times (0,a_3)$, where $a_i > 0$, $i = 1,2,3$, and $\epsilon(\boldsymbol{x} + a_k \hat{\boldsymbol{e}}_k) = \epsilon(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{R}^3$ and for every $k = 1,2,3$. A similar result holds for the permeability μ.

[5]The electric and magnetic fields are scaled as above (see footnote 1), and the SI-unit flux densities $\boldsymbol{D}_{\mathrm{SI}}$ and $\boldsymbol{B}_{\mathrm{SI}}$ are related to the fields $\boldsymbol{D}$ and $\boldsymbol{B}$ used in this paper by

$$\boldsymbol{D}_{\mathrm{SI}}(\boldsymbol{x}) = \sqrt{\epsilon_0}\boldsymbol{D}(\boldsymbol{x}), \qquad \boldsymbol{B}_{\mathrm{SI}}(\boldsymbol{x}) = \sqrt{\mu_0}\boldsymbol{B}(\boldsymbol{x}).$$

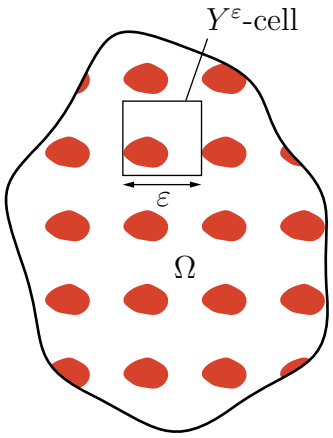


FIG. 4.1. *Typical periodic geometry of the material parameters.*

almost everywhere, with boundary conditions given by (2.6). By using the constitutive relations for the periodic material,

$$\begin{cases} \boldsymbol{D}^\varepsilon(\boldsymbol{x}) = \epsilon(\boldsymbol{x}/\varepsilon)\cdot\boldsymbol{E}^\varepsilon(\boldsymbol{x}), \\ \boldsymbol{B}^\varepsilon(\boldsymbol{x}) = \mu(\boldsymbol{x}/\varepsilon)\cdot\boldsymbol{H}^\varepsilon(\boldsymbol{x}), \end{cases} \qquad \boldsymbol{x}\in\Omega,$$

we eliminate $\boldsymbol{D}^\varepsilon$, $\boldsymbol{B}^\varepsilon$ and obtain

$$\begin{cases} \nabla\times\boldsymbol{E}^\varepsilon(\boldsymbol{x}) = ik_0\mu(\boldsymbol{x}/\varepsilon)\cdot\boldsymbol{H}^\varepsilon(\boldsymbol{x}), \\ \nabla\times\boldsymbol{H}^\varepsilon(\boldsymbol{x}) = -ik_0\epsilon(\boldsymbol{x}/\varepsilon)\cdot\boldsymbol{E}^\varepsilon(\boldsymbol{x}), \\ \nabla\cdot\{\epsilon(\boldsymbol{x}/\varepsilon)\cdot\boldsymbol{E}^\varepsilon(\boldsymbol{x})\} = 0, \\ \nabla\cdot\{\mu(\boldsymbol{x}/\varepsilon)\cdot\boldsymbol{H}^\varepsilon(\boldsymbol{x})\} = 0, \end{cases} \qquad \boldsymbol{x}\in\Omega, \tag{4.1}$$

where the solution $(\boldsymbol{E}^\varepsilon, \boldsymbol{H}^\varepsilon)$ is in $H(\mathrm{rot},\Omega)\times H(\mathrm{rot},\Omega)$ and belongs to a family of solutions, one for each ε. In the homogenization procedure we identify the limit of the fields $\boldsymbol{E}^\varepsilon, \boldsymbol{H}^\varepsilon$ when $\varepsilon\to 0$. This limit satisfies the homogenized system with constant coefficients, which is a model of a homogeneous material.

4.1.1. A priori estimate. We note that the heterogeneous system in (4.1) is of the same form as (2.3) and that the constitutive relations satisfy the same assumptions as in section 2.2. A weak formulation of the two first equations in (4.1) supplied with boundary conditions (2.6) reads

$$a^\varepsilon(\boldsymbol{E}^\varepsilon,\boldsymbol{v}) = \iint_{\partial\Omega} (\hat{\boldsymbol{\nu}}\times\boldsymbol{H}_i - C^e(\hat{\boldsymbol{\nu}}\times\boldsymbol{E}_i))\cdot\boldsymbol{v}^*\,dS \quad \text{for all } \boldsymbol{v}\in H(\mathrm{rot},\Omega), \tag{4.2}$$

where

$$\begin{aligned} a^\varepsilon(\boldsymbol{u},\boldsymbol{v}) = -\iiint_\Omega \Big\{ &\frac{1}{ik_0}(\nabla\times\boldsymbol{v}^*)\cdot\mu^{-1}(\boldsymbol{x}/\varepsilon)\cdot(\nabla\times\boldsymbol{u}) \\ &+ ik_0\boldsymbol{v}^*\cdot\epsilon(\boldsymbol{x}/\varepsilon)\cdot\boldsymbol{u}\Big\}\,dv - \iint_{\partial\Omega} C^e(\hat{\boldsymbol{\nu}}\times\boldsymbol{u})\cdot\boldsymbol{v}^*\,dS. \end{aligned} \tag{4.3}$$

We have the following a priori estimate.

THEOREM 4.1. *Let $\boldsymbol{E}^\varepsilon, \boldsymbol{H}^\varepsilon$ be a solution of (4.2); then*

$$\|\boldsymbol{E}^\varepsilon\|_{H(\mathrm{rot},\Omega)} + \|\boldsymbol{H}^\varepsilon\|_{H(\mathrm{rot},\Omega)} \leq C,$$

where the constant C depends only on the domain Ω, the material parameters in Ω, and the strength of the incident field.

Proof of Theorem 4.1. The sesquilinear form $a^\varepsilon(\boldsymbol{u},\boldsymbol{v})$ is coercive (cf. (3.2)), and the weak formulation (4.2) gives

$$\begin{aligned}
C\|\boldsymbol{E}^\varepsilon\|^2_{H(\mathrm{rot},\Omega)} &\leq \Re a^\varepsilon(\boldsymbol{E}^\varepsilon,\boldsymbol{E}^\varepsilon) \leq |a^\varepsilon(\boldsymbol{E}^\varepsilon,\boldsymbol{E}^\varepsilon)| \\
&= \left| \iint_{\partial\Omega} (\hat{\boldsymbol{\nu}} \times \boldsymbol{H}_i - C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_i)) \cdot (\boldsymbol{E}^\varepsilon)^* \, dS \right| \\
&\leq \left(\|\hat{\boldsymbol{\nu}} \times \boldsymbol{H}_i\|_{H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)} + \|C^e(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_i)\|_{H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)} \right) \|\boldsymbol{E}^\varepsilon\|_{H^{-\frac{1}{2}}(\mathrm{rot},\partial\Omega)} \\
&\leq C' \left(\|\hat{\boldsymbol{\nu}} \times \boldsymbol{H}_i\|_{H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)} + \|(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_i)\|_{H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)} \right) \|\boldsymbol{E}^\varepsilon\|_{H(\mathrm{rot},\Omega)}
\end{aligned}$$

by Minkowski's inequality, duality [9, p. 38], and the continuous dependence of the trace norm on the norm of the corresponding function space. It follows now that

$$\|\boldsymbol{E}^\varepsilon\|_{H(\mathrm{rot},\Omega)} \leq C' \left(\|\hat{\boldsymbol{\nu}} \times \boldsymbol{H}_i\|_{H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)} + \|(\hat{\boldsymbol{\nu}} \times \boldsymbol{E}_i)\|_{H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)} \right) \leq C$$

by the assumption of the incident field. The bound of $\boldsymbol{E}^\varepsilon$ can now be used in (4.1) to get the estimate of $\boldsymbol{H}^\varepsilon$. ∎

4.2. Homogenized problem.

THEOREM 4.2. *The sequence of solutions $(\boldsymbol{E}^\varepsilon, \boldsymbol{H}^\varepsilon)$ of (4.1) converges weakly in $H(\mathrm{rot},\Omega) \times H(\mathrm{rot},\Omega)$ to $(\boldsymbol{E},\boldsymbol{H}) \in H(\mathrm{rot},\Omega) \times H(\mathrm{rot},\Omega)$, the unique solution of the homogenized Maxwell equations*

$$\begin{cases}
\nabla \times \boldsymbol{E}(\boldsymbol{x}) = ik_0\mu^h \cdot \boldsymbol{H}(\boldsymbol{x}), \\
\nabla \times \boldsymbol{H}(\boldsymbol{x}) = -ik_0\epsilon^h \cdot \boldsymbol{E}(\boldsymbol{x}), \\
\nabla \cdot \boldsymbol{B}(\boldsymbol{x}) = 0, \\
\nabla \cdot \boldsymbol{D}(\boldsymbol{x}) = 0,
\end{cases} \tag{4.4}$$

which is coupled to the exterior problem (2.4)–(2.5) via the boundary conditions (2.6). The homogenized permeability and permittivity ϵ^h and μ^h are defined by

$$\begin{cases}
\epsilon^h = \iiint_Y \epsilon(\boldsymbol{y}) \cdot (\mathbf{I}_3 - \nabla_{\boldsymbol{y}} \boldsymbol{\chi}_e(\boldsymbol{y})) \, dv_{\boldsymbol{y}}, \\
\mu^h = \iiint_Y \mu(\boldsymbol{y}) \cdot (\mathbf{I}_3 - \nabla_{\boldsymbol{y}} \boldsymbol{\chi}_h(\boldsymbol{y})) \, dv_{\boldsymbol{y}},
\end{cases}$$

$$\boldsymbol{\chi}_e(\boldsymbol{y}) = \sum_{i=1}^3 \chi_e^i(\boldsymbol{y})\hat{\boldsymbol{e}}_i, \qquad \boldsymbol{\chi}_h(\boldsymbol{y}) = \sum_{i=1}^3 \chi_h^i(\boldsymbol{y})\hat{\boldsymbol{e}}_i, \tag{4.5}$$

where $\chi_e^i(\boldsymbol{y})$ and $\chi_h^i(\boldsymbol{y})$, $i = 1,2,3$, in $H^1_\#(Y)/\mathbb{C}$ solve the local elliptic problems

$$\begin{cases}
\iiint_Y \nabla_{\boldsymbol{y}} w(\boldsymbol{y}) \cdot \epsilon(\boldsymbol{y}) \cdot \left(\hat{\boldsymbol{e}}_i - \nabla_{\boldsymbol{y}} \chi_e^i(\boldsymbol{y}) \right) dv_{\boldsymbol{y}} = 0, \\
\iiint_Y \nabla_{\boldsymbol{y}} w(\boldsymbol{y}) \cdot \mu(\boldsymbol{y}) \cdot \left(\hat{\boldsymbol{e}}_i - \nabla_{\boldsymbol{y}} \chi_h^i(\boldsymbol{y}) \right) dv_{\boldsymbol{y}} = 0
\end{cases} \tag{4.6}$$

for all $w \in H^1_\#(Y)$.

We note that the weak convergence is sharp in the sense that it never converges strongly in $H(\mathrm{rot},\Omega)$ except in the electrostatic case (see the note after Theorem B.3). However, we can get strong convergence by the use of corrector functions; see section 4.2.2. These functions contain the fine-scale information in the problem and yield strong convergence when scaled and added to the homogenized solution.

Proof of Theorem 4.2. We use the concept of two-scale convergence; see Appendix B. Due to the a priori estimates there exists a subsequence which converges in the two-scale sense. We will keep the index ε for this subsequence. In the end we conclude that the whole original sequence converges due to the fact that the homogenized system has a unique solution. Let $\boldsymbol{\phi}(\boldsymbol{x}) = \varepsilon w(\boldsymbol{x}/\varepsilon)\boldsymbol{v}(\boldsymbol{x})$, where $w \in H^1_{\#}(Y)$ and $\boldsymbol{v} \in C_0^\infty(\Omega;\mathbb{C}^3)$. Then $\boldsymbol{\phi} \in H(\mathrm{rot},\Omega)$ and is an admissible test function. We get in (4.1)

$$\begin{cases} \displaystyle\iiint_\Omega \boldsymbol{E}^\varepsilon(\boldsymbol{x})\cdot\{\varepsilon w(\boldsymbol{x}/\varepsilon)\nabla_{\boldsymbol{x}}\times\boldsymbol{v}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}} w(\boldsymbol{x}/\varepsilon)\times\boldsymbol{v}(\boldsymbol{x})\}\,dv \\ \qquad\qquad - ik_0\displaystyle\iiint_\Omega \varepsilon w(\boldsymbol{x}/\varepsilon)\boldsymbol{v}(\boldsymbol{x})\cdot\{\mu(\boldsymbol{x}/\varepsilon)\cdot\boldsymbol{H}^\varepsilon(\boldsymbol{x})\}\,dv = 0, \\ \displaystyle\iiint_\Omega \boldsymbol{H}^\varepsilon(\boldsymbol{x})\cdot\{\varepsilon w(\boldsymbol{x}/\varepsilon)\nabla_{\boldsymbol{x}}\times\boldsymbol{v}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}} w(\boldsymbol{x}/\varepsilon)\times\boldsymbol{v}(\boldsymbol{x})\}\,dv \\ \qquad\qquad + ik_0\displaystyle\iiint_\Omega \varepsilon w(\boldsymbol{x}/\varepsilon)\boldsymbol{v}(\boldsymbol{x})\cdot\{\epsilon(\boldsymbol{x}/\varepsilon)\cdot\boldsymbol{E}^\varepsilon(\boldsymbol{x})\}\,dv = 0. \end{cases}$$

In the limit $\varepsilon \searrow 0$ we get

$$\begin{cases} \displaystyle\iiint_\Omega \boldsymbol{E}^\varepsilon(\boldsymbol{x})\cdot(\nabla_{\boldsymbol{y}} w(\boldsymbol{x}/\varepsilon)\times\boldsymbol{v}(\boldsymbol{x}))\,dv \to 0, \\ \displaystyle\iiint_\Omega \boldsymbol{H}^\varepsilon(\boldsymbol{x})\cdot(\nabla_{\boldsymbol{y}} w(\boldsymbol{x}/\varepsilon)\times\boldsymbol{v}(\boldsymbol{x}))\,dv \to 0, \end{cases}$$

since $\boldsymbol{E}^\varepsilon$ and $\boldsymbol{H}^\varepsilon$ are uniformly bounded in ε in the $L^2(\Omega;\mathbb{C}^3)$-norm. By the use of Theorem B.6, we get

$$\begin{cases} \displaystyle\iiint_\Omega\iiint_Y \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y})\cdot(\nabla_{\boldsymbol{y}} w(\boldsymbol{y})\times\boldsymbol{v}(\boldsymbol{x}))\,dv_{\boldsymbol{y}}\,dv_{\boldsymbol{x}} = 0, \\ \displaystyle\iiint_\Omega\iiint_Y \boldsymbol{H}_0(\boldsymbol{x},\boldsymbol{y})\cdot(\nabla_{\boldsymbol{y}} w(\boldsymbol{y})\times\boldsymbol{v}(\boldsymbol{x}))\,dv_{\boldsymbol{y}}\,dv_{\boldsymbol{x}} = 0, \end{cases}$$

which implies after cyclic permutation that

$$\begin{cases} \displaystyle\iiint_Y \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y})\times\nabla_{\boldsymbol{y}} w(\boldsymbol{y})\,dv_{\boldsymbol{y}} = \boldsymbol{0}, \\ \displaystyle\iiint_Y \boldsymbol{H}_0(\boldsymbol{x},\boldsymbol{y})\times\nabla_{\boldsymbol{y}} w(\boldsymbol{y})\,dv_{\boldsymbol{y}} = \boldsymbol{0}, \end{cases} \qquad \boldsymbol{x}\in\Omega \text{ a.e.}$$

for all $w \in H^1_{\#}(Y)$. The functions $\boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y})$ and $\boldsymbol{H}_0(\boldsymbol{x},\boldsymbol{y})$ both belong to the space $L^2(\Omega;L^2_{\#}(Y;\mathbb{C}^3))$. From Lemma B.5 we conclude that the fields $\boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y})$ and $\boldsymbol{H}_0(\boldsymbol{x},\boldsymbol{y})$ can be decomposed as

$$\begin{cases} \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{E}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}}\Phi_1(\boldsymbol{x},\boldsymbol{y}), \\ \boldsymbol{H}_0(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{H}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}}\Psi_1(\boldsymbol{x},\boldsymbol{y}), \end{cases}$$

where

$$\boldsymbol{E}(\boldsymbol{x}) = \langle \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y})\rangle = \iiint_Y \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y})\,dv_{\boldsymbol{y}}$$

and similarly for the field $\boldsymbol{H}_0(\boldsymbol{x},\boldsymbol{y})$. In summary,

$$\begin{cases} \boldsymbol{E}^\varepsilon(\boldsymbol{x}) \overset{2\text{-s}}{\rightharpoonup} \boldsymbol{E}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}}\Phi_1(\boldsymbol{x},\boldsymbol{y}), \\ \boldsymbol{H}^\varepsilon(\boldsymbol{x}) \overset{2\text{-s}}{\rightharpoonup} \boldsymbol{H}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}}\Psi_1(\boldsymbol{x},\boldsymbol{y}). \end{cases}$$

Multiplication of (4.1) by the admissible test functions $\boldsymbol{\phi} \in C_0^\infty(\Omega;\mathbb{C}^3)$ gives

$$\begin{cases} \iiint_\Omega \nabla_{\boldsymbol{x}} \times \boldsymbol{E}^\varepsilon(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x})\, dv - ik_0 \iiint_\Omega \boldsymbol{\phi}(\boldsymbol{x}) \cdot \{\mu(\boldsymbol{x}/\varepsilon) \cdot \boldsymbol{H}^\varepsilon(\boldsymbol{x})\}\, dv = 0, \\ \iiint_\Omega \nabla_{\boldsymbol{x}} \times \boldsymbol{H}^\varepsilon(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x})\, dv + ik_0 \iiint_\Omega \boldsymbol{\phi}(\boldsymbol{x}) \cdot \{\epsilon(\boldsymbol{x}/\varepsilon) \cdot \boldsymbol{E}^\varepsilon(\boldsymbol{x})\}\, dv = 0. \end{cases}$$

In the limit $\varepsilon \searrow 0$ we get

(4.7)
$$\begin{cases} \iiint_\Omega \nabla_{\boldsymbol{x}} \times \boldsymbol{E}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x})\, dv_{\boldsymbol{x}} \\ \qquad - ik_0 \iiint_\Omega \iiint_Y \boldsymbol{\phi}(\boldsymbol{x}) \cdot \mu(\boldsymbol{y}) \cdot (\boldsymbol{H}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}}\Psi_1(\boldsymbol{x},\boldsymbol{y}))\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}} = 0, \\ \iiint_\Omega \nabla_{\boldsymbol{x}} \times \boldsymbol{H}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x})\, dv_{\boldsymbol{x}} \\ \qquad + ik_0 \iiint_\Omega \iiint_Y \boldsymbol{\phi}(\boldsymbol{x}) \cdot \epsilon(\boldsymbol{y}) \cdot (\boldsymbol{E}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}}\Phi_1(\boldsymbol{x},\boldsymbol{y}))\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}} = 0. \end{cases}$$

Here we have used Theorem B.8, which states that

$$\nabla \times \boldsymbol{E}^\varepsilon \overset{2\text{-s}}{\rightharpoonup} \nabla_{\boldsymbol{x}} \times \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y}) + \nabla_{\boldsymbol{y}} \times \boldsymbol{E}_1(\boldsymbol{x},\boldsymbol{y}),$$

which gives the weak limit $\nabla_{\boldsymbol{x}} \times \boldsymbol{E}(\boldsymbol{x})$ since the admissible test function $\boldsymbol{\phi}$ does not depend on $\boldsymbol{y}$.

The divergence equations are multiplied by $v(\boldsymbol{x}) = \varepsilon\psi(\boldsymbol{x})\phi(\boldsymbol{x}/\varepsilon)$, where $\psi \in C_0^\infty(\Omega)$, $\phi \in H^1_\#(Y)$. We note that $w_\epsilon(\boldsymbol{y}) = \hat{\boldsymbol{e}}_i \cdot \epsilon(\boldsymbol{y}) \cdot \hat{\boldsymbol{e}}_j \in L^\infty_\#(Y)$ and $w_\mu(\boldsymbol{y}) = \hat{\boldsymbol{e}}_i \cdot \mu(\boldsymbol{y}) \cdot \hat{\boldsymbol{e}}_j \in L^\infty_\#(Y)$, which implies that $w_\epsilon(\boldsymbol{y})\nabla_{\boldsymbol{y}}\phi$ and $w_\mu(\boldsymbol{y})\nabla_{\boldsymbol{y}}\phi \in L^2_\#(Y;\mathbb{C}^3)$. Theorem B.3 and an integration by parts give

$$\begin{aligned} &\lim_{\varepsilon\searrow 0} \iiint_\Omega \nabla \cdot \{\epsilon(\boldsymbol{x}/\varepsilon) \cdot \boldsymbol{E}^\varepsilon(\boldsymbol{x})\}\, \varepsilon\psi(\boldsymbol{x})\phi(\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} \\ &\quad = -\lim_{\varepsilon\searrow 0} \iiint_\Omega \{\varepsilon\nabla\psi(\boldsymbol{x})\phi(\boldsymbol{x}/\varepsilon) + \psi(\boldsymbol{x})\nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}/\varepsilon)\} \cdot \{\epsilon(\boldsymbol{x}/\varepsilon) \cdot \boldsymbol{E}^\varepsilon(\boldsymbol{x})\}\, dv_{\boldsymbol{x}} \\ &\quad = -\iiint_\Omega \iiint_Y \psi(\boldsymbol{x})\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y}) \cdot \epsilon(\boldsymbol{y}) \cdot \{\boldsymbol{E}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}}\Phi_1(\boldsymbol{x},\boldsymbol{y})\}\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}} = 0 \end{aligned}$$

for all $\phi \in H^1_\#(Y)$ and all $\psi \in H^1_0(\Omega)$. Using similar arguments for the magnetic field, we get the local equations

(4.8)
$$\begin{cases} \iiint_Y \nabla_{\boldsymbol{y}}\phi(\boldsymbol{y}) \cdot \epsilon(\boldsymbol{y}) \cdot \{\boldsymbol{E}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}}\Phi_1(\boldsymbol{x},\boldsymbol{y})\}\, dv_{\boldsymbol{y}} = 0, \\ \iiint_Y \nabla_{\boldsymbol{y}}\phi(\boldsymbol{y}) \cdot \mu(\boldsymbol{y}) \cdot \{\boldsymbol{H}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}}\Psi_1(\boldsymbol{x},\boldsymbol{y})\}\, dv_{\boldsymbol{y}} = 0, \end{cases} \qquad \boldsymbol{x} \in \Omega \text{ a.e.}$$

Define the vector fields

$$\boldsymbol{\chi}_e(\boldsymbol{y}) = \sum_{i=1}^3 \chi_e^i(\boldsymbol{y})\hat{\boldsymbol{e}}_i, \qquad \boldsymbol{\chi}_h(\boldsymbol{y}) = \sum_{i=1}^3 \chi_h^i(\boldsymbol{y})\hat{\boldsymbol{e}}_i.$$

The variables can be separated by using the ansatz

$$\begin{cases} \nabla_{\boldsymbol{y}}\Phi_1(\boldsymbol{x},\boldsymbol{y}) = -\nabla_{\boldsymbol{y}}\boldsymbol{\chi}_e(\boldsymbol{y})\cdot\boldsymbol{E}(\boldsymbol{x}), \\ \nabla_{\boldsymbol{y}}\Psi_1(\boldsymbol{x},\boldsymbol{y}) = -\nabla_{\boldsymbol{y}}\boldsymbol{\chi}_h(\boldsymbol{y})\cdot\boldsymbol{H}(\boldsymbol{x}) \end{cases}$$

inserted into (4.8), which gives

$$\begin{cases} \langle\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y})\cdot(\epsilon(\boldsymbol{y}) - \epsilon(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}\boldsymbol{\chi}_e(\boldsymbol{y}))\rangle\cdot\boldsymbol{E}(\boldsymbol{x}) = 0, \\ \langle\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y})\cdot(\mu(\boldsymbol{y}) - \mu(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}\boldsymbol{\chi}_h(\boldsymbol{y}))\rangle\cdot\boldsymbol{H}(\boldsymbol{x}) = 0 \end{cases}$$

for all $\phi \in H^1_{\#}(Y)$, i.e.,

$$\begin{cases} \nabla_{\boldsymbol{y}}\cdot(\epsilon(\boldsymbol{y}) - \epsilon(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}\boldsymbol{\chi}_e(\boldsymbol{y})) = 0, \\ \nabla_{\boldsymbol{y}}\cdot(\mu(\boldsymbol{y}) - \mu(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}\boldsymbol{\chi}_h(\boldsymbol{y})) = 0 \end{cases}$$

a.e. in $\Omega \times Y$. Inserting the solutions of the local equations into (4.7) yields the macroscopic homogenized equations

$$\begin{cases} \displaystyle\iiint_\Omega \nabla_{\boldsymbol{x}}\times\boldsymbol{E}(\boldsymbol{x})\cdot\boldsymbol{\phi}(\boldsymbol{x})\,dv_{\boldsymbol{x}} \\ \quad \displaystyle - ik_0\iiint_\Omega\iiint_Y \boldsymbol{\phi}(\boldsymbol{x})\cdot(\mu(\boldsymbol{y}) - \mu(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}\boldsymbol{\chi}_h(\boldsymbol{y}))\,dv_{\boldsymbol{y}}\cdot\boldsymbol{H}(\boldsymbol{x})\,dv_{\boldsymbol{x}} = 0, \\ \displaystyle\iiint_\Omega \nabla_{\boldsymbol{x}}\times\boldsymbol{H}(\boldsymbol{x})\cdot\boldsymbol{\phi}(\boldsymbol{x})\,dv_{\boldsymbol{x}} \\ \quad \displaystyle + ik_0\iiint_\Omega\iiint_Y \boldsymbol{\phi}(\boldsymbol{x})\cdot(\epsilon(\boldsymbol{y}) - \epsilon(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}\boldsymbol{\chi}_e(\boldsymbol{y}))\,dv_{\boldsymbol{y}}\cdot\boldsymbol{E}(\boldsymbol{x})\,dv_{\boldsymbol{x}} = 0 \end{cases}$$

and

$$\begin{cases} \nabla\cdot\boldsymbol{B}(\boldsymbol{x}) = 0, \\ \nabla\cdot\boldsymbol{D}(\boldsymbol{x}) = 0, \end{cases}$$

which defines the homogenized permeability and permittivity as

$$\begin{cases} \epsilon^h = \displaystyle\iiint_Y \epsilon(\boldsymbol{y})\cdot(\mathbf{I}_3 - \nabla_{\boldsymbol{y}}\boldsymbol{\chi}_e(\boldsymbol{y}))\,dv_{\boldsymbol{y}}, \\ \mu^h = \displaystyle\iiint_Y \mu(\boldsymbol{y})\cdot(\mathbf{I}_3 - \nabla_{\boldsymbol{y}}\boldsymbol{\chi}_h(\boldsymbol{y}))\,dv_{\boldsymbol{y}}, \end{cases}$$

i.e., $\boldsymbol{B} = \mu^h\cdot\boldsymbol{H}$ and $\boldsymbol{D} = \epsilon^h\cdot\boldsymbol{E}$. The existence of a unique solution of the homogenized system follows from the properties of the homogenized permeability and permittivity, μ^h and ϵ^h, respectively (see section 4.2.1), which satisfies the same assumptions as the material properties for the heterogeneous system. □

4.2.1. The properties of the homogenized parameters. An immediate consequence of Theorem 4.2 is that the homogenized parameters are independent of the properties of the domain Ω and of the properties of the incident field. Moreover, the homogenized material properties satisfy the same assumptions as the heterogeneous parameters do, i.e., they are coercive and bounded. Coercivity and boundedness follow from the fact that the homogenized parameters are bounded from below and above by the harmonic and arithmetic averages of the heterogeneous parameters; hence the

homogenized parameters are bounded from below and above (e.g., see [5] or [24]). If the heterogeneous material parameters are symmetric (reciprocal material), then the homogenized parameters are also symmetric as proved below.

PROPOSITION 4.3. *The homogenized permeability and permittivity are symmetric, provided the heterogeneous parameters are symmetric.*

Proof of Proposition 4.3. We restrict ourselves to the electric parameters since the arguments for the permeability are the same. By assumption the material parameters are symmetrical, i.e., $\epsilon(\boldsymbol{y}) = \epsilon^t(\boldsymbol{y})$ and $\mu(\boldsymbol{y}) = \mu^t(\boldsymbol{y})$.

We define the average over the Y-cell by

$$\langle f \rangle = \iiint_Y f(\boldsymbol{y})\, dv_{\boldsymbol{y}}.$$

The local problem, (4.6), can be written as ($i = 1, 2, 3$)

$$\langle \nabla_{\boldsymbol{y}} w(\boldsymbol{y}) \cdot \epsilon(\boldsymbol{y}) \cdot \hat{\boldsymbol{e}}_i \rangle = \left\langle \nabla_{\boldsymbol{y}} w(\boldsymbol{y}) \cdot \epsilon(\boldsymbol{y}) \cdot \nabla_{\boldsymbol{y}} \chi_e^i(\boldsymbol{y}) \right\rangle$$

for all $w \in H^1_{\#}(Y)$. We rewrite these equations in one set of equations (see (4.5))

$$\langle \nabla_{\boldsymbol{y}} w(\boldsymbol{y}) \cdot \epsilon(\boldsymbol{y}) \rangle = \langle \nabla_{\boldsymbol{y}} w(\boldsymbol{y}) \cdot \epsilon(\boldsymbol{y}) \cdot \nabla_{\boldsymbol{y}} \boldsymbol{\chi}_e(\boldsymbol{y}) \rangle$$

for all $w \in H^1_{\#}(Y)$. Due to the symmetry in ϵ we get

$$\langle \epsilon(\boldsymbol{y}) \cdot \nabla_{\boldsymbol{y}} \boldsymbol{\chi}(\boldsymbol{y}) \rangle = \left\langle (\nabla_{\boldsymbol{y}} \boldsymbol{\chi}_e(\boldsymbol{y}))^t \cdot \epsilon(\boldsymbol{y}) \cdot \nabla_{\boldsymbol{y}} \boldsymbol{\chi}_e(\boldsymbol{y}) \right\rangle$$

if we choose $w = \chi_e^i$.

The homogenized parameters in (4.4) are

$$\begin{aligned} \epsilon^h &= \langle \epsilon(\boldsymbol{y}) \rangle - \langle \epsilon(\boldsymbol{y}) \cdot \nabla_{\boldsymbol{y}} \boldsymbol{\chi}_e(\boldsymbol{y}) \rangle \\ &= \langle \epsilon(\boldsymbol{y}) \rangle - \left\langle (\nabla_{\boldsymbol{y}} \boldsymbol{\chi}_e(\boldsymbol{y}))^t \cdot \epsilon(\boldsymbol{y}) \cdot \nabla_{\boldsymbol{y}} \boldsymbol{\chi}_e(\boldsymbol{y}) \right\rangle, \end{aligned}$$

which proves that ϵ^h is symmetric. ∎

4.2.2. Correctors. This section is concluded by the proof of a new result on correctors.

We begin with the two-scale limit of the heterogeneous system (4.1), which is given by

$$\text{(4.9)} \quad \begin{cases} \displaystyle\iiint_\Omega \iiint_Y (\nabla_{\boldsymbol{x}} \times \boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{y}) + \nabla_{\boldsymbol{y}} \times \boldsymbol{E}_1(\boldsymbol{x}, \boldsymbol{y})) \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\, dv_{\boldsymbol{y}} dv_{\boldsymbol{x}} \\ \displaystyle = ik_0 \iiint_\Omega \iiint_Y \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \cdot \mu(\boldsymbol{y}) \cdot \boldsymbol{H}_0(\boldsymbol{x}, \boldsymbol{y})\, dv_{\boldsymbol{y}} dv_{\boldsymbol{x}}, \\ \displaystyle\iiint_\Omega \iiint_Y (\nabla_{\boldsymbol{x}} \times \boldsymbol{H}_0(\boldsymbol{x}, \boldsymbol{y}) + \nabla_{\boldsymbol{y}} \times \boldsymbol{H}_1(\boldsymbol{x}, \boldsymbol{y})) \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\, dv_{\boldsymbol{y}} dv_{\boldsymbol{x}} \\ \displaystyle = -ik_0 \iiint_\Omega \iiint_Y \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \cdot \epsilon(\boldsymbol{y}) \cdot \boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{y})\, dv_{\boldsymbol{y}} dv_{\boldsymbol{x}} \end{cases}$$

for all $\boldsymbol{\phi} \in D(\Omega; C^\infty_{\#}(Y; \mathbb{C}^3))$. These equations follow from the fact that (see Appendix B)

$$\boldsymbol{E}^\varepsilon(\boldsymbol{x}) \overset{\text{2-s}}{\rightharpoonup} \boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{y})$$

and

$$\nabla \times \boldsymbol{E}^\varepsilon(\boldsymbol{x}) \overset{\text{2-s}}{\rightharpoonup} \nabla_{\boldsymbol{x}} \times \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y}) + \nabla_{\boldsymbol{y}} \times \boldsymbol{E}_1(\boldsymbol{x},\boldsymbol{y}),$$

where

$$\begin{cases} \boldsymbol{E}_0 \in L^2(\Omega; L^2_{\#}(Y;\mathbb{C}^3)), \\ \nabla_{\boldsymbol{x}} \times \boldsymbol{E}_0 \in L^2(\Omega; L^2_{\#}(Y;\mathbb{C}^3)), \\ \boldsymbol{E}_1 \in L^2(\Omega; H_{\#}(\mathrm{rot}, Y)/\mathbb{C}). \end{cases}$$

The system (4.9) contains macroscopic and microscopic information which gives the homogenized system when averaged over the local scale. The local equations and the two-scale limit system (4.9) provide us with the following correctors in the case when the solution of the homogenized system is smooth enough.

THEOREM 4.4. *Let $\boldsymbol{E}^\varepsilon, \boldsymbol{H}^\varepsilon$ be the solution of* (4.1)*, let $\boldsymbol{E}, \boldsymbol{H}$ be the solution of the homogenized Maxwell equations* (4.4)*, and let $\boldsymbol{E}_1, \boldsymbol{H}_1$ solve the two-scale limit system* (4.9)*. If $\boldsymbol{E}_0$, $\boldsymbol{H}_0$, $\boldsymbol{E}_1$, $\boldsymbol{H}_1$, $\nabla_{\boldsymbol{x}} \times \boldsymbol{E}_0$, $\nabla_{\boldsymbol{x}} \times \boldsymbol{H}_0$, $\nabla_{\boldsymbol{x}} \times \boldsymbol{E}_1$, $\nabla_{\boldsymbol{x}} \times \boldsymbol{H}_1$, $\nabla_{\boldsymbol{y}} \times \boldsymbol{E}_1$, and $\nabla_{\boldsymbol{y}} \times \boldsymbol{H}_1$ are admissible test functions, then*

$$\begin{cases} \lim_{\varepsilon\to 0} \|\boldsymbol{E}^\varepsilon(\boldsymbol{x}) - \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{x}/\varepsilon) - \varepsilon \boldsymbol{E}_1(\boldsymbol{x},\boldsymbol{x}/\varepsilon)\|_{H(\mathrm{rot},\Omega)} = 0, \\ \lim_{\varepsilon\to 0} \|\boldsymbol{H}^\varepsilon(\boldsymbol{x}) - \boldsymbol{H}_0(\boldsymbol{x},\boldsymbol{x}/\varepsilon) - \varepsilon \boldsymbol{H}_1(\boldsymbol{x},\boldsymbol{x}/\varepsilon)\|_{H(\mathrm{rot},\Omega)} = 0, \end{cases}$$

where

$$\begin{cases} \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{E}(\boldsymbol{x}) - \nabla_{\boldsymbol{y}} \boldsymbol{\chi}_e(\boldsymbol{y}) \cdot \boldsymbol{E}(\boldsymbol{x}), \\ \boldsymbol{H}_0(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{H}(\boldsymbol{x}) - \nabla_{\boldsymbol{y}} \boldsymbol{\chi}_h(\boldsymbol{y}) \cdot \boldsymbol{H}(\boldsymbol{x}), \end{cases}$$

$$\boldsymbol{\chi}_e(\boldsymbol{y}) = \sum_{i=1}^{3} \chi_e^i(\boldsymbol{y}) \hat{\boldsymbol{e}}_i, \qquad \boldsymbol{\chi}_h(\boldsymbol{y}) = \sum_{i=1}^{3} \chi_h^i(\boldsymbol{y}) \hat{\boldsymbol{e}}_i,$$

and where $\chi_e^i(\boldsymbol{y})$ and $\chi_h^i(\boldsymbol{y})$, $i = 1,2,3$, in $H^1_{\#}(Y)$ solve the local problems (4.6)*.*

Proof. The assumptions imply that (see Theorem B.8)

$$\begin{cases} \boldsymbol{E}^\varepsilon \overset{\text{2-s}}{\rightharpoonup} \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y}), \\ \nabla \times \boldsymbol{E}^\varepsilon \overset{\text{2-s}}{\rightharpoonup} \nabla_{\boldsymbol{x}} \times \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y}) + \nabla_{\boldsymbol{y}} \times \boldsymbol{E}_1(\boldsymbol{x},\boldsymbol{y}) \end{cases}$$

and $\nabla_{\boldsymbol{y}} \times \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0}$.

The proof is carried out using the sesquilinear form

$$Q^\varepsilon(\boldsymbol{u},\boldsymbol{v}) = -\iiint_\Omega \left\{ \frac{1}{ik_0} (\nabla \times \boldsymbol{v}^*) \cdot \mu^{-1}(\boldsymbol{x}/\varepsilon) \cdot (\nabla \times \boldsymbol{u}) + ik_0 \boldsymbol{v}^* \cdot \epsilon(\boldsymbol{x}/\varepsilon) \cdot \boldsymbol{u} \right\} dv,$$

which is identical to (4.3) but without the surface integral term.

The coercivity assumption, (2.1), implies

$$C\|\boldsymbol{u}(\boldsymbol{x})\|^2_{H(\mathrm{rot},\Omega)} \le \Re Q^\varepsilon(\boldsymbol{u},\boldsymbol{u}).$$

We get

$$C\|\boldsymbol{E}^\varepsilon(\boldsymbol{x}) - \boldsymbol{E}_0(\boldsymbol{x},\boldsymbol{x}/\varepsilon) - \varepsilon \boldsymbol{E}_1(\boldsymbol{x},\boldsymbol{x}/\varepsilon)\|^2_{H(\mathrm{rot},\Omega)} \le I_1^\varepsilon + I_2^\varepsilon,$$

where

$$\begin{cases} I_1^\varepsilon = \Re Q^\varepsilon \left(\boldsymbol{E}^\varepsilon(\boldsymbol{x}), \boldsymbol{A}_\varepsilon(\boldsymbol{x}) \right), \\ I_2^\varepsilon = -\Re Q^\varepsilon \left(\boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) + \varepsilon \boldsymbol{E}_1(\boldsymbol{x}, \boldsymbol{x}/\varepsilon), \boldsymbol{A}_\varepsilon(\boldsymbol{x}) \right), \end{cases}$$

where, for short, we denote $\boldsymbol{A}_\varepsilon(\boldsymbol{x}) = \boldsymbol{E}^\varepsilon(\boldsymbol{x}) - \boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) - \varepsilon \boldsymbol{E}_1(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)$. Due to the assumptions of the fields in $\boldsymbol{A}_\varepsilon(\boldsymbol{x})$, we have

$$\begin{cases} \boldsymbol{A}_\varepsilon \overset{\text{2-s}}{\rightharpoonup} \boldsymbol{0}, \\ \nabla \times \boldsymbol{A}_\varepsilon \overset{\text{2-s}}{\rightharpoonup} \boldsymbol{0}, \end{cases}$$

since

$$\begin{cases} \boldsymbol{E}^\varepsilon \overset{\text{2-s}}{\rightharpoonup} \boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{y}), \\ \nabla \times \boldsymbol{E}^\varepsilon \overset{\text{2-s}}{\rightharpoonup} \nabla_{\boldsymbol{x}} \times \boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{y}) + \nabla_{\boldsymbol{y}} \times \boldsymbol{E}_1(\boldsymbol{x}, \boldsymbol{y}) \end{cases}$$

and $\nabla_{\boldsymbol{y}} \times \boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}$.

We start by analyzing the first integral I_1^ε. Since $\boldsymbol{E}^\varepsilon$ satisfies the Maxwell equations, (4.1), we get

$$I_1^\varepsilon = \Re \iiint_\Omega (\nabla \times \boldsymbol{H}^\varepsilon(\boldsymbol{x})) \cdot \boldsymbol{A}_\varepsilon(\boldsymbol{x})^* \, dv - \Re \iiint_\Omega \boldsymbol{H}^\varepsilon(\boldsymbol{x}) \cdot (\nabla \times \boldsymbol{A}_\varepsilon(\boldsymbol{x}))^* \, dv.$$

We now use $\nabla \cdot (\nabla \times \boldsymbol{H}^\varepsilon) = 0$ and $\nabla \cdot (\nabla \times \boldsymbol{A}_\varepsilon) = 0$ and, moreover, the fact that $\nabla \times \boldsymbol{H}^\varepsilon \in L^2(\Omega; \mathbb{C}^3)$ and $\nabla \times \boldsymbol{A}_\varepsilon \in L^2(\Omega; \mathbb{C}^3)$. The div-curl lemma (see [24, 25]) can be used and the limit is zero, since

$$\boldsymbol{A}_\varepsilon(\boldsymbol{x}) \rightharpoonup \boldsymbol{0} \text{ and } \nabla \times \boldsymbol{A}_\varepsilon(\boldsymbol{x}) \rightharpoonup \boldsymbol{0}$$

weakly in $L^2(\Omega; \mathbb{C}^3)$.

The second integral is now analyzed:

$$\begin{aligned} I_2^\varepsilon = -\Re \iiint_\Omega \Big\{ & \frac{1}{ik_0} \left(\nabla \times \boldsymbol{A}_\varepsilon(\boldsymbol{x}) \right)^* \cdot \mu^{-1}(\boldsymbol{x}/\varepsilon) \\ & \cdot \left(\nabla_{\boldsymbol{x}} \times \boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) + \varepsilon \nabla_{\boldsymbol{x}} \times \boldsymbol{E}_1(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) + \nabla_{\boldsymbol{y}} \times \boldsymbol{E}_1(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) \right) \\ & + ik_0 \boldsymbol{A}_\varepsilon(\boldsymbol{x})^* \cdot \epsilon(\boldsymbol{x}/\varepsilon) \cdot \left(\boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) + \varepsilon \boldsymbol{E}_1(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) \right) \Big\} \, dv_{\boldsymbol{x}}. \end{aligned}$$

We pass to the limit, $\varepsilon \searrow 0$, and use that $\mu^{-1}(\boldsymbol{x}/\varepsilon) \cdot (\nabla_{\boldsymbol{x}} \times \boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) + \varepsilon \nabla_{\boldsymbol{x}} \times \boldsymbol{E}_1(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) + \nabla_{\boldsymbol{y}} \times \boldsymbol{E}_1(\boldsymbol{x}, \boldsymbol{x}/\varepsilon))$ and $\epsilon(\boldsymbol{x}/\varepsilon) \cdot (\boldsymbol{E}_0(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) + \varepsilon \boldsymbol{E}_1(\boldsymbol{x}, \boldsymbol{x}/\varepsilon))$ are admissible test functions and obtain

$$\lim_{\varepsilon \searrow 0} I_2^\varepsilon = 0;$$

the theorem is proved. ∎

Remark 4.1. It is still an open question how irregular a function can be and still be an admissible test function. However, if the homogenized solution $\boldsymbol{E} \in C(\overline{\Omega}; \mathbb{C}^3)$, then $\boldsymbol{E}_0 \in L^2_{\#}(Y; C(\overline{\Omega}; \mathbb{C}^3))$ is admissible (see Appendix B). Further, if $\boldsymbol{E} \in C(\overline{\Omega}; \mathbb{C}^3)$, then $\boldsymbol{H} \in C(\overline{\Omega}; \mathbb{C}^3)$ by symmetry, and via (4.9) we find that $\nabla_{\boldsymbol{x}} \times \boldsymbol{E}_0 + \nabla_{\boldsymbol{y}} \times \boldsymbol{E}_1$ is smooth in x, and for sufficient smoothness $\nabla_{\boldsymbol{x}} \times \boldsymbol{E}_1$ is also an admissible test function. To the knowledge of the authors there exist no results about regularity of the solutions of the Maxwell equations in the anisotropic, constant coefficient case. However, we believe that for sufficient regular boundary and incident fields, the solutions are admissible test functions.

5. Examples. In this section, we give two explicit examples of the exterior Calderón operator.

5.1. Plane boundary. The general representation of the solution to Problem (R) in (2.7) in a region $x_3 > c$ (plane interface Ω, $x_3 = c$) is found by a Fourier transform in the lateral coordinates x_1 and x_2.

The Fourier transform $\boldsymbol{E}(\boldsymbol{\xi}, x_3)$ of the electric field $\boldsymbol{E}(\boldsymbol{x})$, $\boldsymbol{x} = \hat{\boldsymbol{e}}_1 x_1 + \hat{\boldsymbol{e}}_2 x_2 + \hat{\boldsymbol{e}}_3 x_3$, with respect to the lateral position vector $\boldsymbol{\rho} = \hat{\boldsymbol{e}}_1 x_1 + \hat{\boldsymbol{e}}_2 x_2$ is defined by

$$\boldsymbol{E}(\boldsymbol{\xi}, x_3) = \iint_{\mathbb{R}^2} \boldsymbol{E}(\boldsymbol{x}) e^{-i\boldsymbol{\xi}\cdot\boldsymbol{\rho}}\, d\boldsymbol{\rho},$$

where the Fourier variable $\boldsymbol{\xi}$ is

$$\boldsymbol{\xi} = \hat{\boldsymbol{e}}_1 \xi_1 + \hat{\boldsymbol{e}}_2 \xi_2$$

and $d\boldsymbol{\rho} = dx_1\, dx_2$. The modulus of this vector is denoted ξ, i.e.,

$$\xi = \sqrt{\xi_1^2 + \xi_2^2}.$$

By the Fourier inversion formula,

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{1}{4\pi^2} \iint_{\mathbb{R}^2} \boldsymbol{E}(\boldsymbol{\xi}, x_3) e^{i\boldsymbol{\xi}\cdot\boldsymbol{\rho}}\, d\boldsymbol{\xi},$$

where $d\boldsymbol{\xi} = d\xi_1\, d\xi_2$. Specifically, the tangential electric field on the surface $\partial\Omega$ is

$$-\,\hat{\boldsymbol{e}}_3 \times (\hat{\boldsymbol{e}}_3 \times \boldsymbol{E}(\boldsymbol{x}))|_{\partial\Omega} = \frac{1}{4\pi^2} \iint_{\mathbb{R}^2} \boldsymbol{A}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}\cdot\boldsymbol{\rho}}\, d\boldsymbol{\xi},$$

where $\boldsymbol{A}(\boldsymbol{\xi})$ is the Fourier transform of the trace of the tangential electric field.

The general form of the solution to Problem (R) in (2.7) in a region $x_3 > c$ is (see [16])

$$\begin{cases} \boldsymbol{E}(\boldsymbol{x}) = \frac{1}{4\pi^2} \iint_{\mathbb{R}^2} \left(\mathbf{I}_2 - \dfrac{\xi}{\xi_3} \hat{\boldsymbol{e}}_3 \hat{\boldsymbol{e}}_\parallel \right) \cdot \boldsymbol{A}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}\cdot\boldsymbol{\rho} + i\xi_3(x_3 - c)}\, d\boldsymbol{\xi}, \\ \boldsymbol{H}(\boldsymbol{x}) = \frac{1}{4\pi^2} \iint_{\mathbb{R}^2} \left(\dfrac{\boldsymbol{\xi}}{k_0} + \dfrac{\xi_3}{k_0} \hat{\boldsymbol{e}}_3 \right) \times \left(\mathbf{I}_2 - \dfrac{\xi}{\xi_3} \hat{\boldsymbol{e}}_3 \hat{\boldsymbol{e}}_\parallel \right) \cdot \boldsymbol{A}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}\cdot\boldsymbol{\rho} + i\xi_3(x_3 - c)}\, d\boldsymbol{\xi}, \end{cases}$$

where $\mathbf{I}_2$ is the identity dyadic in $\mathbb{R}^2$, and a pertinent orthogonal basis in $\mathbb{R}^2$ is $\{\hat{\boldsymbol{e}}_\parallel, \hat{\boldsymbol{e}}_\perp\}$, defined by

$$\hat{\boldsymbol{e}}_\parallel = \boldsymbol{\xi}/\xi, \qquad \hat{\boldsymbol{e}}_\perp = \hat{\boldsymbol{e}}_3 \times \hat{\boldsymbol{e}}_\parallel$$

and where

$$\xi_3 = \left(k_0^2 - \xi^2\right)^{1/2} = \begin{cases} \sqrt{k_0^2 - \xi^2} & \text{for } \xi < k_0, \\ i\sqrt{\xi^2 - k_0^2} & \text{for } \xi > k_0 \end{cases}$$

and the standard convention of the square root of a nonnegative argument is intended.

The representation of the fields can be simplified using dyadic calculus:

$$\begin{cases} \boldsymbol{E}(\boldsymbol{x}) = \dfrac{1}{4\pi^2} \iint_{\mathbb{R}^2} \left(\mathbf{I}_2 - \dfrac{\xi}{\xi_3} \hat{\boldsymbol{e}}_3 \hat{\boldsymbol{e}}_\parallel \right) \cdot \boldsymbol{A}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}\cdot\boldsymbol{\rho} + i\xi_3(x_3 - c)}\, d\boldsymbol{\xi}, \\ \boldsymbol{H}(\boldsymbol{x}) = \dfrac{1}{4\pi^2} \iint_{\mathbb{R}^2} \left(\dfrac{\xi}{k_0} \hat{\boldsymbol{e}}_3 \hat{\boldsymbol{e}}_\perp + \dfrac{k_0}{\xi_3} \hat{\boldsymbol{e}}_\perp \hat{\boldsymbol{e}}_\parallel - \dfrac{\xi_3}{k_0} \hat{\boldsymbol{e}}_\parallel \hat{\boldsymbol{e}}_\perp \right) \cdot \boldsymbol{A}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}\cdot\boldsymbol{\rho} + i\xi_3(x_3 - c)}\, d\boldsymbol{\xi}. \end{cases}$$

From these relations the exterior Calderón operator is the transformation from

$$\hat{\boldsymbol{e}}_3 \times \boldsymbol{E}(\boldsymbol{x})|_{\partial\Omega} = \frac{1}{4\pi^2} \iint_{\mathbb{R}^2} \hat{\boldsymbol{e}}_3 \times \boldsymbol{A}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}\cdot\boldsymbol{\rho}}\, d\boldsymbol{\xi}$$

to

$$\hat{\boldsymbol{e}}_3 \times \boldsymbol{H}(\boldsymbol{x})|_{\partial\Omega} = -\frac{1}{4\pi^2} \iint_{\mathbb{R}^2} \left(\frac{k_0}{\xi_3} \hat{\boldsymbol{e}}_\parallel \hat{\boldsymbol{e}}_\parallel - \frac{\xi_3}{k_0} \hat{\boldsymbol{e}}_\perp \hat{\boldsymbol{e}}_\perp \right) \cdot \boldsymbol{A}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}\cdot\boldsymbol{\rho}}\, d\boldsymbol{\xi},$$

where the vector field $\boldsymbol{A}(\boldsymbol{\xi})$ is determined from $\hat{\boldsymbol{e}}_3 \times \boldsymbol{E}(\boldsymbol{x})|_{\partial\Omega}$ by

$$\boldsymbol{A}(\boldsymbol{\xi}) = -\iint_{\mathbb{R}^2} \hat{\boldsymbol{e}}_3 \times (\hat{\boldsymbol{e}}_3 \times \boldsymbol{E}(\boldsymbol{x}))|_{\partial\Omega}\, e^{-i\boldsymbol{\xi}\cdot\boldsymbol{\rho}}\, d\boldsymbol{\rho}.$$

We note that in this example the domain and the boundary are unbounded, which yields other function spaces for the traces. We refer to [9] for the details.

5.2. Spherical boundary. For a spherical boundary, $x = a$, the exterior Calderón operator can be represented in a series of vector spherical waves; see Appendix C.

The general form of the solution to Problem (R) in (2.7) in a region $x > a$ is (see (C.2) and (C.3))

$$\begin{cases} \boldsymbol{E}(\boldsymbol{x}) = \displaystyle\sum_{\tau n} a_{\tau n} \boldsymbol{u}_{\tau n}(k_0 \boldsymbol{x}), \\ \boldsymbol{H}(\boldsymbol{x}) = -i \displaystyle\sum_{\tau n} a_{\tau n} \boldsymbol{u}_{\bar{\tau} n}(k_0 \boldsymbol{x}), \end{cases}$$

where the index $\bar{\tau}$ is the dual index of τ, defined by $\bar{1} = 2$ and $\bar{2} = 1$.

The traces of the electric and the magnetic fields are ($\kappa = k_0 a$)

$$\begin{cases} \hat{\boldsymbol{x}} \times \boldsymbol{E}(\boldsymbol{x})|_{\partial\Omega} = \displaystyle\sum_n \left(a_{1n} h_l^{(1)}(\kappa) \boldsymbol{A}_{2n}(\hat{\boldsymbol{x}}) - a_{2n} \frac{(\kappa h_l^{(1)}(\kappa))'}{\kappa} \boldsymbol{A}_{1n}(\hat{\boldsymbol{x}}) \right), \\ \hat{\boldsymbol{x}} \times \boldsymbol{H}(\boldsymbol{x})|_{\partial\Omega} = -i \displaystyle\sum_n \left(a_{2n} h_l^{(1)}(\kappa) \boldsymbol{A}_{2n}(\hat{\boldsymbol{x}}) - a_{1n} \frac{(\kappa h_l^{(1)}(\kappa))'}{\kappa} \boldsymbol{A}_{1n}(\hat{\boldsymbol{x}}) \right). \end{cases}$$

For a given tangential field $\hat{\boldsymbol{x}} \times \boldsymbol{E}(\boldsymbol{x})|_{\partial\Omega}$, the expansion coefficients $a_{\tau n}$ are found by the orthogonality relation (see (C.1))

$$\begin{cases} a_{1n} = \dfrac{1}{h_l^{(1)}(\kappa)} \displaystyle\iint_\gamma \boldsymbol{A}_{2n}(\hat{\boldsymbol{x}}) \cdot (\hat{\boldsymbol{x}} \times \boldsymbol{E}(\boldsymbol{x})|_{\partial\Omega}), \\ a_{2n} = -\dfrac{\kappa}{(\kappa h_l^{(1)}(\kappa))'} \displaystyle\iint_\gamma \boldsymbol{A}_{1n}(\hat{\boldsymbol{x}}) \cdot (\hat{\boldsymbol{x}} \times \boldsymbol{E}(\boldsymbol{x})|_{\partial\Omega}). \end{cases}$$

The exterior Calderón mapping is the mapping from $\hat{\boldsymbol{x}} \times \boldsymbol{E}(\boldsymbol{x})|_{\partial\Omega}$ (which determines the expansion coefficients $a_{\tau n}$ uniquely) to $\hat{\boldsymbol{x}} \times \boldsymbol{H}(\boldsymbol{x})|_{\partial\Omega}$.

6. Conclusions. This paper analyzes the homogenization of the Maxwell equations for a material with periodic microscale. The material can be anisotropic and satisfies a coercivity condition (passive material), and the sources of the excitation are located in the region outside the heterogeneous material in Ω. We utilize the concept of two-scale convergence. A new a priori estimate is established, and a proof of strong convergence of the corrector fields is presented. The homogenized parameters are shown to be independent of the properties of the domain Ω and of the properties of the incident field.

Appendix A. Function spaces. In this appendix, we list the various function spaces used in this paper. Let Ω be a bounded, open, simply connected set in $\mathbb{R}^3$ with Lipschitz boundary $\partial\Omega$. A Y-periodic function, f, is defined as $f(\boldsymbol{x}+\hat{\boldsymbol{e}}_k)=f(\boldsymbol{x})$ for every $k=1,2,3$, where $\hat{\boldsymbol{e}}_k$, $k=1,2,3$, is the canonical basis in $\mathbb{R}^3$.

The space $C(\Omega)$ is the space of continuous functions in Ω. We also use $C_0(\overline{\Omega})$, which consists of all uniformly continuous functions which are zero at the boundary. The space $C^\infty(\Omega)$ is the space of infinitely continuously differentiable functions in Ω, and $C_0^\infty(\Omega)$ are the functions in this space with compact support in Ω, which we also denote $D(\Omega)$. Moreover,

$$C_\#^\infty(Y)=\left\{\phi\in C^\infty(\mathbb{R}^3),\phi\ Y\text{-periodic}\right\}.$$

Several function spaces with square integrable functions are used in this paper. The basic space is

$$L^2(\Omega)\stackrel{\text{def}}{=}\left\{u(\boldsymbol{x}):u\text{ Lebesgue integrable},\iiint_\Omega|u(\boldsymbol{x})|^2\,dv_{\boldsymbol{x}}<\infty\right\}$$

with norm

$$\|u\|_{L^2(\Omega)}=\left\{\iiint_\Omega|u(\boldsymbol{x})|^2\,dv_{\boldsymbol{x}}\right\}^{1/2}.$$

Similarly for vector-valued spaces we have the norm

$$\|\boldsymbol{u}\|_{L^2(\Omega;\mathbb{C}^3)}=\left\{\iiint_\Omega|\boldsymbol{u}(\boldsymbol{x})|^2\,dv_{\boldsymbol{x}}\right\}^{1/2}.$$

We also define two function spaces of periodic functions:

$$L_\#^2(Y)\stackrel{\text{def}}{=}\left\{\text{the completion of }C_\#^\infty(Y)\text{ in the }L^2(Y)\text{-norm}\right\}$$

and

$$L_\#^\infty(Y)\stackrel{\text{def}}{=}\left\{\phi\in L^\infty(\mathbb{R}^3),\phi\ Y\text{-periodic}\right\},$$

$$\begin{cases}H(\mathrm{div},\Omega)\stackrel{\text{def}}{=}\left\{\boldsymbol{u}\in L^2(\Omega;\mathbb{C}^3):\nabla\cdot\boldsymbol{u}\in L^2(\Omega)\right\},\\ H(\mathrm{rot},\Omega)\stackrel{\text{def}}{=}\left\{\boldsymbol{u}\in L^2(\Omega;\mathbb{C}^3):\nabla\times\boldsymbol{u}\in L^2(\Omega;\mathbb{C}^3)\right\},\end{cases}$$

which are Hilbert spaces with norms

$$\begin{cases}\|\boldsymbol{u}\|_{H(\mathrm{div},\Omega)}=\left(\|\boldsymbol{u}\|^2_{L^2(\Omega;\mathbb{C}^3)}+\|\nabla\cdot\boldsymbol{u}\|^2_{L^2(\Omega)}\right)^{1/2},\\ \|\boldsymbol{u}\|_{H(\mathrm{rot},\Omega)}=\left(\|\boldsymbol{u}\|^2_{L^2(\Omega;\mathbb{C}^3)}+\|\nabla\times\boldsymbol{u}\|^2_{L^2(\Omega;\mathbb{C}^3)}\right)^{1/2}.\end{cases}$$

The curl and the divergence are defined in the weak sense as

$$\begin{cases}(\nabla\times\boldsymbol{u},\boldsymbol{\phi})=(\boldsymbol{u},\nabla\times\boldsymbol{\phi})\qquad\text{for all }\boldsymbol{\phi}\in D(\Omega;\mathbb{C}^3),\\ (\nabla\cdot\boldsymbol{u},\phi)=-(\boldsymbol{u},\nabla\phi)\qquad\text{for all }\phi\in D(\Omega).\end{cases}$$

In the exterior region, we define spaces of locally integrable functions as

$$\begin{cases} H_{\rm loc}(\mathrm{div},\overline{\Omega}_e) \stackrel{\rm def}{=} \left\{\boldsymbol{u}\in D'(\Omega_e;\mathbb{C}^3) : \xi\boldsymbol{u}\in H(\mathrm{div},\Omega_e) \text{ for all } \xi\in D(\mathbb{R}^3)\right\}, \\ H_{\rm loc}(\mathrm{rot},\overline{\Omega}_e) \stackrel{\rm def}{=} \left\{\boldsymbol{u}\in D'(\Omega_e;\mathbb{C}^3) : \xi\nabla\times\boldsymbol{u}\in H(\mathrm{rot},\Omega_e) \text{ for all } \xi\in D(\mathbb{R}^3)\right\}, \end{cases}$$

where $\Omega_e = \mathbb{R}^3\setminus\overline{\Omega}$ and $D'(\Omega_e)$ is the space of distributions. The appropriate trace spaces used in this paper are $H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega)$ and $H^{-\frac{1}{2}}(\mathrm{rot},\partial\Omega)$ defined by

$$\begin{cases} H^{-\frac{1}{2}}(\mathrm{div},\partial\Omega) \stackrel{\rm def}{=} \left\{\boldsymbol{u}\in H^{-\frac{1}{2}}(\partial\Omega;\mathbb{C}^3),\ \hat{\boldsymbol{\nu}}\cdot\boldsymbol{u}=0,\ \mathrm{div}_{\partial\Omega}\boldsymbol{u}\in H^{-\frac{1}{2}}(\partial\Omega)\right\}, \\ H^{-\frac{1}{2}}(\mathrm{rot},\partial\Omega) \stackrel{\rm def}{=} \left\{\boldsymbol{u}\in H^{-\frac{1}{2}}(\partial\Omega;\mathbb{C}^3),\ \hat{\boldsymbol{\nu}}\cdot\boldsymbol{u}=0,\ \mathrm{rot}_{\partial\Omega}\boldsymbol{u}\in H^{-\frac{1}{2}}(\partial\Omega)\right\}, \end{cases}$$

where the surface divergence, $\mathrm{div}_{\partial\Omega}$, and the surface rotation, $\mathrm{rot}_{\partial\Omega}$, are defined by duality and restriction,

$$\begin{cases} (\mathrm{div}_{\partial\Omega}\boldsymbol{u},\phi) = -(\boldsymbol{u},\mathrm{grad}_{\partial\Omega}\phi) \qquad \text{for all } \phi\in D(\partial\Omega), \\ \mathrm{rot}_{\partial\Omega}\boldsymbol{u} = \hat{\boldsymbol{\nu}}\cdot(\nabla\times\boldsymbol{u})|_{\partial\Omega}, \end{cases}$$

and the surface gradient, $\mathrm{grad}_{\partial\Omega}$, is defined by the orthogonal projection of ∇ on the surface $\partial\Omega$.

We also define the function spaces

$$\begin{cases} H_{\#}(\mathrm{div},Y) \stackrel{\rm def}{=} \{\boldsymbol{u}\in H(\mathrm{div},Y), \boldsymbol{u}\ Y\text{-periodic}\}, \\ H_{\#}(\mathrm{rot},Y) \stackrel{\rm def}{=} \{\boldsymbol{u}\in H(\mathrm{rot},Y), \boldsymbol{u}\ Y\text{-periodic}\} \end{cases}$$

and

$$\begin{cases} H^1_{\#}(Y) \stackrel{\rm def}{=} \left\{\text{the completion of } C^\infty_{\#}(Y) \text{ in the } H^1(Y)\text{-norm}\right\}, \\ H^1_{\#}(Y)/\mathbb{C} \stackrel{\rm def}{=} \left\{\phi\in H^1_{\#}(Y),\ \text{equivalent up to a complex constant}\right\}. \end{cases}$$

If γ denotes the unit sphere in $\mathbb{R}^3$, the following norms are used in the paper:

$$\begin{cases} \|\boldsymbol{u}\|_\gamma = \left\{\iint_\gamma |\boldsymbol{u}(\hat{\boldsymbol{x}})|^2\ d\gamma\right\}^{1/2}, \\ \|\boldsymbol{u}\|_\infty = \sup_{|\hat{\boldsymbol{x}}|=1}|\boldsymbol{u}(\hat{\boldsymbol{x}})|, \end{cases}$$

and $d\gamma$ denotes the surface measure on the unit sphere in $\mathbb{R}^3$.

We conclude this appendix by stating the Lax–Milgram theorem [13].

THEOREM A.1 (Lax–Milgram). *Assume that H is a Hilbert space with norm $\|\cdot\|$. Moreover, assume that*

$$B : H\times H\to\mathbb{C}$$

is a sesquilinear functional on H, for which there exists constants $a,b>0$ such that

$$|B[u,v]|\le a\|u\|\|v\| \qquad \textit{for all } u,v\in H$$

and

$$b\|u\|^2\le|B[u,u]| \qquad \textit{for all } u\in H.$$

Finally, let $f : H\to\mathbb{C}$ be a bounded linear functional on H.

Then there exists a unique $u \in H$ such that

$$B[u,v] = f(v) \quad \textit{for all } v \in H.$$

Appendix B. Two-scale convergence.

DEFINITION B.1. *A sequence $\{\boldsymbol{u}^\varepsilon\}$ in $L^2(\Omega;\mathbb{C}^3)$ two-scale converges to $\boldsymbol{u}_0 \in L^2(\Omega \times Y;\mathbb{C}^3)$ if*

$$\lim_{\varepsilon \searrow 0} \iiint_\Omega \boldsymbol{u}^\varepsilon(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} = \iiint_\Omega \iiint_Y \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y}) \cdot \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

for every $\boldsymbol{\phi} \in D(\Omega; C^\infty_\#(Y;\mathbb{C}^3))$. We denote this by $\boldsymbol{u}^\varepsilon \overset{2\text{-}s}{\rightharpoonup} \boldsymbol{u}_0$.

The class of test functions can be enlarged to all admissible test functions defined below [2].

DEFINITION B.2. *We say that $\boldsymbol{\phi} \in L^2(\Omega; L^2_\#(Y;\mathbb{C}^3))$ is an admissible test function if $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)$ is measurable and*

$$\lim_{\varepsilon \searrow 0} \|\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)\|_{L^2(\Omega;\mathbb{C}^3)} = \|\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y})\|_{L^2(\Omega \times Y;\mathbb{C}^3)}.$$

Remark B.1. Some examples of admissible test functions are $L^2(\Omega; C_\#(Y;\mathbb{C}^3))$ and for Ω bounded $L^2_\#(Y; C(\overline{\Omega};\mathbb{C}^3))$.

We cite two important theorems by Nguetseng [19].

THEOREM B.3 (Nguetseng [19]). *Let $u^\varepsilon \in L^2(\Omega)$. Suppose that there exists a constant $C > 0$ such that*

$$\|u^\varepsilon\|_{L^2(\Omega)} \le C \textit{ for all } \varepsilon.$$

Then a subsequence (still denoted by ε) can be extracted from ε such that, letting $\varepsilon \searrow 0$,

$$\iiint_\Omega u^\varepsilon(\boldsymbol{x}) \Psi(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} \to \iiint_\Omega \iiint_Y u_0(\boldsymbol{x},\boldsymbol{y}) \Psi(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

for all $\Psi \in C_0(\overline{\Omega}; C_\#(Y))$, where $u_0 \in L^2(\Omega; L^2_\#(Y))$. Moreover,

$$\iiint_\Omega u^\varepsilon(\boldsymbol{x}) v(\boldsymbol{x}) w(\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} \to \iiint_\Omega \iiint_Y u_0(\boldsymbol{x},\boldsymbol{y}) v(\boldsymbol{x}) w(\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

for all $v \in C_0(\overline{\Omega})$ and all $w \in L^2_\#(Y)$.

We note that if u^ε is a sequence in $L^2(\Omega)$, which two-scale converges to the limit $u_0 \in L^2(\Omega \times Y)$, then u^ε also converges to $u(\boldsymbol{x}) = \iiint_Y u_0(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}$ in $L^2(\Omega)$ weakly [2]. Moreover, if u^ε converges strongly to $u(\boldsymbol{x})$ in $L^2(\Omega)$, then u^ε two-scale converges to the same limit $u(\boldsymbol{x})$. The second theorem is the following.

THEOREM B.4 (Nguetseng). *Let $u^\varepsilon \in H^1(\Omega)$. Suppose that there exists a constant $C > 0$ such that*

$$\|u^\varepsilon\|_{H^1(\Omega)} \le C \textit{ for all } \varepsilon.$$

Then a subsequence (still denoted by ε) can be extracted from ε such that, letting $\varepsilon \searrow 0$,

$$u^\varepsilon \to u \textit{ in } H^1(\Omega)\textit{-weak}$$

and

$$\iiint_\Omega \frac{\partial u^\varepsilon(\boldsymbol{x})}{\partial x_j} v(\boldsymbol{x}) w(\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} \\ \to \iiint_\Omega \iiint_Y \left\{ \frac{\partial u(\boldsymbol{x})}{\partial x_j} + \frac{\partial u_1(\boldsymbol{x},\boldsymbol{y})}{\partial y_j} \right\} v(\boldsymbol{x}) w(\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}},$$

$j = 1, 2, 3$, *for all* $v \in C_0(\overline{\Omega})$ *and all* $w \in L^2_\#(Y)$, *where* $u_1 \in L^2(\Omega; H^1_\#(Y)/\mathbb{C})$.

In addition to these two theorems, we observe that, taking $w = 1$, we get from Theorem B.3

$$\iiint_\Omega u^\varepsilon(\boldsymbol{x}) v(\boldsymbol{x})\, dv_{\boldsymbol{x}} \to \iiint_\Omega u(\boldsymbol{x}) v(\boldsymbol{x})\, dv_{\boldsymbol{x}}$$

for all $v \in C_0(\overline{\Omega})$, where

$$u(\boldsymbol{x}) = \iiint_Y u_0(\boldsymbol{x}, \boldsymbol{y})\, dv_{\boldsymbol{y}}$$

is the usual weak $L^2(\Omega)$-limit of $u^\varepsilon(\boldsymbol{x})$. It follows that u_0 is uniquely expressed in the form

$$u_0(\boldsymbol{x}, \boldsymbol{y}) = u(\boldsymbol{x}) + \tilde{u}_0(\boldsymbol{x}, \boldsymbol{y}),$$

where

$$\iiint_Y \tilde{u}_0(\boldsymbol{x}, \boldsymbol{y})\, dv_{\boldsymbol{y}} = 0.$$

LEMMA B.5. *Let* $\boldsymbol{f} \in H^1_\#(Y; \mathbb{C}^3)$ *and assume that* $\nabla_{\boldsymbol{y}} \times \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{0}$. *Moreover, assume that* $\langle \boldsymbol{f} \rangle = 0$. *Then there exists a unique function* $q \in H^1_\#(Y)/\mathbb{C}$ *such that*

$$\boldsymbol{f}(\boldsymbol{y}) = \nabla_{\boldsymbol{y}} q(\boldsymbol{y}).$$

Proof of Lemma B.5. The periodicity of the function $\boldsymbol{f} \in H^1_\#(Y; \mathbb{C}^3)$ implies that $\boldsymbol{f}$ has a Fourier expansion

$$\boldsymbol{f}(\boldsymbol{y}) = \sum_n \boldsymbol{f}_n e^{i\boldsymbol{k}_n \cdot \boldsymbol{y}},$$

where the vector $\boldsymbol{k}_n$ is defined as

$$\boldsymbol{k}_n = 2\pi n_1 \hat{\boldsymbol{e}}_1 + 2\pi n_2 \hat{\boldsymbol{e}}_2 + 2\pi n_3 \hat{\boldsymbol{e}}_3$$

and where n_1, n_2, n_3 are integers and $n = (n_1, n_2, n_3)$. The sequence $\boldsymbol{f}_n$ belongs to $\left(\ell^2_1\right)^3$. The assumption that $\langle \boldsymbol{f} \rangle = 0$ implies that $\boldsymbol{f}_{(0,0,0)} = 0$. Moreover, the coefficients $\boldsymbol{f}_n$ satisfy

$$\boldsymbol{k}_n \times \boldsymbol{f}_n = \boldsymbol{0} \quad \text{for all } n.$$

Therefore $\boldsymbol{f}_n$ has the form

$$\boldsymbol{f}_n = \hat{\boldsymbol{k}}_n \left(\hat{\boldsymbol{k}}_n \cdot \boldsymbol{f}_n \right).$$

Define q_n as

$$\begin{cases} q_n = -i(\hat{\boldsymbol{k}}_n \cdot \boldsymbol{f}_n)/k_n \text{ for } n \neq (0,0,0), \\ q_{(0,0,0)} \text{ arbitrary}, \end{cases}$$

where $k_n = |\boldsymbol{k}_n|$. The coefficients $q_n \in \left(\ell_1^2\right)^3$,

$$\boldsymbol{f}_n = i\boldsymbol{k}_n q_n \quad \text{for all } n,$$

and

$$\boldsymbol{f}(\boldsymbol{y}) = \sum_n i\boldsymbol{k}_n q_n e^{i\boldsymbol{k}_n \cdot \boldsymbol{y}} = \nabla_{\boldsymbol{y}} q(\boldsymbol{y}),$$

where

$$q(\boldsymbol{y}) = \sum_n q_n e^{i\boldsymbol{k}_n \cdot \boldsymbol{y}} \in H^1_{\#}(Y)/\mathbb{C},$$

since $q_{(0,0,0)}$ is arbitrary and the lemma is proved. ∎

The obvious vector analogous theorems follow.

THEOREM B.6. *Let $\boldsymbol{u}^\varepsilon \in L^2(\Omega;\mathbb{C}^3)$. Suppose that there exists a constant $C > 0$ such that*

$$\|\boldsymbol{u}^\varepsilon\|_{L^2(\Omega;\mathbb{C}^3)} \leq C \text{ for all } \varepsilon.$$

Then a subsequence (still denoted by ε) can be extracted from ε such that, letting $\varepsilon \searrow 0$,

$$\iiint_\Omega \boldsymbol{u}^\varepsilon(\boldsymbol{x}) \cdot \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} \to \iiint_\Omega \iiint_Y \boldsymbol{u}_0(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

for all $\boldsymbol{\Psi} \in C_0(\overline{\Omega}; C_{\#}(Y;\mathbb{C}^3))$, where $\boldsymbol{u}_0 \in L^2(\Omega; L^2_{\#}(Y;\mathbb{C}^3))$. Moreover,

$$\iiint_\Omega \boldsymbol{u}^\varepsilon(\boldsymbol{x}) \cdot \boldsymbol{v}(\boldsymbol{x}) w(\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} \to \iiint_\Omega \iiint_Y \boldsymbol{u}_0(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{v}(\boldsymbol{x}) w(\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

for all $\boldsymbol{v} \in C_0(\overline{\Omega};\mathbb{C}^3)$ and all $w \in L^2_{\#}(Y)$.

The field $\boldsymbol{u}_0$ is uniquely expressed in the form

$$\boldsymbol{u}_0(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{u}(\boldsymbol{x}) + \tilde{\boldsymbol{u}}_0(\boldsymbol{x}, \boldsymbol{y}),$$

where

$$\iiint_Y \tilde{\boldsymbol{u}}_0(\boldsymbol{x}, \boldsymbol{y})\, dv_{\boldsymbol{y}} = \boldsymbol{0}.$$

We have the following results proved in [23].

THEOREM B.7. *Let $\boldsymbol{u}^\varepsilon \in H(\mathrm{div}, \Omega)$. Suppose that there exists a constant $C > 0$ such that*

$$\|\boldsymbol{u}^\varepsilon\|_{H(\mathrm{div},\Omega)} \leq C \quad \text{for all } \varepsilon.$$

Then a subsequence (still denoted by ε) can be extracted from ε such that, letting $\varepsilon \searrow 0$,

$$\boldsymbol{u}^\varepsilon \to \boldsymbol{u} \text{ in } L^2(\Omega;\mathbb{C}^3)\text{-weak}$$

and

$$\iiint_\Omega \nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}^\varepsilon(\boldsymbol{x}) v(\boldsymbol{x}) w(\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}}$$
$$\to \iiint_\Omega \iiint_Y \left\{\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}} \cdot \boldsymbol{u}_1(\boldsymbol{x},\boldsymbol{y})\right\} v(\boldsymbol{x}) w(\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

for all $v \in C_0(\overline{\Omega})$ *and all* $w \in L^2_\#(Y)$, *where* $\boldsymbol{u}(\boldsymbol{x}) = \iiint_Y \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}$, $\boldsymbol{u}_0$ *is the two-scale limit of* $\boldsymbol{u}^\varepsilon$, *and* $\boldsymbol{u}_1 \in L^2(\Omega; H_\#(\mathrm{div}, Y))$.

THEOREM B.8. *Let* $\boldsymbol{u}^\varepsilon \in H(\mathrm{rot}, \Omega)$. *Suppose that there exists a constant* $C > 0$ *such that*

$$\|\boldsymbol{u}^\varepsilon\|_{H(\mathrm{rot},\Omega)} \le C \quad \textit{for all } \varepsilon.$$

Then a subsequence (still denoted by ε*) can be extracted from* ε *such that, letting* $\varepsilon \searrow 0$,

$$\boldsymbol{u}^\varepsilon \to \boldsymbol{u}_0 \textit{ in } L^2(\Omega;\mathbb{C}^3)\textit{-weak}$$

and

$$\iiint_\Omega \nabla \times \boldsymbol{u}^\varepsilon(\boldsymbol{x}) \cdot v(\boldsymbol{x}) \boldsymbol{w}(\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}}$$
$$\to \iiint_\Omega \iiint_Y \left\{\nabla_{\boldsymbol{x}} \times \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y}) + \nabla_{\boldsymbol{y}} \times \boldsymbol{u}_1(\boldsymbol{x},\boldsymbol{y})\right\} \cdot v(\boldsymbol{x}) \boldsymbol{w}(\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

for all $v \in C_0(\overline{\Omega})$ *and all* $\boldsymbol{w} \in L^2_\#(Y;\mathbb{C}^3)$, *where* $\boldsymbol{u}_1 \in L^2(\Omega; H_\#(\mathrm{rot}, Y))$.

Proof of Theorem B.8. From Theorem B.6 we get

$$\iiint_\Omega \boldsymbol{u}^\varepsilon(\boldsymbol{x}) \cdot \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} \to \iiint_\Omega \iiint_Y \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y}) \cdot \boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

and

$$\iiint_\Omega \nabla \times \boldsymbol{u}^\varepsilon(\boldsymbol{x}) \cdot \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} \to \iiint_\Omega \iiint_Y \boldsymbol{\chi}_0(\boldsymbol{x},\boldsymbol{y}) \cdot \boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

for all $\boldsymbol{\Psi} \in C_0(\overline{\Omega}; C_\#(Y;\mathbb{C}^3))$, where $\boldsymbol{u}_0, \boldsymbol{\chi}_0 \in L^2(\Omega; L^2_\#(Y;\mathbb{C}^3))$. Choose test functions $\boldsymbol{\Psi} \in C_0(\overline{\Omega}; C_\#(Y;\mathbb{C}^3))$ such that $\nabla_{\boldsymbol{y}} \times \boldsymbol{\Psi} = 0$. We get by integration by parts

$$\iiint_\Omega \nabla \times \boldsymbol{u}^\varepsilon(\boldsymbol{x}) \cdot \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} = \iiint_\Omega \boldsymbol{u}^\varepsilon(\boldsymbol{x}) \cdot \nabla \times \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}}$$
$$= \iiint_\Omega \boldsymbol{u}^\varepsilon(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} \times \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}}$$
$$\to \iiint_\Omega \iiint_Y \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y}) \cdot \nabla_{\boldsymbol{x}} \times \boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$
$$= \iiint_\Omega \iiint_Y \nabla_{\boldsymbol{x}} \times \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y}) \cdot \boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}.$$

This means that

$$\iiint_\Omega \iiint_Y (\boldsymbol{\chi}_0(\boldsymbol{x},\boldsymbol{y}) - \nabla_{\boldsymbol{x}} \times \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y})) \cdot \boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}} = 0$$

for all $\boldsymbol{\Psi} \in C_0(\overline{\Omega}; C_{\#}(Y;\mathbb{C}^3))$ such that $\nabla_{\boldsymbol{y}} \times \boldsymbol{\Psi} = 0$. By the decomposition of $L^2(\Omega;\mathbb{C}^3)$ (e.g., see [9]) there exists a function $\boldsymbol{u}_1 \in L^2(\Omega; H_{\#}(\mathrm{rot}, Y))$ such that

$$\nabla_{\boldsymbol{y}} \times \boldsymbol{u}_1 = \boldsymbol{\chi}_0(\boldsymbol{x},\boldsymbol{y}) - \nabla_{\boldsymbol{x}} \times \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y}). \qquad \square$$

THEOREM B.9 (see Wellander [26] or [27]). *Let $\boldsymbol{u}^\varepsilon \in H(\mathrm{rot},\Omega)$. Suppose that there exists a constant $C > 0$ such that*

$$\|\boldsymbol{u}^\varepsilon\|_{H(\mathrm{rot},\Omega)} \le C \quad \text{for all } \varepsilon.$$

Then a subsequence (still denoted by ε) can be extracted from ε such that, letting $\varepsilon \searrow 0$,

$$\boldsymbol{u}^\varepsilon \overset{2\text{-}s}{\rightharpoonup} \boldsymbol{u}(\boldsymbol{x}) + \nabla_{\boldsymbol{y}}\phi(\boldsymbol{x},\boldsymbol{y}),$$

where $\phi \in L^2(\Omega; H^1_{\#}(Y))$ is a scalar-valued function satisfying

$$\iiint_Y \nabla_{\boldsymbol{y}}\phi(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}} = \boldsymbol{0}.$$

Moreover,

$$\nabla \times \boldsymbol{u}^\varepsilon \rightharpoonup \nabla \times \boldsymbol{u}(\boldsymbol{x}) \text{ in } L^2(\Omega;\mathbb{C}^3).$$

THEOREM B.10 (see Wellander [26] or [27]). *Let $\boldsymbol{u}^\varepsilon \in H(\mathrm{div},\Omega)$. Suppose that there exists a constant $C > 0$ such that*

$$\|\boldsymbol{u}^\varepsilon\|_{H(\mathrm{div},\Omega)} \le C \quad \text{for all } \varepsilon.$$

Then a subsequence (still denoted by ε) can be extracted from ε such that, letting $\varepsilon \searrow 0$,

$$\boldsymbol{u}^\varepsilon \overset{2\text{-}s}{\rightharpoonup} \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y})$$

and

$$\varepsilon\nabla\cdot\boldsymbol{u}^\varepsilon \overset{2\text{-}s}{\rightharpoonup} \nabla_{\boldsymbol{y}}\cdot\boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y}).$$

Proof of Theorem B.10. From Theorem B.6 we get

$$\iiint_\Omega \boldsymbol{u}^\varepsilon(\boldsymbol{x})\cdot\boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} \to \iiint_\Omega \iiint_Y \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y})\cdot\boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

and

$$\iiint_\Omega \nabla\cdot\boldsymbol{u}^\varepsilon(\boldsymbol{x})\Psi(\boldsymbol{x},\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} \to \iiint_\Omega \iiint_Y \chi_0(\boldsymbol{x},\boldsymbol{y})\Psi(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}$$

for all $\boldsymbol{\Psi} \in C_0(\overline{\Omega}; C_{\#}(Y;\mathbb{C}^3))$ and $\Psi \in C_0(\overline{\Omega}; C_{\#}(Y))$, where $\boldsymbol{u}_0 \in L^2(\Omega; L^2_{\#}(Y;\mathbb{C}^3))$ and $\chi_0 \in L^2(\Omega; L^2_{\#}(Y))$.

We get by integration by parts

$$\begin{aligned}
&\iiint_\Omega \varepsilon\nabla\cdot\boldsymbol{u}^\varepsilon(\boldsymbol{x})\Psi(\boldsymbol{x},\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} = -\iiint_\Omega \varepsilon\boldsymbol{u}^\varepsilon(\boldsymbol{x})\cdot\nabla\Psi(\boldsymbol{x},\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}}\\
&\quad = -\iiint_\Omega \varepsilon\boldsymbol{u}^\varepsilon(\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\Psi(\boldsymbol{x},\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}} - \iiint_\Omega \boldsymbol{u}^\varepsilon(\boldsymbol{x})\cdot\nabla_{\boldsymbol{y}}\Psi(\boldsymbol{x},\boldsymbol{x}/\varepsilon)\, dv_{\boldsymbol{x}}\\
&\quad \to -\iiint_\Omega \iiint_Y \boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}\Psi(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}\\
&\quad = \iiint_\Omega \iiint_Y \nabla_{\boldsymbol{y}}\cdot\boldsymbol{u}_0(\boldsymbol{x},\boldsymbol{y})\Psi(\boldsymbol{x},\boldsymbol{y})\, dv_{\boldsymbol{y}}\, dv_{\boldsymbol{x}}. \qquad \square
\end{aligned}$$

Appendix C. Vector spherical harmonics. The vector spherical harmonics are defined as (see [8])

$$\begin{cases} \boldsymbol{A}_{1n}(\hat{\boldsymbol{x}}) = \frac{1}{\sqrt{l(l+1)}}\nabla \times (\boldsymbol{x} Y_n(\hat{\boldsymbol{x}})) = \frac{1}{\sqrt{l(l+1)}}\nabla Y_n(\hat{\boldsymbol{x}}) \times \boldsymbol{x}, \\ \boldsymbol{A}_{2n}(\hat{\boldsymbol{x}}) = \frac{1}{\sqrt{l(l+1)}} x \nabla Y_n(\hat{\boldsymbol{x}}), \\ \boldsymbol{A}_{3n}(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}} Y_n(\hat{\boldsymbol{x}}), \end{cases}$$

where the spherical harmonics are denoted $Y_n(\hat{\boldsymbol{x}})$. The index n is a multi-index for the integer indices $l = 0, 1, 2, 3, \ldots$, $m = 0, 1, \ldots, l$, and $\sigma =$ e,o (even and odd in the azimuthal angle). From these definitions we see that the first two vector spherical harmonics, $\boldsymbol{A}_{1n}(\hat{\boldsymbol{x}})$ and $\boldsymbol{A}_{2n}(\hat{\boldsymbol{x}})$, are tangential to the unit sphere γ in $\mathbb{R}^3$ and they are related by

$$\begin{cases} \hat{\boldsymbol{x}} \times \boldsymbol{A}_{1n}(\hat{\boldsymbol{x}}) = \boldsymbol{A}_{2n}(\hat{\boldsymbol{x}}), \\ \hat{\boldsymbol{x}} \times \boldsymbol{A}_{2n}(\hat{\boldsymbol{x}}) = -\boldsymbol{A}_{1n}(\hat{\boldsymbol{x}}). \end{cases}$$

The vector spherical harmonics form an orthonormal set over the unit sphere γ in $\mathbb{R}^3$, i.e.,

$$\iint_\gamma \boldsymbol{A}_{\tau n}(\hat{\boldsymbol{x}}) \cdot \boldsymbol{A}_{\tau' n'}(\hat{\boldsymbol{x}})\, d\gamma = \delta_{nn'}\delta_{\tau\tau'}. \tag{C.1}$$

The radiating solutions to the Maxwell equations in a vacuum are defined as

$$\begin{cases} \boldsymbol{u}_{1n}(k_0\boldsymbol{x}) = h_l^{(1)}(k_0 x)\boldsymbol{A}_{1n}(\hat{\boldsymbol{x}}), \\ \boldsymbol{u}_{2n}(k_0\boldsymbol{x}) = \frac{1}{k_0}\nabla \times \left(h_l^{(1)}(k_0 x)\boldsymbol{A}_{1n}(\hat{\boldsymbol{x}}) \right), \end{cases}$$

where $h_l^{(1)}(k_0 x)$ is the spherical Hankel function of the first kind [1]. These vector waves satisfy

$$\nabla \times (\nabla \times \boldsymbol{u}_{\tau n}(k_0\boldsymbol{x})) - k_0^2 \boldsymbol{u}_{\tau n}(k_0\boldsymbol{x}) = \boldsymbol{0}, \qquad \tau = 1, 2, \tag{C.2}$$

and they also satisfy the radiation condition in (2.5). Another representation of the definition of the vector waves is

$$\begin{cases} \boldsymbol{u}_{1n}(k_0\boldsymbol{x}) = h_l^{(1)}(k_0 x)\boldsymbol{A}_{1n}(\hat{\boldsymbol{x}}), \\ \boldsymbol{u}_{2n}(k_0\boldsymbol{x}) = \frac{(k_0 x h_l^{(1)}(k_0 x))'}{k_0 x}\boldsymbol{A}_{2n}(\hat{\boldsymbol{x}}) + \sqrt{l(l+1)}\frac{h_l^{(1)}(k_0 x)}{k_0 x}\boldsymbol{A}_{3n}(\hat{\boldsymbol{x}}), \end{cases}$$

where $'$ denotes differentiation with respect to the argument of the spherical Hankel function. A simple consequence of these definitions is

$$\begin{cases} \boldsymbol{u}_{1n}(k_0\boldsymbol{x}) = \frac{1}{k_0}\nabla \times \boldsymbol{u}_{2n}(k_0\boldsymbol{x}), \\ \boldsymbol{u}_{2n}(k_0\boldsymbol{x}) = \frac{1}{k_0}\nabla \times \boldsymbol{u}_{1n}(k_0\boldsymbol{x}). \end{cases} \tag{C.3}$$

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Applied Mathematics Series 55, National Bureau of Standards, Washington, DC, 1970.

[2] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[3] Y. Amirat, K. Hamdache, and A. Ziani, *Homogenization of degenerate wave equations with periodic coefficients*, SIAM J. Math. Anal., 24 (1993), pp. 1226–1253.
[4] M. Artola, *Homogenization and electromagnetic wave propagation in composite media with high conductivity inclusions*, in Proceedings of the Second Workshop on Composite Media and Homogenization Theory, G. Dal Maso and G. Dell'Antonio, eds., World Scientific, Singapore, 1995, pp. 1–15.
[5] H. Attouch, *Variational Convergence of Functions and Operators*, Pitman, London, 1984.
[6] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North–Holland, Amsterdam, 1978.
[7] A. Bossavit, *On the homogenization of Maxwell equations*, COMPEL, 14 (1995), pp. 23–26.
[8] A. Boström, G. Kristensson, and S. Ström, *Transformation properties of plane, spherical and cylindrical scalar and vector wave functions*, in Field Representations and Introduction to Scattering, V. V. Varadan, A. Lakhtakia, and V. K. Varadan, eds., Acoustic, Electromagnetic and Elastic Wave Scattering 1, North–Holland, Amsterdam, 1991, pp. 165–210.
[9] M. Cessenat, *Mathematical Methods in Electromagnetism*, Ser. Adv. Math. Appl. Sci. 41, World Scientific, Singapore, 1996.
[10] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford University Press, Oxford, 1999.
[11] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1992.
[12] C. Conca and M. Vanninathan, *Homogenization of periodic structures via Bloch decomposition*, SIAM J. Appl. Math., 57 (1997), pp. 1639–1659.
[13] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.
[14] A. Holmbom, *Homogenization of parabolic equations: An alternative approach and some corrector-type results*, Appl. Math., 42 (1997), pp. 321–343.
[15] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.
[16] G. Kristensson, S. Poulsen, and S. Rikte, *Propagators and Scattering of Electromagnetic waves in Planar Bianisotropic Slabs—An Application to Frequency Selective Structures*, Technical report LUTEDX/(TEAT-7099)/1–32/(2001), Lund Institute of Technology, Department of Electroscience, Lund, Sweden, 2001.
[17] D. Lukkassen, G. Nguetseng, and P. Wall, *Two scale convergence*, J. Pure Appl. Math., 2 (2002), pp. 35–86.
[18] P. A. Markowich and F. Poupaud, *The Maxwell equation in a periodic medium: Homogenization of the energy density*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 23 (1996), pp. 301–324.
[19] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.
[20] E. Sanchez-Palencia, *Non-Homogeneous Media and Vibration Theory*, Lecture Notes in Phys. 127, Springer-Verlag, Berlin, 1980.
[21] S. Shkoller and G. Hegemier, *Homogenization of plane weave composites using two-scale convergence*, Internat. J. Solids Structures, 32 (1995), pp. 783–794.
[22] A. Sihvola, *Electromagnetic Mixing Formulae and Applications*, IEE Electromagnet. Waves Ser. 47, IEE, London, 1999.
[23] N. Svanstedt and N. Wellander, *A Note on Two-Scale Limits of Differential Operators*, Technical report 19, Department of Mathematics, Chalmers University of Technology, Göteborg, Sweden, 2001.
[24] L. Tartar, *Cours peccot au Collège de France*, manuscript, 1977.
[25] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. 4, Res. Notes in Math. 39, Pitman, London, 1979, pp. 136–212.
[26] N. Wellander, *Homogenization of Some Linear and Nonlinear Partial Differential Equations*, Ph.D. thesis, Luleå University of Technology, Luleå, Sweden, 1998.
[27] N. Wellander, *Homogenization of the Maxwell equations: Case* I. *Linear theory*, Appl. Math., 46 (2001), pp. 29–51.
[28] N. Wellander, *Homogenization of the Maxwell equations: Case* II. *Nonlinear conductivity*, Appl. Math., 47 (2002), pp. 255–283.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 196–215

A HOMOGENIZATION TECHNIQUE FOR THE BOLTZMANN EQUATION FOR LOW PRESSURE CHEMICAL VAPOR DEPOSITION*

MATTHIAS K. GOBBERT† AND CHRISTIAN RINGHOFER‡

Abstract. We present a homogenization technique for rarefied gas flow over a microstructured surface consisting of patterns of periodic features. The length scale of the model domain is comparable to the mean free path of the molecules, while the scale of the surface patterns is much smaller. The flow is modeled by a system of linear Boltzmann equations with a diffusive boundary condition at the patterned surface. The resulting homogenized boundary condition holds at a virtual flat surface and incorporates the microscopic geometry information about the surface structure on the macroscopic level. Numerical results validate the approach. The setup models low pressure chemical vapor deposition processes in the manufacturing of integrated circuits.

Key words. Boltzmann equation, rarefied gas dynamics, boundary homogenization, microstructured surface, chemical vapor deposition

AMS subject classifications. 65M06, 76P05, 35B27, 76M45, 80A30

DOI. 10.1137/S0036139902393476

1. Introduction.

Low pressure chemical vapor deposition is used in the manufacturing of integrated circuits to deposit a thin layer of material onto the surface of a silicon wafer. The deposition surface necessarily involves a microstructure given by the electrical components of the future microchip. Classical models for this process include reactor scale models [17] with a typical length scale of over 10 cm, which model the gas flow throughout the chemical reactor, and feature scale models [3] with a typical length scale of under 1 μm, which focus on the evolution of the film profile inside an individual feature.

In more detail, the process works as follows. Molecules of the species to be deposited are carried inside the chemical reactor by an inert carrier gas to a microstructured surface, where they are partially absorbed and partially reflected at a certain rate. The length scale of the surface structure is several orders of magnitude smaller than that of the reactor and therefore cannot be reasonably resolved on the reactor scale. On the other hand, this structure will influence the gas flow through the boundary conditions; i.e., adsorption on the microstructured surface will result in a different behavior of the gas flow than adsorption on a flat surface, even on the macroscopic reactor scale. We therefore have to solve a homogenization problem at the boundary by deriving a boundary condition for the flow problem which incorporates the microscopic geometric information about the surface structure into the macroscopic flow picture.

*Received by the editors October 11, 2002; accepted for publication (in revised form) April 15, 2003; published electronically November 19, 2003. This research was supported in part by the Institute for Mathematics and its Applications, with funds provided by the National Science Foundation. This research was also supported by the Wittgenstein Award 2000 of Peter Markowich, financed by the Austrian Research Fund.

http://www.siam.org/journals/siap/64-1/39347.html

†Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250 (gobbert@math.umbc.edu). The research of this author was supported in part by the National Science Foundation under grant DMS-9805547.

‡Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804 (ringhofer@asu.edu). The research of this author was supported in part by the National Science Foundation under grant DMS-9706792.

Based on analytic work in [2, 5, 6, 8, 11], the authors and coworkers have previously introduced [7, 12] a mesoscopic scale model on a length scale intermediate to those of the classical reactor scale and feature scale models, which was designed to provide information on the effects of feature clustering on a length scale of about 1.0 cm at comparably high pressures of at least 1 torr (where 760 torr = 1 atm). In that regime, the mean free path of the gas molecules was well inside the domain size, and the gas flow could be modeled as diffusion-dominated [8, 9, 11, 12, 18]. While [8, 11] derive the homogenized model formally for the concrete problem of interest, [2, 5, 6] handle the analysis rigorously for a more general class of models.

However, as individual feature sizes decrease below 1 μm, the length scale of interest for clustering effects also decreases. Therefore, this work considers a mesoscopic scale model with a domain with typical length scale on the order of 0.01 cm. Using this together with typical values for the total pressure of 1 torr or less, the mean free path of the gas molecules is of length comparable to the typical length scale. The Knudsen number Kn, which is defined as the ratio of the mean free path and the typical length scale, is on the order of unity, and the process lies in the transition regime for gaseous flow modeling [15].

The proper mathematical model for a gas flow in the transition regime is given by the Boltzmann equation of gas dynamics for the (scaled) probability density $f(x, v, t)$ that there is a molecule in the region $[x_1, x_1 + dx_1] \times [x_2, x_2 + dx_2] \times [x_3, x_3 + dx_3]$ with velocity in $[v_1, v_1 + dv_1] \times [v_2, v_2 + dv_2] \times [v_3, v_3 + dv_3]$ during time $[t, t + dt]$,

$$\frac{\partial f}{\partial t} + v \cdot \nabla f = Q(f, f), \tag{1.1}$$

with the collision operator

$$Q(f,f)(x,v,t) = \iiint \big[f(x,v',t)f(x,v'_*,t) - f(x,v,t)f(x,v_*,t)\big] B(\vartheta, |V|)\, d\vartheta\, d\varepsilon\, dv_*, \tag{1.2}$$

where $v' = v - n(n \cdot V)$ and $v'_* = v_* + n(n \cdot V)$ denote the precollision velocities, with $V = v - v_*$ and $n = (\sin\vartheta\cos\varepsilon, \sin\vartheta\sin\varepsilon, \cos\vartheta)^T$ (see [4, 15]).

A complete model for chemical vapor deposition will consist of one Boltzmann equation for each gaseous species $f_i(x, v, t)$, $i = 0, 1, \ldots, n_s$, to form the system

$$\frac{\partial f_i}{\partial t} + v \cdot \nabla_x f_i = \sum_{j=0}^{n_s} Q_{ij}(f_i, f_j), \quad i = 0, 1, \ldots, n_s, \tag{1.3}$$

where the collision operators $Q_{ij}(f_i, f_j)$ model the collisions between molecules of species i and species j for $0 \le i, j \le n_s$. This model includes the inert background gas $f_0(x, v, t)$ of the manufacturing process, which is a rarefied gas itself but still much denser than the reacting species. Under these assumptions, the collisions of a reacting species with the background gas will be much more frequent than collisions *among* the reacting species, and it is legitimate to neglect all collisions except the ones with the background species $j = 0$ on the right-hand side of (1.3); this also decouples the equation for the background species f_0 from the other equations, which can hence be solved for independently from the other solutions f_i, $i = 1, \ldots, n_s$.

In the classical derivation [4, Chapter IV], it is additionally assumed that the background gas is in equilibrium and spatially homogeneously distributed. Then its

distribution function $f_0(x, v, t)$ is given by a Maxwellian distribution of the form

$$f_0(x, v, t) = M(v) := \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{|v|^2}{2}\right), \tag{1.4}$$

where $d \in \{1, 2, 3\}$ denotes the dimension of the velocity space under consideration. Then, the time-evolution of the probability distribution of a typical reacting species is given by the system of linear Boltzmann equations

$$\frac{\partial f_i}{\partial t} + v \cdot \nabla f_i = Q_i(f_i), \quad i = 1, \ldots, n_s, \tag{1.5}$$

with the linear collision operators

$$Q_i(f_i)(x, v, t) = \int S_i(v, v') \left[\frac{f_i(x, v', t)}{M(v')} - \frac{f_i(x, v, t)}{M(v)}\right] dv'; \tag{1.6}$$

see [4, Chapter IV] for a detailed derivation. Here, $S_i(v, v') = S_i(v', v)$ denotes the scattering cross section, which describes the probability that a molecule with velocity v' before a collision scatters to a velocity v after the collision. A generalization to spatially varying background gases is possible. We will assume that the reacting species are introduced into the reactor chamber starting at the beginning of the processing step, i.e., that $f_i = 0$, $i = 1, \ldots, n_s$, at $t = 0$.

For the analysis, we therefore consider the representative equation for $f(x, y, v, t)$:

$$\frac{\partial f}{\partial t} + v_1 \frac{\partial f}{\partial x_1} + v_2 \frac{\partial f}{\partial x_2} + v_3 \frac{\partial f}{\partial y} = Q(f). \tag{1.7}$$

Here, $x \equiv (x_1, x_2)^T$ counts across the reacting surface, and $y \equiv x_3$ points into the gaseous domain perpendicular to the surface. That is, molecules with velocities $v = (v_1, v_2, v_3)^T$ travel towards the mean wafer surface when $v_3 < 0$. More precisely, the microstructured surface Γ_ε is for this paper assumed to be given by a function

$$y = \tilde{h}(x) = \varepsilon\, h(x, \tfrac{x}{\varepsilon}), \tag{1.8}$$

with the small parameter $0 < \varepsilon \ll 1$. This parameter represents the ratio of the typical feature mouth (or, more precisely, the so-called pitch, i.e., the distance from the center of one feature to the next one), e.g., 1 μm, to the typical length scale of the mesoscopic scale model, e.g., 100 μm.

In this dimensionless form, $h(x, \xi)$ is periodic in $\xi \equiv (\xi_1, \xi_2)^T$ with period 1 in each component; that is,

$$h(x, \xi) = h(x, \xi + e_1) = h(x, \xi + e_2) \quad \text{for all } (x, \xi), \tag{1.9}$$

where $e_1 = (1, 0)^T$ and $e_2 = (0, 1)^T$ denote the unit vectors in two dimensions. Physically, this reflects the fact that the microscopic features on a computer chip are not random but arranged in clusters of hundreds or thousands of (by design) identical features, periodic with the feature pitch as period. However, this periodic structure will be different in different regions of the chip, and therefore we allow the function h to depend on the slow spatial variable x as well. Realistic values for ε include the range from 10^{-4} to 10^{-3}.

The model is complemented by a boundary condition on the reacting surface for all molecules that flow into the gaseous domain, i.e., that satisfy $n \cdot v \leq 0$, where $n =$

$n(x,y)$ denotes the unit outward normal vector at position $(x,y)^T \in \Gamma_\varepsilon$. Specifically, we assume that the reinsertion occurs with Maxwellian distributed velocities and set

(1.10)
$$f(x,y,v,t) = M(v) \int_{n\cdot w>0} a(x,\tfrac{x}{\varepsilon},w)\, f(x,y,w,t)\, dw \quad \text{for } n\cdot v \le 0 \text{ and } (x,y)^T \in \Gamma_\varepsilon,$$

where $a(x,\frac{x}{\varepsilon},w) \ge 0$ denotes a given function, and n is the unit outward normal vector on the surface Γ_ε. This boundary condition reflects a pseudo–steady-state assumption, that is, the deposition of molecules on the surface progresses several orders of magnitude more slowly than the flow of the rarefied gas; therefore neither the functions $a(x,\frac{x}{\varepsilon},w)$ nor the geometry of the surface $y=\tilde{h}(x)$ depends on the time t of the gas flow under consideration. However, they are certainly allowed to depend on x; that is, different regions of the surface can see different deposition conditions; this is important in actual applications.

The model in its present form with a microstructured surface is not numerically tractable because of the high cost of resolving the domain close to the rough surface Γ_ε. The goal of this work is to obtain a model with a reduced boundary condition on a flat surface Γ_0 given by $y=0$ that gives equivalent results for $f(x,y,v,t)$ in the bulk of the gaseous domain away from the surface in an asymptotic sense using the expansion parameter ε.

To this end, we make the ansatz

$$f(x,y,v,t) = \tilde{f}(x,y,v,t) + \hat{f}(\tfrac{x}{\varepsilon},\tfrac{y}{\varepsilon},\tfrac{t}{\varepsilon},x,v,t) + o(1), \tag{1.11}$$

where $\tilde{f}$ denotes the bulk variable, for which we wish to derive a numerically tractable model, and $\hat{f}$ is the small-scale correction, which is assumed to be periodic in $\xi = \frac{x}{\varepsilon}$ in the same way as the surface function $h(x,\xi)$. There is only one scale for the velocity v of the molecules; hence there have to be pairs of corresponding length and time scales on both the long (x and t) and the short scales ($\xi = \frac{x}{\varepsilon}$ and $\tau = \frac{t}{\varepsilon}$). Note that we have assumed the Knudsen number to be of order $O(1)$ in the bulk of the mesoscopic scale model, whose solution is $f(x,y,v,t)$. This means that on the $O(\varepsilon)$ spatial scale, the feature scale, collisions will be negligible, and we obtain free transport inside the features of the surface. Therefore, due to the hyperbolic nature of the Boltzmann equation, we can assume only that $\hat{f}$ decays weakly with ε at any fixed distance from the surface; that is, we assume that small-scale fluctuations in the inner solution $\hat{f}$ average out to zero at any fixed finite distance above the surface as $\varepsilon \to 0$. That is, we require that

$$\lim_{\varepsilon\to 0} \iint \hat{f}(\tfrac{x}{\varepsilon},\tfrac{y}{\varepsilon},\tfrac{t}{\varepsilon},x,v,t)\, \psi(x,t)\, dx\, dt = 0 \tag{1.12}$$

for all test functions $\psi(x,t)$.

Based on these assumptions, we will derive the appropriate reduced boundary condition for the bulk term $\tilde{f}(x,y,v,t)$ on the *flat* surface $y=0$

$$\tilde{f}(x,y,v,t) = M(v) \int \tilde{a}(x,v,w)\, \tilde{f}(x,y,w,t)\, dw \quad \text{for } v_3 > 0 \text{ and } y = 0, \tag{1.13}$$

where the integral kernel $\tilde{a}(x,v,w)$ incorporates the information about the microscopic surface geometry into the flow equations. This problem is tractable numerically, since it is posed on a domain with a flat reacting surface, and the values of $\tilde{a}(x,v,w)$ can be

precomputed for all times, since $\tilde{a}$ does not depend on the macroscopic time t. This effective boundary condition still reflects the assumption of a pseudo–steady state in the original model, because $\tilde{a}$ does not depend on the time t that is the relevant time scale for the gas flow on the mesoscopic scale.

We remark that in the previous work [2, 5, 6, 8, 11], where the flow was assumed to be Maxwellian inside the features as well, the only information needed about the surface geometry was the ratio of surface areas between the microstructured surface and the flat surface. This ratio is now replaced by the integral kernel $\tilde{a}$, which contains much more information about the actual shape of the surface and is necessary for the non-Maxwellian picture. This approach has to be viewed as an alternative to the work presented in [1], which deals with specular reflections on a random surface, as opposed to random reflections on a deterministic surface. The result is, however, quite different for the obvious reason that we allow for absorption into the surface; i.e., in our resulting homogenized problem the total mass inside the gas phase domain will not be conserved.

Section 2 details the analytical derivation of the reduced boundary condition, where we will restrict ourselves to the two-dimensional case for notational simplicity. The generalization to three dimensions is straightforward. Section 3 provides a numerical demonstration of the result for a setup that closely resembles the structure of the application problem under consideration by using a periodic boundary geometry.

2. Analysis. This section considers the two-dimensional linear Boltzmann equation

$$\frac{\partial f}{\partial t} + v_1 \frac{\partial f}{\partial x} + v_2 \frac{\partial f}{\partial y} = Q(f) \tag{2.1}$$

with boundary condition (1.10) at the reacting wafer surface. Note that we have changed the notation slightly in going to a two-dimensional problem: The gas phase domain is now given by $y > \varepsilon\, h(x, \frac{x}{\varepsilon})$, and molecules travel towards the surface for velocities $v = (v_1, v_2)^T$ with $v_2 < 0$. We introduce the surface density ρ as

$$\rho(\tfrac{x}{\varepsilon}, \tfrac{t}{\varepsilon}, x, t) = \int_{n\cdot w \geq 0} a(x, \tfrac{x}{\varepsilon}, w) \left(\tilde{f}(x, 0, w, t) + \hat{f}(\tfrac{x}{\varepsilon}, h(x, \tfrac{x}{\varepsilon}), \tfrac{t}{\varepsilon}, x, w, t) \right) dw + o(1) \tag{2.2}$$

for the integral on the right-hand side of (1.10). Using the ansatz $f = \tilde{f} + \hat{f} + o(1)$, the boundary condition (1.10) then reads

$$\tilde{f}(x, 0, v, t) + \hat{f}(\xi, h(x, \xi), \tau, x, v, t) = M(v)\, \rho(\xi, \tau, x, t) + o(1) \quad \text{for } \sigma \leq 0, \tag{2.3}$$

using also the shorthand notation $\sigma(x, \xi, v) = n \cdot v = v_1 \partial_\xi h(x, \xi) - v_2 + \mathcal{O}(\varepsilon)$. Here we have already replaced the argument $y = \varepsilon\, h(x, \xi)$ by 0 in the evaluation of $\tilde{f}$ at the boundary. The homogenized boundary condition (1.13) on the limiting flat surface $y = 0$ will be as derived follows.

Tracing back the characteristics, we will first solve the boundary layer problem for the inner solution $\hat{f}$ in terms of boundary data consisting of the outer solution $\tilde{f}$ and the boundary density ρ defined in (2.2). This result is given in Theorem 2.2 in subsection 2.1. The surface density ρ involves both the inner and the outer solutions and forms the connection between them. Its behavior in the limit of the fast time variables $\tau \to \infty$ is analyzed in Theorem 2.4.

In subsection 2.2, using the weak decay condition (1.12), we derive the homogenized boundary condition for $\tilde{f}$ on the flat surface $y = 0$ in terms of the surface density ρ and finally in terms of $\tilde{a}$ as in (1.13). This quantity will contain only information about the microscopic and quasi-periodic geometry of the surface and can be solved beforehand for a given surface structure.

2.1. Solution to the inner equation. The asymptotic ansatz applied to (2.1) leads to the inner problem for the layer correction term $\hat{f}(\xi, \eta, \tau, x, v, t)$ as a function of the layer variables ξ, η, τ:

$$\frac{\partial \hat{f}}{\partial \tau} + v_1 \frac{\partial \hat{f}}{\partial \xi} + v_2 \frac{\partial \hat{f}}{\partial \eta} = 0 \tag{2.4}$$

for any fixed x, v, t. Using the method of characteristics with $\xi' = v_1$ and $\eta' = v_2$ yields

$$\hat{f}(\xi, \eta, \tau, x, v, t) = \hat{f}(\xi - sv_1, \eta - sv_2, \tau - s, x, v, t)$$

for all parameters s sufficiently small. We can then follow the characteristics back to the boundary or back to the initial condition $\hat{f} = 0$ at $\tau = 0$ to obtain

$$\begin{aligned} &\hat{f}(\xi, \eta, \tau, x, v, t) \\ &\quad = \hat{f}(\xi - v_1 s, \eta - v_2 s, \tau - s, x, v, t) \\ &\quad = \begin{cases} \hat{f}(\xi - v_1\phi_0, \eta - v_2\phi_0, \tau - \phi_0, x, v, t) & \text{if } \phi_0 < \tau, \\ 0 & \text{if } \phi_0 \geq \tau \end{cases} \\ &\quad = H(\tau - \phi_0)\, \hat{f}(\xi - v_1\phi_0, \eta - v_2\phi_0, \tau - \phi_0, x, v, t), \end{aligned} \tag{2.5}$$

with

$$\phi_0(\xi, \eta, x, v) = \begin{cases} \min\{s > 0 \,:\, \eta - v_2 s = h(x, \xi - v_1 s)\}, \\ \infty \quad \text{if } \eta - v_2 s \neq h(x, \xi - v_1 s) \text{ for all } s > 0, \end{cases} \tag{2.6}$$

and using the Heaviside function

$$H(z) = \begin{cases} 0 & \text{for } z < 0, \\ 1 & \text{for } z \geq 0. \end{cases}$$

The function ϕ_0 denotes the intersection time, i.e., the time it takes for a molecule emitted from the surface with velocity v to reach the point (ξ, η). The boundary condition for the inner equation then reads

$$\hat{f}(\xi, h(x, \xi), \tau, x, v, t) = M(v)\, \rho(\xi, \tau, x, t) - \tilde{f}(x, 0, v, t) \quad \text{for } \sigma(x, \xi, v) \leq 0. \tag{2.7}$$

In order to apply the boundary condition to the solution in (2.5), we need to guarantee that $\sigma(x, \xi - v_1\phi_0, v) \leq 0$ in the case when the characteristic traces back to the boundary, i.e., ϕ_0 is finite.

LEMMA 2.1. *If $\phi_0(\xi, \eta, x, v)$ is finite and* either (i) $\eta = h(x, \xi)$ *and* $\sigma(x, \xi, v) > 0$ or (ii) $\eta > h(x, \xi)$, *then it holds that* $\sigma(x, \xi - v_1\phi_0, v) \leq 0$.

Proof. Define

$$g(s) = \eta - sv_2 - h(x, \xi - sv_1).$$

By definition of ϕ_0, the first positive root of $g(s)$ is given by $s = \phi_0$. Then

$$g'(s) = -v_2 + v_1 \partial_\xi h(x, \xi - s v_1) = \sigma(x, \xi - s v_1, v),$$

and we have to show that $g'(\phi_0) = \sigma(x, \xi - v_1 \phi_0, v) \leq 0$.

Case (i). If $\eta = h(x, \xi)$ and $\sigma(x, \xi, v) > 0$, then

$$g(0) = 0, \quad g'(0) > 0 \quad \Rightarrow \quad g(s) > 0 \quad \text{for } 0 < s < \phi_0(\xi, \eta, x, v) \quad \Rightarrow \quad g'(\phi_0) \leq 0.$$

Case (ii). If $\eta > h(x, \xi)$, then

$$g(0) > 0 \quad \Rightarrow \quad g(s) > 0 \quad \text{for } s < \phi_0(\xi, \eta, x, v), \quad g(\phi_0) = 0 \quad \Rightarrow \quad g'(\phi_0) \leq 0. \qquad \square$$

THEOREM 2.2. *The solution $\hat{f}(\xi, \eta, \tau, x, v, t)$ to the inner problem* (2.4) *with boundary condition* (2.7) *and initial condition $\hat{f} = 0$ at $\tau = 0$ is given by*

$$\hat{f}(\xi, \eta, \tau, x, v, t) = \begin{cases} M(v)\rho(\xi, \tau, x, t) - \tilde{f}(x, y = 0, v, t) \\ \qquad \textit{if } \eta = h(x, \xi) \textit{ and } \sigma(x, \xi, v) \leq 0, \\ H(\tau - \phi_0)(M(v)\rho(\xi - v_1\phi_0, \tau - \phi_0, x, t) - \tilde{f}(x, y = 0, v, t)) \\ \qquad \textit{if (i) } \eta = h(x, \xi) \textit{ and } \sigma(x, \xi, v) > 0 \text{ or (ii) } \eta > h(x, \xi), \end{cases}$$

with the intersection time ϕ_0 given by (2.6).

Proof. Define the characteristics

$$g(s, \xi, \eta, \tau, x, v, t) := \hat{f}(\xi - s v_1, \eta - s v_2, \tau - s, x, v, t);$$

then $dg/ds = 0$. Follow the characteristics in four possible cases, as follows.

Case 1. If $\eta > h(x, \xi)$ and the line $(\xi - s v_1, \eta - s v_2, \tau - s)$ intersects $\tau = 0$ first $(\tau < \phi_0(\xi, \eta, x, v))$, then

$$\hat{f}(\xi, \eta, \tau, x, v, t) = \hat{f}(\xi - \tau v_1, \eta - \tau v_2, \tau = 0, x, v, t) = 0.$$

Case 2. If $\eta > h(x, \xi)$ and the line $(\xi - s v_1, \eta - s v_2, \tau - s)$ intersects $\eta = h(x, \xi)$ first $(\tau > \phi_0(\xi, \eta, x, v))$, then

$$\hat{f}(\xi, \eta, \tau, x, v, t) = \hat{f}(\xi - v_1\phi_0, \eta - v_2\phi_0, \tau - \phi_0, x, v, t), \quad \phi_0 = \phi_0(\xi, \eta, x, v).$$

Then $\eta - v_2\phi_0(\xi, \eta, x, v) = h(x, \xi - v_1\phi_0(\xi, \eta, x, v))$, and hence

$$\hat{f}(\xi, \eta, \tau, x, v, t) = M(v)\rho(\xi - v_1\phi_0, \tau - \phi_0, x, t) - \tilde{f}(x, 0, v, t),$$

because $\sigma(x, \xi - \phi_0 v_1, v) \leq 0$ by Lemma 2.1.

Case 3. If $\eta = h(x, \xi)$ and $\sigma(x, \xi, v) \leq 0$ (boundary condition), then

$$\hat{f}(\xi, h(x, \xi), \tau, x, v, t) = M(v)\rho(\xi, \tau, x, t) - \tilde{f}(x, 0, v, t).$$

Case 4. If $\eta = h(x, \xi)$ and $\sigma(x, \xi, v) > 0$, then follow the ray $(\xi - s v_1, \eta - s v_2, \tau - s)$ back until either $\tau - s = 0$ or $\eta - s v_2 = h(x, \xi - s v_1)$. We distinguish the following two subcases.

(a) If $\eta = h(x, \xi)$ and $\sigma(x, \xi, v) > 0$ and $\tau < \phi_0(\xi, \eta, x, v)$ (that is, $\tau = 0$ is intersected first), then

$$\hat{f}(\xi, \eta, \tau, x, v, t) = \hat{f}(\xi - \tau v_1, h - \tau v_2, \tau = 0, x, v, t) = 0.$$

(b) If $\eta = h(x,\xi)$ and $\sigma(x,\xi,v) > 0$ and $\tau > \phi_0(\xi,\eta,x,v)$ (that is, $\eta = h(x,\xi)$ is intersected first), then

$$\hat{f}(\xi,\eta,\tau,x,v,t) = \hat{f}(\xi - v_1\phi_0, h - v_2\phi_0, \tau - \phi_0, x, v, t),$$
$$\phi_0 = \phi_0(\xi, h(x,\xi), x, v),$$

and

$$\hat{f}(\xi,\eta,\tau,x,v,t) = M(v)\rho(\xi - v_1\phi_0, \tau - \phi_0, x, t) - \tilde{f}(x,0,v,t),$$

because $\sigma(x, \xi - \phi_0 v_1, v) \le 0$ by Lemma 2.1. □

We now turn to the evolution of the surface density ρ defined in (2.2). On the one hand, according to (2.2), ρ is defined in terms of $\tilde{f}$ and $\hat{f}$ evaluated at the surface $y = \varepsilon\, h(x, \frac{x}{\varepsilon})$. On the other hand, according to Theorem 2.2, the inner solution $\hat{f}$ at the surface is given in terms of ρ and $\tilde{f}$. Combining these two formulas, we are able to write an evolution equation for ρ in terms of the outer solution $\tilde{f}$ alone. In other words, we are able to write $\rho = F[\tilde{f}]$ with some integral operator to be defined. Moreover, for the computation of the reduced boundary condition (1.13), we will need to make a statement about the behavior of $\rho(\xi,\tau,x,t)$ for the fast time variable $\tau \to \infty$. To this end, it is convenient to make the following definitions. First we introduce for convenience a new symbol for the function ϕ_0 in Theorem 2.2 evaluated at the boundary by defining $\phi_1(\xi,x,v) := \phi_0(\xi, h(x,\xi), x, v)$, that is,

$$\phi_1(\xi,x,v) = \begin{cases} \min\{s > 0 \,:\, h(x,\xi) - v_2 s = h(x, \xi - v_1 s)\}, \\ \infty \quad \text{if } h(x,\xi) - v_2 s \neq h(x,\xi - v_1 s) \text{ for all } s > 0. \end{cases} \tag{2.8}$$

The function ϕ_1 denotes the time taken by molecules emitted from the boundary to reach another point at the boundary, and is formally set to ∞ if they never do. Furthermore, we define the indicator function on the set of all ξ, x, v for which, formally, ϕ_1 is infinite:

$$\chi(\xi,x,v) = \begin{cases} 0 & \text{if } \phi_1(\xi,x,v) < \infty, \\ 1 & \text{if } \phi_1(\xi,x,v) = \infty. \end{cases} \tag{2.9}$$

To analyze the limiting behavior of the surface density ρ as $\tau \to \infty$ we will need that the corresponding limiting problem actually has a solution. This is the statement of the following lemma.

LEMMA 2.3. *The integral equation*

$$Z(\xi,x) = g(\xi,x) + \int H(\sigma(x,\xi,w))a(x,\xi,w)\,(1 - \chi(\xi,x,w))\,M(w)\,Z(\xi - w_1\phi_1(\xi,x,w), x)\,dw$$

has a solution $Z(\xi,x)$ for any function $g(\xi,x)$, provided that

$$\int H(\sigma(x,\xi,w))a(x,\xi,w)\,(1 - \chi(\xi,x,w))\,M(w)\,dw \le C \tag{2.10}$$

for some constant $0 \le C < 1$.

Proof. Introduce the notation

$$A(x,\xi,w) := H(\sigma(x,\xi,w))a(x,\xi,w)\,(1 - \chi(\xi,x,w))\,M(w),$$

which satisfies $A(x,\xi) \geq 0$ since $a(x,\xi,w)$ and all other terms are nonnegative. Then we consider the integral equation

$$Z(\xi,x) = \int A(x,\xi,w)\, Z(\xi - w_1\phi_1(\xi,x,w),x)\, dw + g(\xi,x).$$

To compute the solution iteratively, introduce the fixed-point iteration for $\{Z_n(\xi,x)\}$ by

$$Z_{n+1}(\xi,x) = \int A(x,\xi,w)\, Z_n(\xi - w_1\phi_1(\xi,x,w),x)\, dw + g(\xi,x) \quad \text{for } n = 0,1,2,\ldots,$$

with initial iterate $Z_0(\xi,x) = 0$. Hence, the difference between successive iterates satisfies

$$(Z_{n+1} - Z_n)(\xi,x) = \int A(x,\xi,w)\ (Z_n - Z_{n-1})(\xi - w_1\phi_1(\xi,x,w),x)\, dw,$$

and we can bound it as

$$\|Z_{n+1} - Z_n\|_\infty \leq \int A(x,\xi,w)\, dw\, \|Z_n - Z_{n-1}\|_\infty.$$

The convergence of this sequence is guaranteed if condition (2.10) is satisfied. ∎

Remark 2.1. Condition (2.10) constitutes a restriction on the function $a(x,\xi,w)$ in the boundary condition chosen in the application. It will be verified in section 3 for our choice of $a(x,\xi,w)$.

THEOREM 2.4. *If $a(x,\xi,w)$ satisfies* (2.10)*, then the boundary density ρ is of the form*

$$\rho(\xi,\tau,x,t) = F_\infty[\tilde{f}](\xi,x,t) + \rho_1(\xi,\tau,x,t),$$

where $F_\infty[\tilde{f}] = \rho_\infty$ is given implicitly by the integral equation

$$\begin{aligned}\rho_\infty(\xi,x,t) = &\int H(\sigma(x,\xi,w))a(x,\xi,w)\chi(\xi,x,w)\, \tilde{f}(x,0,w,t)\, dw \\ &+ \int H(\sigma(x,\xi,w))a(x,\xi,w)(1-\chi(\xi,x,w))M(w)\, \rho_\infty(\xi - w_1\phi_1,x,t)\, dw,\end{aligned} \tag{2.11}$$

and the remainder term ρ_1 satisfies $\int_0^\infty \rho_1(\xi,\tau,x,t)\, d\tau < \infty$.

Proof. Inserting the expression for $\hat{f}(\xi,h(x,\xi),\tau,x,v,t)$ for $\sigma(x,\xi,v) > 0$ from Theorem 2.2 into the definition (2.2) for ρ gives

(2.12)

$$\begin{aligned}\rho(\xi,\tau,x,t) = &\int H(\sigma(x,\xi,w))a(x,\xi,w)\,[1 - H(\tau - \phi_1(\xi,x,w))]\ \tilde{f}(x,0,w,t)\, dw \\ &+ \int H(\sigma(x,\xi,w))a(x,\xi,w)H(\tau - \phi_1(\xi,x,w))M(w)\, \rho(\xi - w_1\phi_1,\tau - \phi_1,x,t)\, dw.\end{aligned}$$

We are interested in the limiting behavior for $\tau \to \infty$. To this end, we have to distinguish the cases when $\phi_1 < \infty$ holds, which means that the Heaviside function in (2.12) will become equal to unity for τ sufficiently large, and the case $\phi_1 = \infty$ for

which the Heaviside function will always be zero. Therefore we write

$$
\begin{aligned}
&\rho(\xi,\tau,x,t)\\
&\quad=\int H(\sigma(x,\xi,w))a(x,\xi,w)\left[1-(1-\chi(\xi,x,w))H(\tau-\phi_1(\xi,x,w))\right]\,\tilde{f}(x,0,w,t)\,dw\\
&\quad+\int H(\sigma(x,\xi,w))a(x,\xi,w)(1-\chi(\xi,x,w))H(\tau-\phi_1(\xi,x,w))M(w)\\
&\qquad\qquad\times\rho(\xi-w_1\phi_1,\tau-\phi_1,x,t)\,dw,
\end{aligned}
\tag{2.13}
$$

with χ defined as in (2.9). Letting, formally, τ tend to infinity gives $\rho=\rho_\infty+\rho_1$ with (2.11) for $\rho_\infty(\xi,x,t)$. The solution ρ_∞ exists due to Lemma 2.3. This is, of course, only a formal definition for ρ_∞; the key is to estimate the remainder term ρ_1. The remainder term ρ_1 satisfies the integral equation

$$
\begin{aligned}
&\rho_1(\xi,\tau,x,t)\\
&\quad=\int H(\sigma(x,\xi,w))a(x,\xi,w)(1-\chi(\xi,x,w))H(\tau-\phi_1(\xi,x,w))M(w)\\
&\qquad\qquad\times\rho_1(\xi-w_1\phi_1,\tau-\phi_1,x,t)\,dw\\
&\quad+\int H(\sigma(x,\xi,w))a(x,\xi,w)(1-\chi(\xi,x,w))(1-H(\tau-\phi_1(\xi,x,w)))\\
&\qquad\qquad\times(\tilde{f}(x,0,w,t)-M(w)\rho_\infty(\xi-w_1\phi_1,x,t))\,dw.
\end{aligned}
\tag{2.14}
$$

Let $\mathcal{L}[\rho_1](\xi,s,x,t):=\int_0^\infty \rho_1(\xi,\tau,x,t)\,e^{-s\tau}\,d\tau$, with $0\le s<\infty$, denote the Laplace transform of ρ_1. The transform of (2.14) is then given by

$$
\begin{aligned}
&\mathcal{L}[\rho_1](\xi,s,x,t)\\
&\quad=\int H(\sigma(x,\xi,w))a(x,\xi,w)(1-\chi(\xi,x,w))M(w)e^{-s\phi_1(x,\xi,w)}\\
&\qquad\qquad\times\mathcal{L}[\rho_1](\xi-w_1\phi_1,s,x,t)\,dw\\
&\quad+\int H(\sigma(x,\xi,w))a(x,\xi,w)(1-\chi(\xi,x,w))\,\frac{1}{s}\,\left[1-e^{-s\phi_1(x,\xi,w)}\right]\\
&\qquad\qquad\times(\tilde{f}(x,0,w,t)-M(w)\rho_\infty(\xi-w_1\phi_1,x,t))\,dw.
\end{aligned}
$$

The variable s appears as a parameter only in the integral equation, so in order to find $\mathcal{L}[\rho_1](\xi,0,x,t)$, we can solve directly

$$
\begin{aligned}
&\mathcal{L}[\rho_1](\xi,0,x,t)\\
&\quad=\int H(\sigma(x,\xi,w))a(x,\xi,w)(1-\chi(\xi,x,w))M(w)\,\mathcal{L}[\rho_1](\xi-w_1\phi_1,0,x,t)\,dw\\
&\quad+\int H(\sigma(x,\xi,w))a(x,\xi,w)(1-\chi(\xi,x,w))\,\frac{1}{\phi_1}\\
&\qquad\qquad\times(\tilde{f}(x,0,w,t)-M(w)\rho_\infty(\xi-w_1\phi_1,x,t))\,dw.
\end{aligned}
\tag{2.15}
$$

By Lemma 2.3, the solution $\mathcal{L}[\rho_1](\xi,0,x,t)$ exists, and we therefore have

$$
\int_0^\infty \rho_1(\xi,\tau,x,t)\,d\tau=\mathcal{L}[\rho_1](\xi,0,x,t)<\infty
$$

by definition of the transform. □

Remark 2.2. Theorem 2.4 establishes the limiting behavior of ρ for $\tau \to \infty$ in a weak sense, namely, that $\rho_1(\xi, \frac{t}{\varepsilon}, x, t)$ will go to zero when integrated with respect to t against a test function. This is precisely the property that will be needed for the derivation of the boundary condition.

2.2. Boundary condition for the outer solution. This subsection will use the weak convergence property for the inner solution to derive the boundary condition for the outer solution. Because the layer equation (2.4) does not have any damping, $\hat{f}$ will give a contribution throughout the half-space $y > 0$. However, this contribution will result in high frequency oscillations only for $y > 0$. Therefore, we formulate the Boltzmann equation in the weak sense and require that $\hat{f}$ tend to zero weakly for fixed $y > 0$ and $\varepsilon \to 0$. Thus, we require in a weak sense for y, v fixed that

$$\lim_{\varepsilon \to 0} \int_0^\infty \int_{-\infty}^{+\infty} \hat{f}\left(\tfrac{x}{\varepsilon}, \tfrac{y}{\varepsilon}, \tfrac{t}{\varepsilon}, x, v, t\right) \psi(x,t)\, dx\, dt = 0 \tag{2.16}$$

for all test functions $\psi(x,t)$.

THEOREM 2.5. *The integral in* (2.16) *goes to zero for* $\varepsilon \to 0$ *and all fixed* $y > 0$, v *if and only if*

$$\tilde{f}(x,0,v,t) = M(v) \int_0^1 F_\infty[\tilde{f}](\xi - v_1\phi_3, x, t)\, d\xi \quad \textit{for all } v_2 > 0$$

holds with

$$\phi_3(\xi, x, v) = \min\{s \in \mathbb{R} : -v_2 s = h(x, \xi - v_1 s)\}$$

and the operator $F_\infty[\tilde{f}]$ *defined as in Theorem* 2.4.

Proof. Let $I_\varepsilon := \iint \hat{f}(\frac{x}{\varepsilon}, \frac{y}{\varepsilon}, \frac{t}{\varepsilon}, x, v, t)\, \psi(x,t)\, dx\, dt$ denote the integral in (2.16). We need to estimate I_ε for velocities v with $v_2 > \kappa(\varepsilon)$, where κ is some function with $\kappa(\varepsilon) > 0$ and $\kappa(\varepsilon) \to 0$ as well as $\varepsilon/\kappa(\varepsilon) \to 0$ as $\varepsilon \to 0$. Since we are interested in the bulk solution $\tilde{f}(x, y, v, t)$, we need to consider only $\eta = \frac{y}{\varepsilon} > h(x, \frac{x}{\varepsilon})$.

We have from Theorem 2.2 for $\frac{y}{\varepsilon} > h(x, \frac{x}{\varepsilon})$ and $v_2 > \kappa(\varepsilon) > 0$

$$\hat{f}(\tfrac{x}{\varepsilon}, \tfrac{y}{\varepsilon}, \tfrac{t}{\varepsilon}, x, v, t) = H(\tfrac{t}{\varepsilon} - \phi_0)\Big(M(v)\rho(\tfrac{x}{\varepsilon} - v_1\phi_0, \tfrac{t}{\varepsilon} - \phi_0, x, t) - \tilde{f}(x, 0, v, t)\Big),$$

with $\phi_0(\frac{x}{\varepsilon}, \frac{y}{\varepsilon}, x, v) = \min\big\{s > 0 : \frac{y}{\varepsilon} - v_2 s = h(x, \frac{x}{\varepsilon} - v_1 s)\big\}$, where this simplified (compared to (2.6)) definition for ϕ_0 is possible because the existence of the minimum is guaranteed for $v_2 > \kappa(\varepsilon) > 0$. To transform

$$\phi_0(\tfrac{x}{\varepsilon}, \tfrac{y}{\varepsilon}, x, v) = \frac{y}{v_2\varepsilon} + \phi_2(\tfrac{x}{\varepsilon} - \tfrac{v_1 y}{v_2 \varepsilon}, \tfrac{y}{\varepsilon}, x, v),$$

define

$$\phi_2(\tfrac{x}{\varepsilon} - \tfrac{v_1 y}{v_2\varepsilon}, \tfrac{y}{\varepsilon}, x, v) = \min\left\{s > -\frac{y}{v_2\varepsilon} \,:\, -v_2 s = h(x, \tfrac{x}{\varepsilon} - \tfrac{v_1 y}{v_2\varepsilon} - v_1 s)\right\}. \tag{2.17}$$

This gives

$$\begin{aligned}&\hat{f}(\tfrac{x}{\varepsilon}, \tfrac{y}{\varepsilon}, \tfrac{t}{\varepsilon}, x, v, t)\\ &= H(\tfrac{t}{\varepsilon} - \tfrac{y}{v_2\varepsilon} - \phi_2)\left(M(v)\rho(\tfrac{x}{\varepsilon} - \tfrac{v_1 y}{v_2\varepsilon} - v_1\phi_2, \tfrac{t}{\varepsilon} - \tfrac{y}{v_2\varepsilon} - \phi_2, x, t) - \tilde{f}(x, 0, v, t)\right),\end{aligned}$$

with $\phi_2 \equiv \phi_2(\frac{x}{\varepsilon} - \frac{v_1 y}{v_2 \varepsilon}, \frac{y}{\varepsilon}, x, v)$ for $v_2 > \kappa(\varepsilon)$.

We introduce the transformation

$$x = x_{j\xi} := \frac{v_1}{v_2} y + \varepsilon j + \varepsilon \xi, \quad j \in \mathbb{Z}, \quad \xi \in [0,1],$$

to rewrite I_ε into

$$I_\varepsilon = \varepsilon \sum_{j=-\infty}^{+\infty} \int_0^\infty \int_0^1 H(\tfrac{t}{\varepsilon} - \tfrac{y}{v_2\varepsilon} - \phi_2)\Big(M(v)\rho(j + \xi - v_1\phi_2, \tfrac{t}{\varepsilon} - \tfrac{y}{v_2\varepsilon} - \phi_2, x_{j\xi}, t) - \tilde{f}(x_{j\xi}, 0, v, t)\Big)\, \psi(x_{j\xi}, t)\, d\xi\, dt,$$

with $\phi_2 \equiv \phi_2(j + \xi, \frac{y}{\varepsilon}, x_{j\xi}, v) = \min\{s > -\frac{y}{v_2\varepsilon} \,:\, -v_2 s = h(x_{j\xi}, j + \xi - v_1 s)\}$. Because $h(x,\xi)$ and hence also ρ and ϕ_2 are 1-periodic in ξ, we can drop the j from their first arguments. Since ρ, $\tilde{f}$, and ψ vary only slowly in x and since ξ varies only in $[0,1]$, we make an $\mathcal{O}(\varepsilon)$ perturbation by replacing $x_{j\xi}$ by $x_j := \frac{v_1}{v_2} y + \varepsilon j$. Together this gives

$$I_\varepsilon = \varepsilon \sum_{j=-\infty}^{+\infty} \int_0^\infty \int_0^1 H(\tfrac{t}{\varepsilon} - \tfrac{y}{v_2\varepsilon} - \phi_2)\Big(M(v)\rho(\xi - v_1\phi_2, \tfrac{t}{\varepsilon} - \tfrac{y}{v_2\varepsilon} - \phi_2, x_j, t) - \tilde{f}(x_j, 0, v, t)\Big)\, \psi(x_j, t)\, d\xi\, dt + \mathcal{O}(\varepsilon),$$

with $\phi_2 \equiv \phi_2(\xi, \frac{y}{\varepsilon}, x_j, v) = \min\{s > -\frac{y}{v_2\varepsilon} \,:\, -v_2 s = h(x_j, \xi - v_1 s)\}$. The sum over the j forms a Riemann sum for an integral with approximation error $\mathcal{O}(\varepsilon)$. Thus, we obtain

(2.18)
$$I_\varepsilon = \int_0^\infty \int_{-\infty}^{+\infty} \int_0^1 H(\tfrac{t}{\varepsilon} - \tfrac{y}{v_2\varepsilon} - \phi_2)\Big(M(v)\rho(\xi - v_1\phi_2, \tfrac{t}{\varepsilon} - \tfrac{y}{v_2\varepsilon} - \phi_2, z, t) - \tilde{f}(z, 0, v, t)\Big)\, \psi(z, t)\, d\xi\, dz\, dt + \mathcal{O}(\varepsilon),$$

with $\phi_2 \equiv \phi_2(\xi, \frac{y}{\varepsilon}, z, v) = \min\{s > -\frac{y}{v_2\varepsilon} \,:\, -v_2 s = h(z, \xi - v_1 s)\}$.

Now we replace $\rho = \rho_\infty + \rho_1$ with $F_\infty[\tilde{f}] = \rho_\infty$, as defined in Theorem 2.4. For the remainder term involving ρ_1 this gives

$$E_\varepsilon := \int_0^\infty \int_{-\infty}^{+\infty} \int_0^1 H(\tfrac{t}{\varepsilon} - \tfrac{y}{v_2\varepsilon} - \phi_2) M(v) \rho_1(\xi - v_1\phi_2, \tfrac{t}{\varepsilon} - \tfrac{y}{v_2\varepsilon} - \phi_2, z, t) \times \psi(z, t)\, d\xi\, dz\, dt.$$

We transform again to a fast variable $\tau := \frac{t}{\varepsilon} - \frac{y}{v_2\varepsilon} - \phi_2$ and obtain

$$E_\varepsilon = \varepsilon M(v) \int_0^\infty \int_{-\infty}^{+\infty} \int_0^1 \rho_1(\xi - v_1\phi_2, \tau, z, \tfrac{y}{v_2} + \varepsilon\tau + \varepsilon\phi_2) \times \psi(z, \tfrac{y}{v_2} + \varepsilon\tau + \varepsilon\phi_2)\, d\xi\, dz\, d\tau.$$

Since the test function ψ is smooth and has compact support, and since Theorem 2.4 guarantees that $\int_0^\infty \rho_1(\xi - v_1\phi_2, \tau, z, t)\, d\tau$ is bounded, the integrals in E_ε remain bounded, and $E_\varepsilon \to 0$, as $\varepsilon \to 0$.

In the remaining parts of (2.18), we use the fact that the Heaviside function is scaling-invariant to get $H(\frac{t}{\varepsilon} - \frac{y}{v_2\varepsilon} - \phi_2) = H(t - \frac{y}{v_2} - \varepsilon\phi_2)$. From the definition (2.17) of ϕ_2 it follows that, for $v_2 > \kappa(\varepsilon) > 0$, $|\phi_2| \leq \max\{h\}/\kappa(\varepsilon)$ holds. Therefore, the dependence of the Heaviside function on $\varepsilon\phi_2$ is negligible, and it can be taken out of the ξ-integral, introducing only another $\mathcal{O}(\frac{\varepsilon}{\kappa})$ error. Also, the integrand involving $\tilde{f}$ does not depend on the fast variable ξ any more, and so the integration over the interval $[0,1]$ yields just unity. This gives after $\varepsilon \to 0$

(2.19)
$$0 = \int_0^\infty \int_{-\infty}^{+\infty} H(t - \tfrac{y}{v_2}) \left(M(v) \int_0^1 \rho_\infty(\xi - v_1\phi_2, z, t)\, d\xi - \tilde{f}(z, 0, v, t) \right) \psi(z, t)\, dz\, dt$$

for any fixed $v_2 > 0$. Taking the limit $\varepsilon \to 0$ in $\phi_2 \equiv \phi_2(\xi, \frac{y}{\varepsilon}, z, v) = \min\{s > -\frac{y}{v_2\varepsilon} : -v_2 s = h(z, \xi - v_1 s)\}$ results in the minimum being taken over the entire real line, since $y > 0$ and $v_2 > 0$. It also makes ϕ_2 independent of ε and y, and we introduce the notation

$$\phi_3 \equiv \phi_3(\xi, z, v) = \min\{s \in \mathbb{R} : -v_2 s = h(z, \xi - v_1 s)\}.$$

Therefore, (2.19) is satisfied in a weak sense if and only if

$$\tilde{f}(z, 0, v, t) = M(v) \int_0^1 \rho_\infty(\xi - v_1\phi_3, z, t)\, d\xi$$

holds for all z, v, t with $v_2 > 0$. ∎

Remark 2.3. For any fixed velocity v with $v_2 > 0$, the function ϕ_3 is guaranteed to exist with $-\max\{h(x,\xi)\}/v_2 \leq \phi_3(\xi, x, v) \leq -\min\{h(x,\xi)\}/v_2$, since the surface function $h(x, \xi)$ is smooth.

Theorem 2.5 essentially yields the reduced boundary condition. In practice, one will not solve the integral equation (2.11) to compute $F_\infty[\tilde{f}]$ at every time step. It is preferable to write the term $F_\infty[\tilde{f}]$ as an integral operator with a time independent integral kernel. A direct calculation leads from (2.11) to

$$F_\infty[\tilde{f}](\xi, x, t) = \int K_\infty(\xi, x, v)\, \tilde{f}(x, 0, v, t)\, dv,$$

where the integral kernel K_∞ satisfies

(2.20)
$$\begin{aligned} K_\infty(\xi, x, v) &= H(\sigma(x, \xi, v)) a(x, \xi, v) \chi(\xi, x, v) \\ &\quad + \int H(\sigma(x, \xi, w)) a(x, \xi, w)\, (1 - \chi(\xi, x, w))\, M(w)\, K_\infty(\xi - w_1\phi_1(\xi, x, w), x, v)\, dw, \end{aligned}$$

and $K_\infty(\xi, x, v)$ again exists because of Lemma 2.3. Note that K_∞ will be non-negative, provided that $a(x, \xi, w)$ is not too large.

In summary, we have obtained the following numerical problem for the bulk solution $\tilde{f}(x, y, v, t)$:

$$\frac{\partial \tilde{f}}{\partial t} + v_1 \frac{\partial \tilde{f}}{\partial x} + v_2 \frac{\partial \tilde{f}}{\partial y} = Q(\tilde{f}), \tag{2.21}$$

with boundary condition for inflowing molecules on the *flat* surface $y = 0$

$$\tilde{f}(x,0,v,t) = M(v) \int \tilde{a}(x,v,w)\tilde{f}(x,0,w,t)dw \quad \text{for } v_2 > 0, \tag{2.22}$$

with

$$\tilde{a}(x,v,w) := \int_0^1 K_\infty(\xi - v_1\phi_3(\xi,x,v),x,w)\, d\xi \tag{2.23}$$

and

$$\phi_3(\xi,x,v) = \min\{s \in \mathbb{R} : -v_2 s = h(x,\xi - v_1 s)\}. \tag{2.24}$$

This problem is tractable numerically, since it is posed on a domain with a flat reacting surface; the effect of the microscopic surface has been integrated into the boundary condition. The numerical approach in practice will be as follows.

Step 1. Given a surface function $h(x,\xi)$, compute the intersection times ϕ_1 in (2.8) and ϕ_3 in (2.24).

Step 2. Solve the integral equation (2.20) for K_∞.

Step 3. Compute the boundary kernel $\tilde{a}(x,v,w)$ in (2.23).

The function $\tilde{a}(x,v,w)$ provides the information about the microscopic surface geometry on the macroscopic level. These steps have to be performed only once for a given surface function $h(x,\xi)$. Following these preprocessing steps, the Boltzmann equation (2.21) with the homogenized boundary condition (2.22) is solved for the bulk solution $\tilde{f}(x,y,v,t)$.

3. Numerical validation. As a validation problem, we consider the linear Boltzmann equation for a single species

$$\frac{\partial f}{\partial t} + v_1 \frac{\partial f}{\partial x} + v_2 \frac{\partial f}{\partial y} = Q(f). \tag{3.1}$$

We use a relaxation time approximation $S(v,v') = (1/\tau)M(v)M(v')$ in the linear collision operator (1.6) to obtain the simple form

$$Q(f)(x,y,v,t) = -\frac{1}{\tau}\,[f(x,y,v,t) - N(x,y,t)\,M(v)]$$

with the constant relaxation time $\tau > 0$. Here and in the following, N denotes the number density given by $N(x,y,t) = \int f(x,y,v,t)\,dv$. The numerical domain is chosen as

$$\Omega_\varepsilon = \{(x,y) \in \mathbb{R}^2 : \tilde{h}(x) < y < 1, \quad 0 < x < 1\}, \tag{3.2}$$

with a microstructured surface Γ_ε at the bottom given by

$$y = \tilde{h}(x) = \frac{\varepsilon}{8}\left(1 + \cos\left(2\pi\frac{x}{\varepsilon}\right)\right)$$

with homogenization parameter $0 < \varepsilon \ll 1$. This model is designed to closely resemble the salient features of the application problem that motivated the model and to support the above analysis; this motivates also the choice of the simple form of the collision operator. Other types of boundary models, e.g., using random reflections [1], could be used but would not be appropriate examples for our application.

The following model is chosen as the boundary condition describing the reactions of the gaseous species at the wafer surface that result in the deposition of the solid film. If $0 \le R \le 1$ denotes the sticking factor (the probability that a molecule sticks to the surface), then the inflow into the gaseous domain is equal to $(1-R)$ times the outflow from the gaseous domain, namely (see [4]),

$$\int_{n\cdot v<0} |n\cdot v|\, f(x,y,v,t)\, dv = (1-R) \int_{n\cdot v>0} |n\cdot v|\, f(x,y,v,t)\, dv, \tag{3.3}$$

where $n = n(x,y)$ denotes the unit outward normal vector at position $(x,y) \in \Gamma_\varepsilon$. To obtain the boundary condition in the form (1.10), assume reinjection with random velocities, that is, $f(x,y,v,t) = b(x,t)M(v)$ for $n\cdot v < 0$; then

$$f(x,y,v,t) = M(v) \int_{n\cdot w>0} a\left(x, \frac{x}{\varepsilon}, w\right) f(x,w,t)\, dw,$$

with

$$a\left(x, \frac{x}{\varepsilon}, w\right) = \frac{1-R}{c} |n\cdot w|, \quad c = \int_{n\cdot v<0} |n\cdot v|\, M(v)\, dv.$$

The problem is completed by choosing Maxwellian inflow at the top and periodic boundary conditions at both sides. This chosen setup of the problem is representative of the application under consideration. We choose $\tau = 1$ and $R = 0.5$ as values.

To check condition (2.10) for this choice of $a(x, \frac{x}{\varepsilon}, w)$, we compute

$$\begin{aligned} I(x, \tfrac{x}{\varepsilon}) &:= \int H(\sigma(x, \tfrac{x}{\varepsilon}, w)) a(x, \tfrac{x}{\varepsilon}, w) \left(1 - \chi(\tfrac{x}{\varepsilon}, x, w)\right) M(w)\, dw \\ &= (1-R)\, \frac{\int_{n\cdot w>0} |n\cdot w| \left(1 - \chi(\frac{x}{\varepsilon}, x, w)\right) M(w)\, dw}{\int_{n\cdot v<0} |n\cdot v|\, M(v)\, dv}. \end{aligned}$$

Since $0 \le \chi \le 1$, the fraction is always bounded by 1, which is seen by using the transformation $w = -v$ in the denominator and using the symmetry of the Maxwellian in (1.4). In the case that $0 < R \le 1$, we have $1-R < 1$ and $I < 1$. In the limiting case $R = 0$, we can still conclude that $I < 1$, because χ does not vanish almost everywhere in reasonable situations in the application, and hence the fraction will be less than 1 except in pathological cases.

The problem was solved numerically by choosing an expansion in velocity space of the form following [14, 16, 19, 20, 21, 22] and the references therein,

$$f(x,y,v,t) = \sum_{k=1}^{K} f_k(x,y,t)\varphi_k(v),$$

where the $\{\varphi_k(v), k = 1,\dots,K\}$ form an orthogonal set of basis functions with respect to the inner product

$$\langle \varphi_k, \varphi_\ell \rangle = \int \varphi_k(v)\varphi_\ell(v)\omega(v)\, dv = \delta_{k\ell},$$

with weight function $\omega(v) = 1/M(v)$. The basis functions are chosen as Maxwellians multiplied by (properly transformed) Hermite polynomials. To arrive at a Gaussian quadrature for the integrals, the discretization points in velocity space are roots of appropriate Hermite polynomials.

A Galerkin discretization by inserting the expansion for f and forming inner products with all basis functions then leads to the system of hyperbolic equations for the expansion coefficients $F = (f_k)$

$$\frac{\partial F}{\partial t} + A^{(1)} \frac{\partial F}{\partial x} + A^{(2)} \frac{\partial F}{\partial y} = CF,$$

with the $K \times K$ matrices $A^{(1)}$, $A^{(2)}$, and C with components

$$A^{(1)}_{k,\ell} = \langle v_1 \varphi_\ell, \varphi_k \rangle , \quad A^{(2)}_{k,\ell} = \langle v_2 \varphi_\ell, \varphi_k \rangle , \quad C_{k,\ell} = \langle Q(\varphi_\ell), \varphi_k \rangle .$$

We actually use an equivalent collocation basis for the Hermite polynomials, which results in the matrices $A^{(1)}$ and $A^{(2)}$ being diagonal; see [10, 13, 23] for more details. This system is solved by a finite-difference method using first-order upwinding and explicit time-stepping. The solution on the homogenized domain

$$\Omega_0 = (0,1) \times (0,1) \tag{3.4}$$

is straightforward (that was the point of the homogenization). The comparison solution on the microstructured domain Ω_ε is obtained by transforming the domain to the unit square. The transformation is designed such that it is an identity in the upper half of the domain, i.e., for $y \geq 0.5$; this is done in order to facilitate the comparison of both numerical solutions there without incurring additional interpolation error from a mesh transformation.

We discretize the velocity space by six basis functions in both x- and y-directions, resulting in a hyperbolic system of 36 equations. The maximum velocity is bounded by 4 in each direction. Using $\Delta x = 1/128$ and $\Delta y = 1/64$, the CFL condition requires a time step of $\Delta t = 1/1024$, accounting for an additional factor of about 4 from the transformation of the domain. Solutions are computed until the final time $t_{fin} = 2$. We compute solutions $f(x, y, v, t)$ with number densities denoted by $N(x, y, t)$ on domains Ω_ε in (3.2), $\varepsilon = 1/4$, $1/8$, $1/16$, and $1/32$. The solution to the homogenized problem is computed on Ω_0 in (3.4) and denoted by $\tilde{f}(x, y, v, t)$, with number density $\tilde{N}(x, y, t)$. Smaller values do not yield reliable solutions for the grid spacing used in the x-direction.

Figures 3.1, 3.2, 3.3, and 3.4 show comparisons of the number density $\tilde{N}(x, y, t)$ at $T = t_{fin} = 2$ on the homogenized domain to the number density $N(x, y, t)$ on the microstructured domain with the indicated values of ε. The plots show the physical behavior of the flow: The reacting chemical is supplied at the top of the domain at $y = 1$, then moves towards the wafer surface at $y = \tilde{h}(x)$, where it gets partially consumed in the surface reaction. Notice that the oscillations of the solution are limited to a boundary layer close to the microstructured surface; this effect is attributable to the smoothing property of the collision operator; that is, the solution is smoother than could be theoretically expected.

Table 3.1 shows the errors between the densities $\tilde{N}(x, y, t)$ and $N(x, y, t)$ of the homogenized and the original problems for various values of ε. They are compared only across the upper half of the domain, i.e., for $y \geq 0.5$, by choosing the subdomain $\tilde{\Omega} = (0,1) \times (0.5, 1)$ in the norms

$$\|N\|_{L^1(\tilde{\Omega})} := \iint_{\tilde{\Omega}} |N(x, y, t)| \, dx \, dy$$

and

$$\|N\|_{L^\infty(\tilde{\Omega})} := \max_{(x,y) \in \tilde{\Omega}} |N(x, y, t)|$$

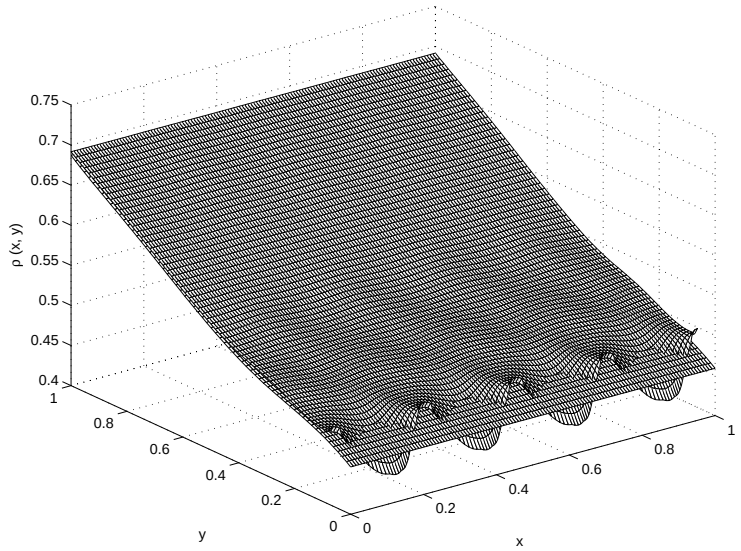


FIG. 3.1. *Comparison of the number densities on the homogenized and the microstructured domains with* $\varepsilon = 1/4$.

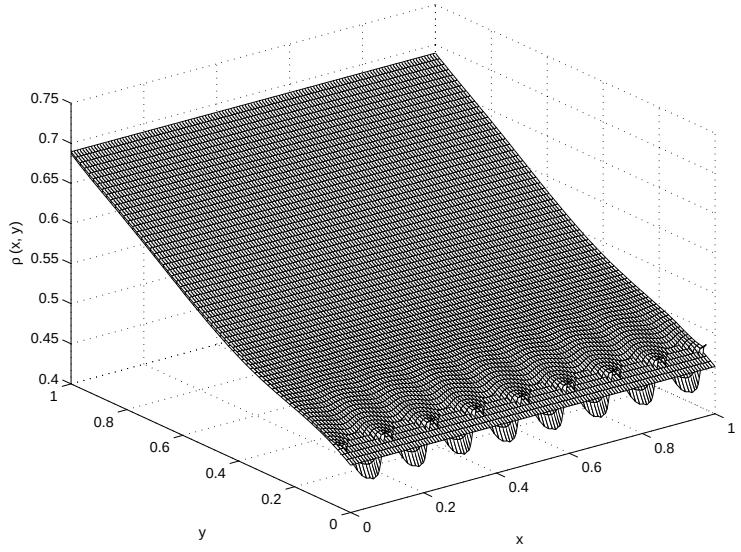


FIG. 3.2. *Comparison of the number densities on the homogenized and the microstructured domains with* $\varepsilon = 1/8$.

for a fixed time t. The table shows results at the final time $T = t_{fin} = 2$. Notice the decrease of all absolute as well as relative errors with ε.

Tables 3.2 and 3.3 study the underlying error in the density function f itself. More precisely, if $\tilde{f}(x, y, v, t)$ denotes the homogenized solution and $f(x, y, v, t)$ the solution

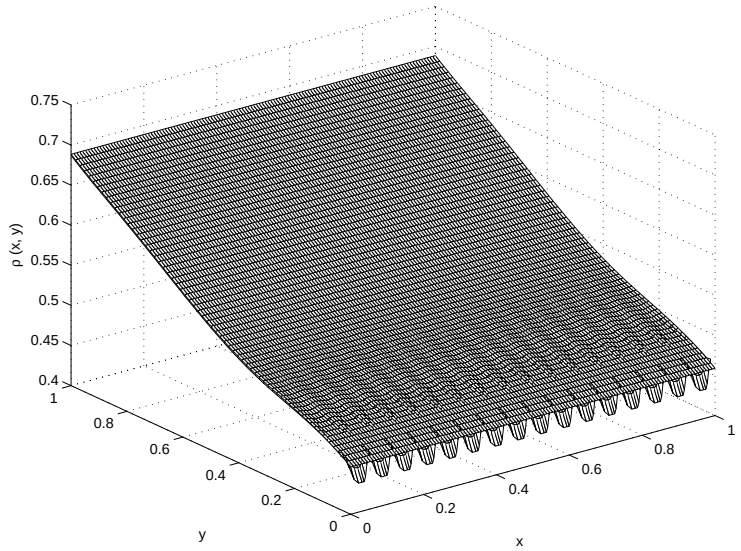


FIG. 3.3. *Comparison of the number densities on the homogenized and the microstructured domains with* $\varepsilon = 1/16$.

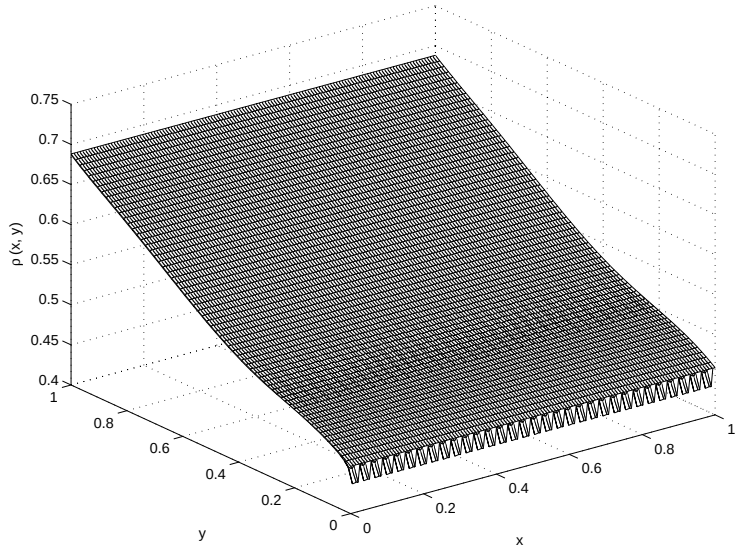


FIG. 3.4. *Comparison of the number densities on the homogenized and the microstructured domains with* $\varepsilon = 1/32$.

on the microstructured domain, then Table 3.2 lists the quantity $I(\tilde{f} - f)(y, v, T)$ with

$$I(f)(y, v, T) := \int_0^T \int_0^1 f(x, y, v, t)\, dx\, dt$$

with $T = t_{fin} = 2$ at $y = 0.875$ (close to the top of the domain) at various points

TABLE 3.1

Errors in density $\tilde{N}$ measured in various norms $\|\tilde{N}-N\|_{L^1(\tilde{\Omega})}$, $\|\tilde{N}-N\|_{L^1(\tilde{\Omega})}/\|N\|_{L^1(\tilde{\Omega})}$, $\|\tilde{N}-N\|_{L^\infty(\tilde{\Omega})}$, and $\|\tilde{N}-N\|_{L^\infty(\tilde{\Omega})}/\|N\|_{L^\infty(\tilde{\Omega})}$ on subdomain $\tilde{\Omega}=(0,1)\times(0.5,1)$ at $T=t_{fin}=2$.

	$\|\tilde{N}-N\|_{L^1}$	$\|\tilde{N}-N\|_{L^1}/\|N\|_{L^1}$	$\|\tilde{N}-N\|_{L^\infty}$	$\|\tilde{N}-N\|_{L^\infty}/\|N\|_{L^\infty}$
$\varepsilon=1/4$	0.0023932	0.0076664	0.0070134	0.0101409
$\varepsilon=1/8$	0.0013703	0.0044039	0.0031563	0.0045777
$\varepsilon=1/16$	0.0008171	0.0026308	0.0017929	0.0026041
$\varepsilon=1/32$	0.0005755	0.0018544	0.0012608	0.0018324

TABLE 3.2

Quantity $I(\tilde{f}-f)(y,v,T)$ with $T=t_{fin}=2$ at $y=0.875$ and velocity with $v_1=-0.6167$ and v_2 as listed.

v_2	−3.3243	−1.8892	−0.6167	0.6167	1.8892	3.3243
$\varepsilon=1/4$	6.8068e-08	4.7740e-06	5.6442e-05	7.2642e-04	3.8061e-04	1.1459e-05
$\varepsilon=1/8$	3.7021e-08	2.5996e-06	3.0958e-05	3.7582e-04	2.1973e-04	6.4414e-06
$\varepsilon=1/16$	2.3066e-08	1.6244e-06	1.9571e-05	2.2059e-04	1.4038e-04	3.6978e-06
$\varepsilon=1/32$	1.8086e-08	1.2793e-06	1.5657e-05	1.5784e-04	1.1535e-04	2.7049e-06

TABLE 3.3

Quantity $I(\tilde{f}-f)(y,v,T)/I(f)(y,v,T)$ with $T=t_{fin}=2$ at $y=0.875$ and velocity with $v_1=-0.6167$ and v_2 as listed.

v_2	−3.3243	−1.8892	−0.6167	0.6167	1.8892	3.3243
$\varepsilon=1/4$	3.4103e-05	5.8409e-05	1.5900e-04	9.9293e-03	2.8628e-02	3.6588e-02
$\varepsilon=1/8$	1.8549e-05	3.1806e-05	8.7218e-05	5.1617e-03	1.6729e-02	2.0902e-02
$\varepsilon=1/16$	1.1557e-05	1.9875e-05	5.5139e-05	3.0362e-03	1.0753e-02	1.2107e-02
$\varepsilon=1/32$	9.0618e-06	1.5652e-05	4.4112e-05	2.1743e-03	8.8528e-03	8.8848e-03

in velocity space given by $v_1=-0.6167$ and v_2 as listed in the table. This quantity mimics the behavior of $\iint \tilde{f}\,\psi\,dx\,dt$ with $\psi\equiv 1$. Table 3.3 shows the corresponding relative quantity $I(\tilde{f}-f)(y,v,T)\,/\,I(f)(y,v,T)$ at the same values of T, y, and v. The convergence rate is not uniform for all velocities. For those components f_k that correspond to velocities pointing towards the wafer surface, i.e., with $v_2<0$, the main portion of the information travels from the given inflow condition, and convergence is good. For the other components corresponding to information traveling back up from the wafer surface, i.e., with $v_2>0$, both absolute and relative errors deteriorate somewhat.

REFERENCES

[1] H. BABOVSKY, *Derivation of stochastic reflection laws from specular reflection*, Transport Theory Statist. Phys., 16 (1987), pp. 113–126.

[2] A. G. BELYAEV, *On Singular Perturbations of Boundary Problems*, Ph.D. thesis, Moscow State University, Moscow, 1990 (in Russian).

[3] T. S. CALE, T. H. GANDY, AND G. B. RAUPP, *A fundamental feature scale model for low pressure deposition processes*, J. Vac. Sci. Technol. A, 9 (1991), pp. 524–529.

[4] C. CERCIGNANI, *The Boltzmann Equation and Its Applications*, Appl. Math. Sci. 67, Springer-Verlag, New York, 1988.

[5] A. FRIEDMAN AND B. HU, *A non-stationary multi-scale oscillating free boundary for the Laplace and heat equations*, J. Differential Equations, 137 (1997), pp. 119–165.

[6] A. FRIEDMAN, B. HU, AND Y. LIU, *A boundary value problem for the Poisson equation with multi-scale oscillating boundary*, J. Differential Equations, 137 (1997), pp. 54–93.

[7] M. K. GOBBERT, T. S. CALE, AND C. A. RINGHOFER, *One approach to combining equipment scale and feature scale models*, in Process Control, Diagnostics, and Modeling in Semi-

conductor Manufacturing, M. Meyyappan, D. J. Economou, and S. W. Butler, eds., The Electrochemical Society Proceedings Series, vol. 95–4, 1995, pp. 553–563.

[8] M. K. Gobbert, T. S. Cale, and C. A. Ringhofer, *The combination of equipment scale and feature scale models for chemical vapor deposition via a homogenization technique*, VLSI Des., 6 (1998), pp. 399–403.

[9] M. K. Gobbert, T. P. Merchant, L. J. Borucki, and T. S. Cale, *A multiscale simulator for low pressure chemical vapor deposition*, J. Electrochem. Soc., 144 (1997), pp. 3945–3951.

[10] M. K. Gobbert and C. Ringhofer, *Mesoscopic scale modeling for chemical vapor deposition in semiconductor manufacturing*, in Dispersive Transport Equations and Multiscale Models, N. B. Abdallah, A. Arnold, P. Degond, I. Gamba, R. Glassey, C. D. Levermore, and C. Ringhofer, eds., IMA Volumes in Mathematics and its Applications 136, Springer-Verlag, New York, 2003, pp. 133–150.

[11] M. K. Gobbert and C. A. Ringhofer, *An asymptotic analysis for a model of chemical vapor deposition on a microstructured surface*, SIAM J. Appl. Math., 58 (1998), pp. 737–752.

[12] M. K. Gobbert, C. A. Ringhofer, and T. S. Cale, *Mesoscopic scale modeling of microloading during low pressure chemical vapor deposition*, J. Electrochem. Soc., 143 (1996), pp. 2624–2631.

[13] M. K. Gobbert, S. G. Webster, J.-F. Remacle, and T. S. Cale, *A Spectral Galerkin Ansatz for the Deterministic Solution of the Boltzmann Equation on Irregular Domains*, Technical report, University of Maryland, Baltimore County, Baltimore, MD, 2002.

[14] H. Grad, *On the kinetic theory of rarefied gases*, Comm. Pure Appl. Math., 2 (1949), pp. 331–407.

[15] A. Kersch and W. J. Morokoff, *Transport Simulation in Microelectronics*, Progress in Numerical Simulation for Microelectronics, 3, Birkhäuser Verlag, Basel, 1995.

[16] A. Klar, *A numerical method for kinetic semiconductor equations in the drift-diffusion limit*, SIAM J. Sci. Comput., 20 (1999), pp. 1696–1712.

[17] C. R. Kleijn and C. Werner, *Modeling of Chemical Vapor Deposition of Tungsten Films*, Progress in Numerical Simulation for Microelectronics, 2, Birkhäuser Verlag, Basel, 1993.

[18] T. P. Merchant, M. K. Gobbert, T. S. Cale, and L. J. Borucki, *Multiple scale integrated modeling of deposition processes*, Thin Solid Films, 365 (2000), pp. 368–375.

[19] C. Ringhofer, *Computational methods for semiclassical and quantum transport in semiconductor devices*, Acta Numer., 6 (1997), pp. 485–521.

[20] C. Ringhofer, *Space-time discretization of series expansion methods for the Boltzmann transport equation*, SIAM J. Numer. Anal., 38 (2000), pp. 442–465.

[21] C. Ringhofer, C. Schmeiser, and A. Zwirchmayr, *Moment methods for the semiconductor Boltzmann equation on bounded position domains*, SIAM J. Numer. Anal., 39 (2001), pp. 1078–1095.

[22] C. Schmeiser and A. Zwirchmayr, *Convergence of moment methods for linear kinetic equations*, SIAM J. Numer. Anal., 36 (1998), pp. 74–88.

[23] S. G. Webster, M. K. Gobbert, J.-F. Remacle, and T. S. Cale, *Parallel numerical solution of the Boltzmann equation for atomic layer deposition*, in Euro-Par 2002 Parallel Processing, B. Monien and R. Feldmann, eds., Lecture Notes in Comput. Sci. 2400, Springer-Verlag, Berlin, 2002, pp. 452–456.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 216–230

NONLINEAR STABILITY OF A LATITUDINAL RING OF POINT-VORTICES ON A NONROTATING SPHERE*

STEFANELLA BOATTO† AND HILDEBERTO E. CABRAL‡

Abstract. We study the nonlinear stability of relative equilibria of configurations of identical point-vortices on the surface of a sphere. In particular, we study how the stability changes as a function of the colatitude θ and of the number of vortices N. By using the integrals of motion, we view the system in a suitable corotating frame where the polygonal vortex configuration is at rest. Then after a sufficient criterion due to Dirichlet, the stability ranges are the θ-intervals for which the Hessian of the Hamiltonian—evaluated at the equilibrium configuration—is positive or negative definite. We find that the stability intervals coincide with those for linear stability determined by Polvani and Dritschel [*J. Fluid Mech.*, 255 (1993), pp. 35–64]. For $N = 3$ we recover the result previously established by Pekarsky and Marsden [*J. Math. Phys.*, 39 (1998), pp. 5894–5907].

Key words. stability, vortex dynamics, relative equilibria

AMS subject classifications. 76, 37, 34

DOI. 10.1137/S0036139902399965

1. Introduction.

In this article, by using a very clear and simple method, we complete a stability analysis started by Thomson over a century ago [35, 36]. In 1883, while studying and modeling the atomic structure, Thomson investigated the *linear* stability of corotating point-vortices in the plane. In particular, his interest was in configurations of identical vortices equally spaced along the circumference of a circle, i.e., located at the vertices of a regular polygon. He proved that for six or fewer vortices the polygonal configurations are stable, while for seven vortices—the Thomson heptagon—he erroneously concluded that the configuration is slightly unstable [26]. It took more than a century to make some progress on this problem. In his Ph.D. thesis, Dritschel succeeded in solving the aspect of the heptagon mystery concerned with its linear stability analysis, leaving open the nonlinear stability question: he proved that the Thomson heptagon is neutrally stable and that for eight or more vortices the corresponding polygonal configurations are linearly unstable [13]. In 1993, Polvani and Dritschel generalized the techniques used in [13] to study the linear stability of a "latitudinal" ring of point vortices (see Figure 1) on the sphere [31], a more relevant problem from the atmospheric modeling point of view. They proved that polygonal configurations are more unstable on the sphere than in the plane. In particular, they showed that at the pole, for $N < 7$ the configuration is stable, for $N = 7$ it is neutrally stable, and for $N > 7$ it is unstable. The ranges of linear stability in the colatitude θ, as a function of N, are summarized in Table 1.

*Received by the editors January 22, 2002; accepted for publication (in revised form) January 10, 2003; published electronically November 19, 2003.

http://www.siam.org/journals/siap/64-1/39996.html

†Astronomie et Systèmes Dynamiques, Institut de Mécanique Céleste, 77 Avenue Denfert-Rochereau, Paris, France, and Département de Mathématiques, Université Paris 13, 99 Av. J.-B. Clément, 93430 Villetaneuse, France. Current address: The Fields Institute for Research in Mathematical Sciences, 222 College Street, Toronto, ON, M5T 3J1 Canada (sboatto@fields.utoronto.ca). The research of this author was supported by the CAPES visiting professor fellowship, via Universidade Federal de Pernambuco, and by the Agreement Brazil/France in Mathematics—Proc. 69.0014/01-5.

‡Departamento de Matemática, Universidade Federal de Pernambuco, Cidade Universitária, CEP 50740-540 Recife, PE, Brazil, and Department of Mathematical Sciences, University of Cincinnati, P.O. Box 210025, Cincinnati, OH 45221-0025 (hild@dmat.ufpe.br).

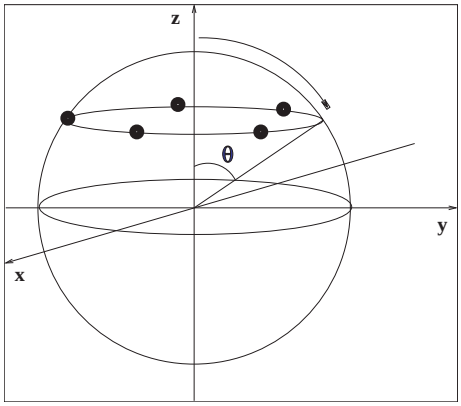


FIG. 1. *Latitudinal ring of point-vortices.*

TABLE 1
Regular polygonal configurations of N vortices. Stability ranges as a function of the colatitude θ and of the number of vortices N.

Number of vortices (N)	Colatitude (θ)
3	$0° \leq \theta \leq 90°$
4	$0° \leq \theta \leq 55°$
5	$0° \leq \theta \leq 45°$
6	$0° \leq \theta \leq 27°$
7	$\theta = 0°$

In 1998, Kidambi and Newton fully studied the motion of three vortices on the sphere and gave a geometrical interpretation of the conserved quantities [18]. By means of the energy momentum method (the Marsden–Meyer–Weistein reduction), Pekarsky and Marsden studied the nonlinear stability analysis for the integrable case of polygonal configurations of three vortices of arbitrary vorticities (k_1, k_2, and k_3) on the sphere, leaving open the stability analysis for nonintegrable vortex systems ($N > 3$) [30]. More recently Cabral and Schmidt completed the linear and nonlinear stability analysis at once for polygonal configurations in the plane [10], leaving untouched the analogous analysis on the sphere. They proved that for seven or fewer vortices the polygonal configurations are *nonlinearly* stable in the plane and, beyond that, they studied the stability of polygonal configurations of identical vortices of strength $k = 1$ and with a central vortex of arbitrary strength K (see Figure 2). They determined the range of stability as a function of K (see Table 2 and [10, Theorems 5.1 and 7.1]) to be

$$\begin{cases} (N^2 - 8N + 8)/16 < K < (N-1)^2/4 & \text{when } N \text{ is even,} \\ (N^2 - 8N + 7)/16 < K < (N-1)^2/4 & \text{when } N \text{ is odd.} \end{cases} \tag{1}$$

The stability of the heptagon, being still a special case, had to be studied by us-

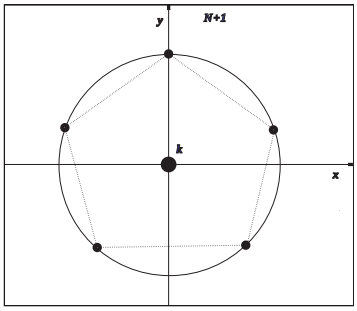


FIG. 2. *Centered pentagonal configuration.*

TABLE 2

Vortex configurations with N identical vortices at the vertices of a regular polygon plus a central vortex of strength K. As a function of N, intervals of vortex strength K assure nonlinear stability of the centered polygonal configurations.

Number of vortices (N)	Stability range for K (the strength of the central vortex)
3	$-0.5 < K < 1.00$
4	$-0.5 < K < 2.25$
5	$-0.5 < K < 4.00$
6	$-0.25 < K < 6.25$
7	$0 = K < 9$

ing normal form reduction and considering higher order terms in the Hamiltonian (see [10]).

In 2001, by means of group theory techniques, Lim, Montaldi, and Roberts exhaustively classified the relative equilibria of a system of identical point-vortices on a sphere [23]. In 2002, Laurent-Polz [21] studied the system formed by $2N$ point-vortices with N vortices of strength $+1$ and N vortices of strength -1. He proved the existence of some fixed and relative equilibria and studied their stability by means of the energy momentum method, as in [30]. Soulière and Tokieda extended the study of periodic vortex motion to various manifolds with symmetry [34], and Montaldi, Soulière, and Tokieda fully determined fixed and relative equilibrium configurations for vortices on the cylinder [25]. We would like to point out that many people are working in this area, both for its mathematical beauty and open questions, and for its great interest as a modeling problem for the vortex dynamics of the earth's atmosphere [26, 31, 11, 14]. We refer the reader to the nice book by Newton for a review on N-vortex dynamics [27] and to the article by H. Aref et al. for a quite complete and updated review on relative vortex configurations [3].

In this article we accomplish the stability analysis for integrable ($N = 3$) and nonintegrable ($3 < N \leq 7$) polygonal configurations of identical point-vortices on the sphere. We introduce a much simpler method than in [10, 30, 21] that at one stroke furnishes a complete linear and nonlinear analysis for the spherical case and that,

moreover, can easily be generalized to an analogous stability analysis for polygonal vortex configurations in other geometries [5, 8], including the planar case previously studied by Cabral and Schmidt [10]. More specifically, by using a *sufficient* criterion due to Dirichlet, we derive the stability ranges as θ-intervals for which the Hessian of the Hamiltonian—evaluated at the equilibrium configuration—is positive or negative definite. At this point we would like to remind the reader that, in general, linear and nonlinear stability ranges may not coincide. As is well known, a system can be linearly stable or unstable and nonlinearly stable or unstable; all four possibilities occur in practice [24, 33, 9]. As a result of our analysis we find that the nonlinear stability ranges (with the exception of the equator where other techniques need to be used to infer stability) coincide with the ranges of linear stability previously determined by Polvani and Dritschel (see Table 2).

Before ending this introduction we would like to make the following remarks:

(a) Relative equilibrium configurations are so called because the motion vanishes in an appropriate rotating frame. The study of such configurations was termed "vortex statics" by Kelvin in 1910 [17], who, with Thomson [35, pp. 94–108], found the relative equilibria of identical vortices. If all circulations have the same sign, relative equilibria are the only stationary configurations possible. Palmore [29] investigated this case and found that many relative equilibria must occur (for more details, see [28, 3]).

(b) When dealing with a system with integrable vortex dynamics (such as three-vortex systems) we can simply make a canonical reduction to a system of one degree of freedom. Therefore relative equilibrium configurations and their stability can be simply inferred by determining the maxima and the minima of the reduced vortex Hamiltonian. For a complete analysis, including the general case of vortices with different vorticities, and for further details about the global dynamics of integrable vortex systems, see Boatto and Laskar [6].

(c) By analogy with the planar case (see Figure 2 and Table 2), in a follow-up article [5] we are generalizing this nonlinear stability analysis to the case of latitudinal polygonal vortex configurations with vortices of strengths k_1 and k_2 at the poles. From the atmospheric dynamics point of view, the central vortex is a model for a polar vortex (southern or northern polar vortex), and "latitudinal" chains of point-vortices are models for planetary waves. Therefore this kind of analysis could help answer questions of the type: Does the presence of a Polar vortex (in the southern or in the northern hemisphere) favor the presence of waves (jets) at a given latitude? To show how the presence of a central polar vortex of strength K dramatically changes the stability ranges, at the end of section 3 we report the results of the stability analysis for the case $N = 3$, for which we obtain fully analytical expressions.

The article is organized as follows. In section 2 we derive the equations of motion, in particular the vortex Hamiltonian, and show the differences between the planar and the spherical cases. In section 3 we summarize the main results and the main tools for proving stability—the Dirichlet theorem on stability and a theorem on positive (negative) quadratic forms (see Gel´fand [15, Theorem 1, p. 49]). In section 4 we derive the stability ranges explicitly.

2. Equations of motion. The starting hypothesis for our derivation is that the fluid can be modeled as a *two-dimensional incompressible* fluid with constant density, i.e., a fluid whose velocity field verifies

$$\nabla \cdot \mathbf{u} = 0, \tag{2}$$

with $\mathbf{u} = (u_x, u_y, u_z) = (\dot{x}, \dot{y}, \dot{z})$. In atmospheric dynamics, a fluid verifying (2) represents the simplest fluid model, called the barotropic model [11]. We are particularly interested in characterizing the fluid dynamics for a given vorticity field, ω:

$$\omega = \nabla \times \mathbf{u}. \tag{3}$$

Regarding this, we notice that in two dimensions our task is simplified since we can recast the fluid equations into a Hamiltonian formalism. In fact, notice that on the plane $\mathbf{u} = (\dot{x}, \dot{y})$, and (2) is still verified if we represent the velocity components (see [19]) as

$$\dot{x} = \frac{\partial \Psi}{\partial y}, \qquad \dot{y} = -\frac{\partial \Psi}{\partial x}, \tag{4}$$

i.e., by means of Ψ, called the *stream function*. Formally Ψ plays the role of a Hamiltonian for the pair of conjugate variables (x, y). By substituting (4) into (3), we obtain

$$\Delta \Psi(\mathbf{r}) = \omega(\mathbf{r}), \tag{5}$$

i.e., a Poisson equation with ω as a source term. Then, once we specify the vorticity field, by inverting (5) we obtain the stream function Ψ as

$$\Psi(\mathbf{r}) = \iint G(\mathbf{r}, \mathbf{r}')\omega(\mathbf{r}) dA', \tag{6}$$

where $G(\mathbf{r}, \mathbf{r}')$ is the Green function. The Green function, both for the plane and for the sphere, is (see [11, 19])

$$G(\mathbf{r}, \mathbf{r}') = -\frac{1}{4\pi} \log \parallel \mathbf{r} - \mathbf{r}' \parallel^2 .$$

Then by (6), once we specify the vorticity field $\omega(\mathbf{r})$, we can compute Ψ, and by (4) we know the velocity field everywhere.

2.1. Point-vortices on the plane. In our specific case we are interested in the simplest of all vorticity fields: a system of point-vortices. A point-vortex can be thought of as an entity in which the vorticity is concentrated into a point. Therefore in the *plane* the vorticity for a system of N point-vortices is

$$\omega(\mathbf{r}) = \sum_{\alpha=1}^{N} k_\alpha \delta(\mathbf{r} - \mathbf{r}_\alpha), \tag{7}$$

where k_α, $\alpha = 1, \ldots, N$, is constant and corresponds to the vorticity of the αth vortex. Therefore by applying the inversion formula (6), the stream function is

$$\Psi(\mathbf{r}) = -\frac{1}{4\pi} \sum_{\alpha=1}^{N} k_\alpha \log \parallel \mathbf{r} - \mathbf{r}_\alpha \parallel^2 . \tag{8}$$

This equation describes, together with (4), the dynamics of a test particle at a point (x, y) in the plane. Analogously, it can be shown that the dynamics of a system of point-vortices in the plane is given by the equations

$$k_\alpha \frac{dX_\alpha}{dt} = \frac{\partial H_p}{\partial Y_\alpha}, \qquad k_\alpha \frac{dY_\alpha}{dt} = -\frac{\partial H_p}{\partial X_\alpha}, \tag{9}$$

where $(q_\alpha, p_\alpha) = (X_\alpha, k_\alpha Y_\alpha)$, $\alpha = 1, \ldots, N$, is a pair of conjugate variables and H_p is the generalization of the stream-function Ψ (8) for the vortex system,

$$H_p = -\frac{1}{4\pi} \sum_{\alpha<\beta} k_\alpha k_\beta \log \parallel \mathbf{r}_\alpha - \mathbf{r}_\beta \parallel^2 . \tag{10}$$

Notice that in addition to the Hamiltonian H_p, a system of point-vortices in the plane has the integrals of motion

$$L = \sum_{\alpha=1}^{N} k_\alpha \parallel \mathbf{r}_\alpha \parallel^2, \quad P_x = \sum_{\alpha=1}^{N} k_\alpha X_\alpha, \quad P_y = \sum_{\alpha=1}^{N} k_\alpha Y_\alpha,$$

expressing, respectively, the conservation of angular momentum and linear momentum in space. Furthermore, by introducing the Poisson bracket

$$[f, g] = \sum_{\alpha=1}^{N} \left(\frac{\partial f}{\partial q_\alpha} \frac{\partial g}{\partial p_\alpha} - \frac{\partial f}{\partial p_\alpha} \frac{\partial g}{\partial q_\alpha} \right) = \sum_{\alpha=1}^{N} \frac{1}{k_\alpha} \left(\frac{\partial f}{\partial X_\alpha} \frac{\partial g}{\partial Y_\alpha} - \frac{\partial f}{\partial Y_\alpha} \frac{\partial g}{\partial X_\alpha} \right),$$

we can show that we can construct three integrals in involution out of the four conserved quantities H_p, L, P_y, and P_y. These are H_p, $P_x^2 + P_y^2$, and L; in fact,

$$[H_p, P_x^2 + P_y^2] = 0, \quad [H_p, L] = 0, \quad [P_x^2 + P_y^2, L] = 0.$$

It is then possible to reduce the system of equations from N to $N - 2$ degrees of freedom. It follows that a system with $N \leq 3$ is integrable, whereas the system of equations of four vortices has been shown by Ziglin to be nonintegrable in the sense that there are no other first integrals analytically dependent on the coordinates and circulations, and functionally independent of L, H_v, P_x, P_y (see [16]). It has been shown, however, that a system of four identical vortices (i.e., $k_\alpha = k$ for $\alpha = 1, \ldots, 4$) can undergo periodic or quasi-periodic motion for special initial conditions. More specifically, the motion of a system of four identical vortices can be periodic, quasi-periodic, or chaotic, depending on the symmetry of the initial configuration [2, 4, 7].

The simplest relative equilibrium is given by a system of two vortices with vorticity of the same sign. For such a system the vortex motion is always periodic, and the orbits are always circles. We do not have the option, as in the Kepler problem, where the two masses can describe elliptical orbits. For further comments comparing vortex dynamics with celestial mechanics, see [1, 27, 6].

2.2. Point-vortices on a sphere. When considering point-vortices on the surface of the sphere, an additional constraint is given by the sphere topology. In fact, the sphere being a compact manifold and, more specifically, a manifold with no boundaries, the Green theorem gives us a constraint on the distribution of vorticity,

$$\iint_S \omega(\mathbf{r}') dA' = \iint_S \nabla \cdot \nabla \Psi(\mathbf{r}') dA' = 0;$$

i.e., the vorticity distribution must have a zero average on the sphere [11]. To fulfill this requirement, it is sufficient to include, in addition to the N point-vortices delta-distributed, an evenly spread constant vorticity Γ, whose magnitude is equal to minus the average of the point-vortices over the sphere, i.e.,

$$\omega(\mathbf{r}) = \sum_{\alpha}^{N} k_\alpha \delta(\mathbf{r} - \mathbf{r}_\alpha) + \Gamma, \tag{11}$$

with $\Gamma = -\sum_{\alpha=1}^{N} k_\alpha/4\pi R^2$, R being the radius of the sphere, $\mathbf{r}$ and $\mathbf{r}_\alpha$ designating, respectively, a generic point and the location of the αth vortex on the sphere. Then, for this case, the vortex Hamiltonian is found to be

$$H_s = \frac{1}{4\pi R^2} \sum_{\alpha<\beta} k_\alpha k_\beta \log \parallel \mathbf{r}_\alpha - \mathbf{r}_\beta \parallel^2, \tag{12}$$

and the corresponding conjugate variables are $p_j = k_j \cos(\theta_j)$ and $q_j = \varphi_j$, $j = 1, \ldots, N$, where φ_j and θ_j are the usual angles of spherical coordinates, respectively the longitude and the colatitude, and k_j is the vortex strength of the jth vortex. Then the dynamics of N point-vortices on the surface of a sphere is given by the equations (see [18, 31])

$$k_j \dot{p}_j = \frac{\partial H_s}{\partial q_j}, \qquad k_j \dot{q}_j = -\frac{\partial H_s}{\partial p_j}.$$

As for the planar case [7], in addition to the Hamiltonian H_s, a system of point-vortices on a sphere has the integrals of motion

$$L = \sum_{\alpha=1}^{N} k_\alpha \parallel \mathbf{r}_\alpha \parallel^2, \quad P_x = \frac{1}{R} \sum_{\alpha=1}^{N} k_\alpha x_\alpha, \quad P_y = \frac{1}{R} \sum_{\alpha=1}^{N} k_\alpha y_\alpha, \quad P_z = \frac{1}{R} \sum_{\alpha=1}^{N} k_\alpha z_\alpha,$$

expressing, respectively, the conservation of angular momentum and linear momentum in the space. Notice that, in the case of vortices on a sphere, the integral of motion L simplifies to $L = R^2 \sum_{\alpha=1}^{N} k_\alpha$ and is therefore redundant. Furthermore, as in section 2.1, by introducing the Poisson bracket $[f, g] = \sum_{\alpha=1}^{N} (\frac{\partial f}{\partial q_\alpha} \frac{\partial g}{\partial p_\alpha} - \frac{\partial f}{\partial p_\alpha} \frac{\partial g}{\partial q_\alpha})$, it can easily be shown that we can construct three integrals in involution out of the four conserved quantities H_s, P_x, P_y, and P_z. These are H_s, $P_x^2 + P_y^2$, and P_z; in fact,

$$[H_s, P_x^2 + P_y^2] = 0, \quad [H_s, P_z] = 0, \quad [P_x^2 + P_y^2, P_z] = 0.$$

It is then possible to reduce the system of equations from N to $N - 2$ degrees of freedom. It follows that a system with $N \leq 3$ is integrable. Notice that while in the planar case [7] we had the freedom of choosing a reference frame within which the origin coincides with the center of vorticity, so that $P_x = P_y = 0$, in the spherical case there is a privileged point in space, i.e., the center of the sphere. In this case, the momentum vector $\mathbf{M} = (P_x, P_y, P_z)$ plays a special role in characterizing the dynamics of the vortex system. We then use the freedom in the axis orientations to select a reference system with the z-axis in the direction of $\mathbf{M}$ (see [18]). This choice still gives $P_x = P_y = 0$, as for the planar case. In what follows there is no loss of generality in considering a sphere of radius one ($R = 1$), and since we consider the case of identical vortices—$k_\alpha = k$, $\alpha = 1, \ldots, N$—we choose $k = 1$. Then the Hamiltonian becomes

$$H_s = \frac{1}{4\pi} \sum_{\alpha<\beta} \log \parallel \mathbf{r}_\alpha - \mathbf{r}_\beta \parallel^2 . \tag{13}$$

We then use the rotational symmetry of the Hamiltonian (12) to view the vortex dynamics in an appropriate corotating frame. In fact, by considering the canonical

transformation

$$\text{(14)}\qquad \begin{cases} \phi_1 = \varphi_1 - \varphi_N & J_1 = p_1 \\ \phi_2 = \varphi_2 - \varphi_N & J_2 = p_2 \\ \vdots & \vdots \\ \phi_{N-1} = \varphi_{N-1} - \varphi_N & J_{N-1} = p_{N-1} \\ \phi_N = \varphi_N & J_N = \sum_{k=1}^{N} p_k, \end{cases}$$

we reduce the problem by one degree of freedom: ϕ_N is a cyclic coordinate, and J_N is the momentum vector $\mathbf{M}$.

3. Main results. To determine the nonlinear stability regions for a system of N identical point-vortices, we use a sufficient condition for stability which may be traced back to Lagrange for potential systems and was later proved by Dirichlet [12] for systems with integrals; this was subsequently generalized by Lyapunov in his direct method (section 3.1). To apply the Dirichlet criterion we need to establish whether the Hessian of the Hamiltonian is positive or negative definite. For this purpose we use the Jacobi method described in section 3.2.

As result of our analysis, we find that the nonlinear stability ranges coincide—apart from the equatorial plane—with the linear stability ranges previously found by Polvani and Dritschel [31], as shown in Table 2. Detailed calculations are done in section 4.

3.1. Nonlinear stability: The Dirichlet criterion. Let x^* be an equilibrium of an autonomous system of ordinary differential equations,

$$\text{(15)}\qquad \frac{dx}{dt} = f(x), \quad x \in \Omega \subset \mathbb{R}^n,$$

that is, $f(x^*) = 0$. Denote by $\phi(t, x)$ the solution of (15) such that $\phi(0, x) = x$. For $r > 0$, denote by $B_r(x)$ the open ball in $\mathbb{R}^n$ with center x and radius r.

Let us recall that a function Ψ is said to be positive definite (resp., negative definite) at a point x^* if $f(x^*) = 0$ and if there exists a neighborhood V of x^* such that $f(x) > 0$ (resp., $f(x) < 0$) for all $x \neq x^*$ in V.

THEOREM 1 (Dirichlet; see [32]). *If there exists a positive (or negative) definite integral Ψ of the system* (15) *in a neighborhood of the equilibrium x^*, then x^* is stable.*

For an autonomous Hamiltonian system, Theorem 1 translates into the following. Let $z^* = 0$ be an equilibrium of the system with an analytic Hamiltonian

$$H(z) = H_2(z) + H_3(z) + \cdots,$$

where $H_k(z)$ is a homogeneous polynomial of degree k in z.

Since H is a first integral of the system, Dirichlet's theorem implies that if, at the equilibrium the quadratic form $H_2(z)$ is positive (or negative) definite, then the equilibrium is stable.

3.2. A theorem on positive (negative) quadratic forms: The method of Jacobi. To determine whether the Hessian of the Hamiltonian is positive or negative definite, we make use of the Jacobi method illustrated in the following theorem (see Gel´fand [15, Theorem 1, p. 49]).

THEOREM 2. *Let $A(x;x)$ be a quadratic form defined relative to some basis $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ by the equation*

$$A(x;x) = \sum_{i,k=1}^{n} a_{ik}\eta_i\eta_k, \qquad a_{ik} = A(\mathbf{f}_i;\mathbf{f}_k),$$

where $\eta_1, \ldots, \eta_n$ are coordinates of x in this basis. Further, let the determinants

$$\Delta_1 = a_{11}, \qquad \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \qquad \ldots,$$

$$\Delta_n = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{11} & a_{12} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}$$

all be different from zero. Then there exists a basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ relative to which $A(x;x)$ is expressed as a sum of squares,

$$A(x;x) = \frac{\Delta_o}{\Delta_1}\zeta_1^2 + \frac{\Delta_1}{\Delta_2}\zeta_2^2 + \cdots + \frac{\Delta_{n-1}}{\Delta_n}\zeta_n^2.$$

Here $\Delta_0 = 1$ and $\zeta_1, \zeta_2, \ldots, \zeta_n$ are the coordinates of x in the basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

4. The Hessian of the Hamiltonian and its principal minors. In this section we simply illustrate the calculations to determine, as a function of N, the region for which the Hamiltonian is negative or positive definite when evaluated at the equilibrium configuration. We remind the reader that, after having performed the canonical transformation (14), we evaluate the principal minor determinants of the Hessian matrix of the Hamiltonian at the relative equilibrium configurations. As announced in section 2, the Hamiltonian is the reduced Hamiltonian for which the vortex dynamics is viewed in the reference system rotating with the relative equilibrium configuration (i.e., with a frequency $\dot{\varphi}_N$).

In the reduced system the Hessian of the Hamiltonian takes the form

$$\mathcal{H}(H_s) = \left(\begin{array}{c|c} \frac{\partial^2 H_s}{\partial\phi_\alpha\partial\phi_\beta} & \frac{\partial^2 H_s}{\partial J_\alpha\partial\phi_\beta} \\ \hline \frac{\partial^2 H_s}{\partial\phi_\alpha\partial J_\beta} & \frac{\partial^2 H_s}{\partial J_\alpha\partial J_\beta} \end{array}\right), \qquad \alpha,\beta = 1,\ldots,N-1,$$

which, evaluated at the equilibrium configuration—$J_{o\alpha} = z_o = \cos\theta_o$ and $\phi_{o\alpha} = 2\pi\,\alpha/N$ for $\alpha = 1,\ldots,N-1$—becomes

$$\mathcal{H}(H_s)_{eq} = b\left(\begin{array}{c|c} \tilde{S}_o - S_o\mathbb{I} & 0 \\ \hline 0 & \frac{\partial^2 H_s}{\partial J_\alpha\partial J_\beta} \end{array}\right), \tag{16}$$

where $b = 1/2\pi$, $\mathbb{I}$, and 0 are, respectively, the $((N-1)\times(N-1))$ identity and the zero matrices,

$$S_o = \sum_{\alpha=1}^{N-1} \frac{1}{1-\cos(\phi_{o\alpha})} = \frac{1}{6}(N^2-1), \qquad \text{with} \quad \phi_{o\alpha} = \frac{2\pi}{N}\alpha,$$

and $\tilde{S}_o$ is a symmetric matrix with zero-diagonal elements and with off-diagonal elements

$$(\tilde{S}_o)_{\alpha\beta} = \frac{1}{1-\cos(\phi_{o\alpha}-\phi_{o\beta})}, \qquad \text{with} \quad \phi_{o\alpha} = \frac{2\pi}{N}\alpha, \quad \phi_{o\beta} = \frac{2\pi}{N}\beta.$$

Then for the determinants of principal minors of (16) we obtain the following:

Case 1. $N = 2$:

$$\Delta_1 = -\frac{1}{2}, \qquad \Delta_2 = +\frac{z_o^2}{2r_o^4}.$$

Then clearly, $\Delta_1 < 0$, $\Delta_2 > 0$ for all $z_o \neq \{0, 1\}$, i.e., for all colatitudes but the equatorial one, and at the poles since $z_o = \cos\theta$. By Theorem 2 (section 3), we can conclude that the Hessian (16) is negative definite at all colatitudes θ, with the exception of the values $\theta = 90^\circ$ (i.e., at the equator) and $\theta = 0^\circ$ (i.e., at the poles). Then by Theorem 1 we can infer that "colatitudinal" relative equilibrium configurations of two vortices are always nonlinearly stable, with the exception of the values $\theta = 90^\circ$ and $\theta = 0^\circ$.

Case 2. $N = 3$:

$$\Delta_1 = -\frac{4}{3}, \qquad \Delta_2 = \frac{4}{3}, \qquad \Delta_3 = -\frac{16z_o^2}{r_o^4}, \qquad \Delta_4 = \frac{16\, z_o^4}{r_o^8}.$$

Then, $\Delta_1 < 0$, $\Delta_2 > 0$, $\Delta_3 < 0$, and $\Delta_4 > 0$ for all $z_o \neq \{0, 1\}$, i.e., for all colatitudes but the equatorial one and at the poles. Then by the same argument as for Case 1 ($N = 2$), we conclude that "colatitudinal" relative equilibrium configurations of three vortices are always nonlinearly stable. The case $\theta = 90^\circ$ needs further analysis; for more details, see Pekarsky and Marsden [30].

Case 3. $N = 4$:

$$\Delta_1 = -5, \qquad \Delta_2 = \frac{21}{4}, \qquad \Delta_3 = -9, \qquad \Delta_4 = 9\frac{6z_o^2-1}{r_o^4},$$

$$\Delta_5 = -27\frac{z_o^2(9z_o^2-2)}{r_o^8}, \quad \text{and} \quad \Delta_6 = 324\frac{z_o^4(3z_o^2-1)}{r_o^{12}}.$$

Then by symmetry restricting the analysis to the domain $0^\circ \le \theta \le 90^\circ$, we find the following:

$$\Delta_4 > 0 \quad \text{if} \quad 0^\circ < \theta \le \arccos\left(\frac{1}{\sqrt{6}}\right) \approx 66^\circ,$$

$$\Delta_5 < 0 \quad \text{if} \quad 0^\circ < \theta \le \arccos\left(\frac{\sqrt{2}}{3}\right) \approx 62^\circ,$$

$$\Delta_6 > 0 \quad \text{if} \quad 0^\circ < \theta \le \arccos\left(\frac{1}{\sqrt{3}}\right) \approx 55^\circ.$$

Then, by the same argument as for the case $N = 2$, we conclude that the Hessian is negative definite if $0° < \theta \leq \arccos(1/\sqrt{3})$; i.e., nonlinear stability holds in the region extending from the pole down to a latitude of about 35°.

Case 4. $N = 5$:

$$\Delta_1 = -4, \quad \Delta_2 = 0.4(37 - \sqrt{5}) > 0, \quad \Delta_3 = -1.92(25 - \sqrt{5}) < 0, \quad \Delta_4 = 115.2,$$

$$\Delta_5 = -46.08\frac{20z_o^2 - 5 - \sqrt{5}}{r_o^4}, \quad \Delta_6 = \frac{1}{5}\frac{(20z_o^2 - 5 + \sqrt{5})(12z_o^2 - \sqrt{5} - 3)}{r_o^8},$$

$$\Delta_7 = -32\,\Delta_4\frac{z_o^2}{r_o^{12}}(2z_o - 1)^2, \quad \text{and} \quad \Delta_8 = \Delta_4\,320\frac{(2z_o^2 - 1)^2 z_o^4}{r_o^{16}}.$$

Then we find the following:

$$\Delta_5 < 0 \quad \text{if} \quad 0° \leq \theta \leq \arccos\left(\frac{\sqrt{25 + 5\sqrt{5}}}{10}\right) \approx 53°,$$

$$\Delta_6 > 0 \quad \text{if} \quad 0° \leq \theta \leq \arccos\left(\sqrt{\frac{3 + \sqrt{5}}{12}}\right) \approx 48°,$$

$$\Delta_7 < 0 \quad \text{if} \quad \theta \neq \arccos\left(\frac{1}{\sqrt{2}}\right) = 45°,$$

$$\Delta_8 > 0 \quad \text{if} \quad \theta \neq \arccos\left(\frac{1}{\sqrt{2}}\right) = 45°,$$

and therefore the Hessian is negative definite if $0° < \theta < 45°$; i.e., nonlinear stability holds in the region extending from the pole down to a latitude of 45°.

Case 5. $N = 6$:

$$\Delta_1 = -\frac{35}{6}, \quad \Delta_2 = \frac{1081}{36}, \quad \Delta_3 = -\frac{10361}{72}, \quad \Delta_4 = \frac{5740}{9}, \quad \Delta_5 = -2400,$$

$$\Delta_6 = -\frac{\Delta_5}{3}\frac{30z_o^2 - 17}{r_o^4}, \quad \Delta_7 = \frac{\Delta_5}{12}\frac{(10z_o^2 - 3)(90z_o^2 - 59)}{r_o^8},$$

$$\Delta_8 = -\frac{\Delta_5}{6}\frac{(20z_o^2 - 9)(150z_o^4 - 125z_o^2 + 12)}{r_o^{12}},$$

$$\Delta_9 = 5\,\Delta_5\frac{z_o^2(5z_o^2 - 3)(125z_o^4 - 125z_o^2 + 24)}{r_o^{16}},$$

$$\Delta_{10} = -150\,\Delta_5\,\frac{z_o^4(5z_o^2 - 4)(5z_o^2 - 3)^2}{r_o^{20}}.$$

Then,

$$\Delta_6 > 0 \quad \text{if} \quad 0° \leq \theta \leq \arccos\left(\sqrt{\frac{17}{30}}\right) \approx 41°,$$

$$\Delta_7 < 0 \quad \text{if} \quad \theta \neq \arccos\left(\sqrt{\frac{3}{10}}\right) \approx 36°,$$

$$\Delta_8 > 0 \quad \text{if} \quad \theta \neq \arccos\left(\sqrt{\frac{9}{20}}\right) \approx 48^\circ,$$

$$\Delta_9 < 0 \quad \text{if} \quad 0 \leq \theta \neq \arccos\left(\sqrt{\frac{1}{2} + \frac{\sqrt{145}}{50}}\right) \approx 30.6^\circ,$$

$$\Delta_{10} > 0 \quad \text{if} \quad \theta \neq \arccos\left(\sqrt{\frac{4}{5}}\right) \approx 27^\circ,$$

so that the Hessian is negative definite for $0^\circ \leq \theta \leq \arccos(\sqrt{4/5})$; i.e., nonlinear stability holds in the region from the pole down to a latitude of approximately 63°.

For the case $N = 7$ we obtain that for all $z_0 \neq \pm 1$ the principal minors have different signs and are not alternating. As for the planar case [10, 5], we can therefore conclude that the Hessian is not definite and further analysis is necessary.

It is shown in Boatto, Cabral, and Simó [5] that a polygonal ring of N identical vortices is stable if

$$r_o^2 < \tfrac{7-N}{4} \qquad \text{for N odd,}$$

$$r_o^2 < -\tfrac{N^2-8N+8}{4(N-1)} \qquad \text{for N even,}$$

where $r_o^2 = 1 - z_o^2$. Therefore a polygonal configuration with $N \geq 7$ is not stable on a sphere.

Before ending this section we would like to make the following remarks:

(a) We remind the reader that for three vortices (Case 2, $N = 3$) the dynamics is integrable (see section 2.2); therefore stability can be simply inferred by reducing the Hamiltonian to that of a system of one degree of freedom,

$$H = B_1 \log\left\{\frac{1}{64}(3 - M^2 + 2\,MJ_1)^2(3 - M^2 + 2\,MJ_2)^2(3 + M^2 - 2\,MJ_2 - 2\,MJ_1)^2\right\}, \tag{17}$$

where $J_1 = z_1$, $J_2 = z_2$, and $M = z_1 + z_2 + z_3$ is the momentum [6]. It can immediately be shown that $(J_1, J_2) = (M/3, M/3)$ is a critical point of (17) and corresponds to a latitudinal equilateral triangle configuration. Furthermore, the Hessian of (17) at the equilibrium ($J_1 = J_2 = z_o = M/3$) takes the form

$$\mathcal{H}(H)\left(J_1 = \frac{M}{3}, J_2 = \frac{M}{3}\right) = -B_1 \frac{72\,M^2}{(M^2-9)^2}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

and therefore we can simply deduce that the equilateral triangle configuration is stable everywhere on the sphere except at the poles ($M = 3$) and at the equator ($M = 0$). Such an analysis can be easily generalized to the case of nonidentical vortices; for details, see [6].

(b) The presence of a polar vortex greatly modifies the stability analysis of the polygonal ring. We have already discussed in the introduction that for the planar case Cabral and Schmidt [10] showed that a central vortex of vortex strength K could stabilize a polygonal configuration of identical vortices, and they provided analytical expressions for the K-intervals which assure stability

for a given N (see (1) and [5]). Similarly, for the spherical case—for a polar vortex of vorticity K and a latitudinal polygonal ring of N identical point-vortices—Boatto and Cabral have found explicit analytical expressions for the θ- and K-intervals which assure stability (for details, see [5]). They have found the following stability ranges for $N = 3$.
For $z \neq \frac{1}{3}$, let $\kappa_1(z)$ and $\kappa_2(z)$ be the minimum and the maximum of the two functions

$$g_2(z) = -\frac{2z}{1+z} \quad \text{and} \quad g_3(z) = \frac{2z(3z^2-4z-3)}{(1+z)^2(1-3z)}.$$

We notice that

$$\kappa_1 = g_3 \quad \text{and} \quad \kappa_2 = g_2 \quad \text{for} \quad -1 < z \leq \frac{1}{3} \quad \text{and} \quad 0 \leq z < \frac{1}{3}$$

and

$$\kappa_1 = g_2 \quad \text{and} \quad \kappa_2 = g_3 \quad \text{for} \quad -\frac{1}{3} \leq z \leq 0 \quad \text{and} \quad \frac{1}{3} < z < 1.$$

Then, according to z ($z = \cos\theta$, θ being the colatitude), we have stability in the following situations:

$$\begin{array}{cccc}
-1 < z < z', & \kappa_1(z) < \kappa < 0 & \text{or} & \kappa > \kappa_2(z) > 0, \\
z = z', & & & \kappa > \kappa_2(z) > 0, \\
z' < z < 0, & 0 < \kappa < \kappa_1(z) & \text{or} & \kappa > \kappa_2(z) > 0, \\
0 \leq z \leq \frac{1}{3}, & & & \kappa > 0, \\
\frac{1}{3} < z < 1, & & & 0 < \kappa < \kappa_2(z),
\end{array}$$

where $z' = -\frac{\sqrt{13}-2}{3}$ is the negative root of the equation $3z^2 - 4z - 3 = 0$.

5. Conclusions. Our nonlinear stability ranges coincide with those computed by Polvani and Dritschel for linear stability [31].

There are many other interesting stability problems concerning relative equilibria, and we refer the reader to the book by Newton [27] and the article by Aref et al. [3] for a review. We would like to stress that our stability approach is rather simple: it consists of viewing the dynamics in a good reference frame and using a Lyapunov-like criterion (the Dirichlet criterion).

Furthermore, our study also has relevance for its modeling aspect. In 1998, a subtropical hexagonal jet was observed in the southern hemisphere [22]. This structure proved to be stable for quite a few days, and the question arises of understanding its persistence. The simplest model of a hexagonal jet is a regular polygonal configuration of point-vortices.

The addition of one or two polar vortices strongly modifies the stability ranges (θ-intervals) for the latitudinal ring. More details will be given in a forthcoming paper [5].

Acknowledgments. We express our thanks to A. Chenciner, N. Lebovitz, and D. S. Schmidt for helpful discussions.

REFERENCES

[1] A. Albouy, *The symmetric central configurations of four equal masses*, Contemp. Math., 198 (1996), pp. 131–135.

[2] H. Aref and N. Pomphrey, *Integrable and chaotic motions of four vortices.* I. *The case of identical vortices*, Proc. Roy. Soc. London Ser. A, 380 (1982), pp. 359–387.

[3] H. Aref, P. K. Newton, M. A. Stremler, T. Tokieda, and D. L. Vainchtein, *Vortex crystals*, Adv. Appl. Math., 39 (2002), pp. 1–79.

[4] S. Boatto, *Formation des amas tourbillonnaires pour des systèmes des vortex ponctuels et dynamique des particules test*, in Proceedings of the Centre National de la Recherche Scientifique (CNRS) winter school, Pralognan, France, 1999, pp. 85–90.

[5] S. Boatto, H. E. Cabral, and C. Simó, *Nonlinear stability of a latitudinal ring with vortices at the poles on a non-rotating sphere*, in preparation.

[6] S. Boatto and J. Laskar, *Point-vortex cluster formation in the plane and on the sphere: An energy bifurcation condition*, Chaos, 13 (2003), pp. 824–835.

[7] S. Boatto and R. T. Pierrehumbert, *Dynamics of a passive tracer in the velocity field of four identical point vortices*, J. Fluid Mech., 394 (1999), pp. 137–174.

[8] S. Boatto and T. Tokieda, *Curvature Perturbations and Vortex Stability*, in preparation.

[9] H. E. Cabral and K. R. Meyer, *Stability of equilibria and fixed points of conservative systems*, Nonlinearity, 12 (1999), pp. 1351–1362.

[10] H. E. Cabral and D. S. Schmidt, *Stability of relative equilibria in the problem of $N+1$ vortices*, SIAM J. Math. Anal., 31 (1999), pp. 231–250.

[11] M. T. DiBattista and L. M. Polvani, *Barotropic vortex pairs on a rotating sphere*, J. Fluid Mech., 358 (1998), pp. 107–133.

[12] P. G. L. Dirichlet, *Werke*, Vol. 2, Georg Reiner, Berlin, 1897, pp. 5–8.

[13] D. G. Dritschel, *The stability and energetics of co-rotating uniform vortices*, J. Fluid Mech., 157 (1985), pp. 95–134.

[14] D. G. Dritschel, *Contour dynamics and contour surgery: Numerical algorithms for extended high-resolution modeling of vortex dynamics in two-dimensional, inviscid, incompressible flows*, Computer Phys. Rep., 10 (1989), pp. 77–146.

[15] I. M. Gel´fand, *Lectures on Linear Algebra*, Tracts in Math. 9, Interscience, New York, London, 1963.

[16] K. M. Kahnin, *Quasi-periodic motion of vortex system*, Phys. D, 4 (1982), pp. 261–269.

[17] Lord Kelvin, *Mathematical and Physical Papers*, Vol. 4, Nos. 10 and 12, Cambridge University Press, Cambridge, UK, 1910.

[18] R. Kidambi and P. K. Newton, *Motion of three point vortices on a sphere*, Phys. D, 116 (1998), pp. 143–175.

[19] Y. Kimura and H. Okamoto, *Vortex motion on a sphere*, J. Phys. Soc. Japan, 56 (1987), pp. 4203–4206.

[20] G. R. Kirchhoff, *Vorlesungen über mathematische Physik. Mechanik*, Teubner, Leipzig, 1876.

[21] F. Laurent-Polz, *Point vorticies on the sphere: A case with opposite vorticites*, Nonlinearity, 15 (2002), pp. 143–171.

[22] B. Legras, *private communication*, Laboratoire de Météorologie Dynamique, Ecole Normale Supérieure, Paris, France, 2001.

[23] C. Lim, J. Montaldi, and M. Roberts, *Relative equilibria of point vortices on the sphere*, Phys. D, 148 (2001), pp. 97–135.

[24] A. P. Markeev, *On the stability of the triangular libration points in the circular bounded three body problem*, Appl. Math. Mech., 33 (1966), pp. 105–110.

[25] J. Montaldi, A. Soulière, and T. Tokieda, *Vortex dynamics on a cylinder*, SIAM J. Appl. Dynam. Sys., 2 (2003), pp. 417–430.

[26] G. K. Morikawa and E. V. Swenson, *Interacting motion of rectilinear geostrophic vortices*, Phys. Fluids, 14 (1971), pp. 1058–1073.

[27] P. K. Newton, *The N-Vortex Problem. Analytical Techniques*, Springer-Verlag, New York, 2001.

[28] K. A. O´Neil, *Stationary configurations of point vortices*, Trans. AMS, 302 (1987), pp. 383–425.

[29] J. I. Palmore, *Relative equilibria of vortices in two dimensions*, Proc. Natl. Acad. Sci. USA, 79 (1982), pp. 716–718.

[30] S. Pekarsky and J. E. Marsden, *Point vortices on a sphere: Stability of relative equilibria*, J. Math. Phys., 39 (1998), pp. 5894–5907.

[31] L. M. Polvani and D. G. Dritschel, *Wave and vortex dynamics on the surface of a sphere*, J. Fluid Mech., 255 (1993), pp. 35–64.

[32] C. L. Siegel and J. K. Moser, *Lectures on Celestial Mechanics*, Springer-Verlag, New York, Heidelberg, Berlin, 1971, section 29, p. 208.
[33] A. G. Sokol'skii, *Stability of the Lagrange solutions of the restricted three body problem for critical ratio of the masses*, Appl. Math. Mech., 39 (1975), pp. 342–345.
[34] A. Soulière and T. Tokieda, *Periodic motions of vortices on surfaces with symmetry*, J. Fluid Mech., 460 (2002), pp. 83–92.
[35] J. J. Thomson, *A Treatise on the Motion of Vortex Rings*, Macmillan, New York, 1883.
[36] J. J. Thomson, *Electricity and Matter*, Westminster Archibald Constable, London, 1904.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 231–259

GEOCHEMICAL PHASE DIAGRAMS AND GALE DIAGRAMS*

P. H. EDELMAN†, S. W. PETERSON‡, V. REINER§, AND J. H. STOUT¶

Abstract. The problem of predicting the possible topologies of a geochemical phase diagram, based on the chemical formula of the phases involved, is shown to be intimately connected with and aided by well-studied notions in discrete geometry: Gale diagrams, triangulations, secondary fans, and oriented matroids.

Key words. Gale diagram, Gale transform, phase diagram, triangulation, secondary polytope, secondary fan, geochemistry, heterogeneous equilibrium, closed net, Euler sphere

AMS subject classifications. 52B35, 52C40, 86A99

DOI. 10.1137/S003613990241182X

1. Introduction.

A central problem in geochemistry has been to understand how the equilibrium state of a chemical system varies with temperature and pressure, and to predict the form of its temperature-pressure phase diagram (hereafter called just the *phase diagram*). The purpose of this paper is to explain how some recently developed tools from discrete geometry (the theory of oriented matroids, triangulations, Gale diagrams, and secondary fans) can be used to elucidate this problem. Our goal is to be comprehensible to both discrete geometers and workers in geochemistry.

Figure 1 illustrates the familiar phase diagram for a simple chemical system that involves three phases (ice, water, steam) of the same chemical compound, H_2O. The topological structure of this diagram is fairly simple: there is a unique point, called a triple point, where all three phases can be present in equilibrium. The triple point lies at the junction of three curves. Along each of these curves, exactly two of the phases are present in equilibrium (either ice + water, or ice + steam, or water + steam), and these three curves separate two-dimensional regions where only one phase (pure ice, pure water, or pure steam) can be present in equilibrium.

This example of a phase diagram is quite an elementary one, in that all the phases have the same underlying chemistry, that of H_2O. Geochemists are interested in phase diagrams as the principal tool in reconstructing the temperature and pressure conditions from rock formations once deep within the Earth but which now reside at its surface. Thus it is important to have accurate phase diagrams involving much more complex systems, in which the phases have different chemistry as well as different states.

To be more precise, a *phase* means a physically homogeneous substance, having its own chemical formula, although different phases within the system can have the same

*Received by the editors July 22, 2002; accepted for publication (in revised form) April 1, 2003; published electronically November 19, 2003.
http://www.siam.org/journals/siap/64-1/41182.html
†Department of Mathematics, Vanderbilt University, Nashville, TN 37240-0001 (edelman@math.vanderbilt.edu).
‡School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (peterson@math.umn.edu). The work of this author was carried out partly as a Masters thesis at the University of Minnesota Center for Industrial Mathematics, and was partly supported by a University of Minnesota Grant-in-Aid of Research.
§School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (reiner@math.umn.edu). The work of this author was partially supported by NSF grant DMS-9877047.
¶Department of Geology and Geophysics, University of Minnesota, Minneapolis, MN 55455 (jstout@umn.edu).

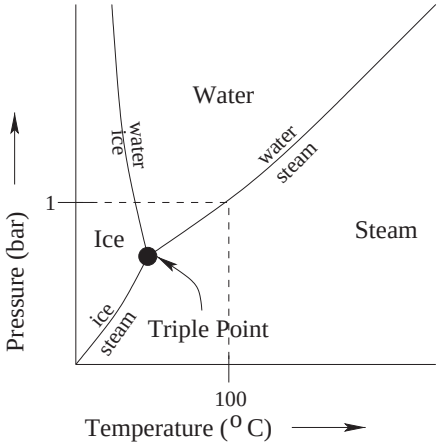


FIG. 1. *The phase diagram for the simple chemical system with phases ice, water, and steam.*

formula (as in the ice-water-steam example). At a particular temperature and pressure, the equilibrium state consists of groups of one or more phases that are referred to as *phase assemblages*. Within a closed system at fixed temperature and pressure, only certain phase assemblages will be stable, namely, those having the lowest possible *Gibbs free energy* under the given conditions. Other assemblages with higher energy than the minimum under those conditions are referred to as *metastable*—these assemblages react spontaneously to produce a stable assemblage and a net decrease in energy. For example, pure water placed in the stability field of ice will spontaneously freeze because a lower Gibbs energy assemblage (ice) is available under those conditions. When there are different chemical formulae present among the phases, more exotic reactions than simple phase changes are possible. The regions of simultaneous stability for various collections of phase assemblages and the chemical reactions that relate them can be conveniently summarized in the phase diagram.

The locus of temperatures and pressures within which a particular phase assemblage is stable is called its *stability field*. The stability field is called *invariant*, *univariant*, or *divariant* depending upon its dimension, that is, the number of degrees of freedom one can vary while staying within that stability field. In the example above, the triple point (ice-water-steam) is an invariant point, there are three univariant curves (ice-water, ice-steam, water-steam), and three divariant stability fields (pure ice, pure water, pure steam). The univariant fields correspond to simple chemical reactions that transform one phase assemblage to another, and hence are sometimes referred to as *reaction lines*. In producing these phase diagrams, a prediction of the possible topologies (i.e., number of invariant points, number of univariant curves joining them, etc.) is indispensable, as the thermodynamic data needed to resolve such topological features can sometimes be difficult to obtain.

It turns out that much of the complexity of the phase diagram for a chemical system is governed by two parameters:

- the number of phases, m, and
- the number of components, n (defined below).

We will see in section 3 that these two parameters correspond to the *size of the ground set* and the *rank* of a related vector configuration (or affine point configuration or oriented matroid) associated with the chemical system. It is well known, in

both geochemistry and discrete geometry, that what matters most in predicting the complexity of the phase diagram is not the sizes of n and m, but rather the sizes of n and $m - n$ (the rank and the *corank*).

The *number of components* n for a chemical system is defined as follows. Think of the chemical formulae of the various phases of the system as vectors in a vector space of all possible such formulae (the *chemical composition space*—see section 3), whose coordinate axes correspond to the elements present on earth. Then the number n of components of the chemical system is simply the dimension of the subspace spanned by the chemical formulae of the phases present in the system.[1] For example, the system of ice, water, and steam from Figure 1 has $m = 3$ and $n = 1$, while the system depicted in Figure 2 (in section 3) has $m = 4$ and $n = 2$.

Phase diagram topologies for chemical systems with $m \leq n + 2$ are fairly well understood, even as m grows large. For $m \leq n+1$ they are essentially trivial, and for $m = n + 2$, they look roughly like Figure 1, having an invariant point surrounded by several univariant reaction curves. The schematic picture of such an invariant point surrounded by reaction curves is referred to as an *invariant point map* [17]. We will explain in section 8 why invariant point maps look roughly like a *two-dimensional Gale diagram*, in concordance with rules for the phase diagram's construction first delineated by Schreinemakers [25] nearly 100 years ago.

However, by $m = n + 3$ (a situation common for chemical systems applicable to the earth) the topology of the phase diagram can become quite complex as m grows large. For example, under certain genericity assumptions about the chemical formulae of the phases, the diagram will contain exactly $n+3$ invariant points located at various temperature and pressure coordinates. These invariant points are connected to one another by various reaction lines to a form a network of points and lines referred to by geochemists as a *petrogenetic grid*. For example, a typical grid for $n = 4$ and $m = 7$ will have seven invariant points connected by 21 different reaction lines.

The phase diagram topology of this (as well as higher order systems) has been represented schematically by geochemists via a *straight line net*, made up of a set of invariant point maps linked together by common reaction lines. In section 9 we will explain how straight line nets for systems with $m = n + 3$ can be constructed using *three-dimensional Gale diagrams* and *secondary fans*, and hence why their phase diagrams strongly resemble the encoding of a three-dimensional Gale diagram as a two-dimensional *affine Gale diagram*.

Within the geochemical literature, there are two general approaches to reconstructing the topology of the phase diagram. The first approach was pioneered by Schreinemakers [25], who reasoned about the relative Gibbs energies of phases to deduce invariant point maps for $m = n + 2$. This method has been extended by various authors (see [29], [31], [32]; [6]; [16], and references therein) to produce viable straight line nets for systems with $m = n + 3$. All feasible topologies are enumerated by this method, and then empirical data is used to eliminate those diagrams which are physically impossible. This approach can be made much more efficient by the methods described in this paper.

The alternative approach is to compute the variation in Gibbs energy directly for every phase of interest. Modern thermodynamic databases (e.g., [13] and references therein) now make these computations possible. The method fails in some cases be-

[1] It will be convenient in section 3 to choose a basis for this space, that is, a minimal set of phases such that all the formulae of the phases can be expressed as linear combinations of these basic components; hence the term *number of components* of the system.

cause the data either lacks the accuracy to distinguish between topologically different diagrams or is simply not available for some phases. The latter situation is becoming more common as new phases are discovered by laboratory synthesis under high pressure conditions. The quantities of these phases are so small that the necessary thermodynamic data will not be available in the foreseeable future. Thus geochemists must rely on the topological approach developed earlier.

We should point out that there have been a few authors [12], [27] who have given a somewhat similar mathematical formulation of this problem, but without taking advantage of the language, techniques, and highly developed theory provided by Gale diagrams and oriented matroids. The applications derived in section 10 of this paper are, as far as we know, new. Furthermore, the theory described in this paper is the basis for JAVA applets written by the second author and available on the web [18], which give practical tools for use in geochemistry for predicting phase diagram topology.

2. A geochemistry-discrete geometry glossary. For the convenience of the reader, and as a guide to what lies in store, we present a (very rough) glossary of corresponding terms.

Geochemistry	Discrete geometry
chemical formula for a phase	vector in composition space
chemical system	acyclic vector configuration $\mathcal{V}$
chemography	affine point configuration $\mathcal{A}$
number of phases m	number of vectors/points
number of components n	rank of vector/point configuration
reactions among phases	linear/affine dependences of $\mathcal{V}/\mathcal{A}$
minimal reactions	circuits
stable assemblage of phases	simplex in a triangulation of $\mathcal{A}$
phase diagram	affine plane slice of secondary fan
phase diagram when $m = n + 2$	2-dimensional (2-D) Gale diagram $\mathcal{A}^*$
reaction half-line for $m = n + 2$	vector in 2-D Gale diagram $\mathcal{A}^*$
phase diagram when $m = n + 3$	2-D affine Gale diagram for $\mathcal{A}^*$
invariant point when $m = n + 3$	vector in 3-D Gale diagram $\mathcal{A}^*$
closed net when $m = n + 3$	spherical representation of $\mathcal{A}^*$
Euler sphere for $m = n + 3$	great circles normal to $\mathcal{A}^*$

3. Chemical composition space. In this section, we introduce the chemical composition space that allows one to associate to each chemical system a configuration of vectors, an affine point configuration, and their oriented matroid. For terminology on vector configurations, point configurations, and oriented matroids, we refer the reader to two excellent references: the bible of the subject, [2] (in particular, its section 1.2), and Ziegler's book, [33, Lecture 6].

DEFINITION 3.1. *The formulae of chemical compounds may be represented in a natural way by vectors in a* chemical composition space *whose axes are indexed by the elements in the periodic table. For example, H_2O has coordinates which are two units on the hydrogen axis, one unit on the oxygen axis, and zero on all other axes. The m phases of a chemical system in this way give rise to a collection $\mathcal{V} = \{v_1, \ldots, v_m\}$ of vectors spanning a subspace of some dimension n, which is called the* number of components *of the system. By picking a basis for this subspace, we can identify it with*

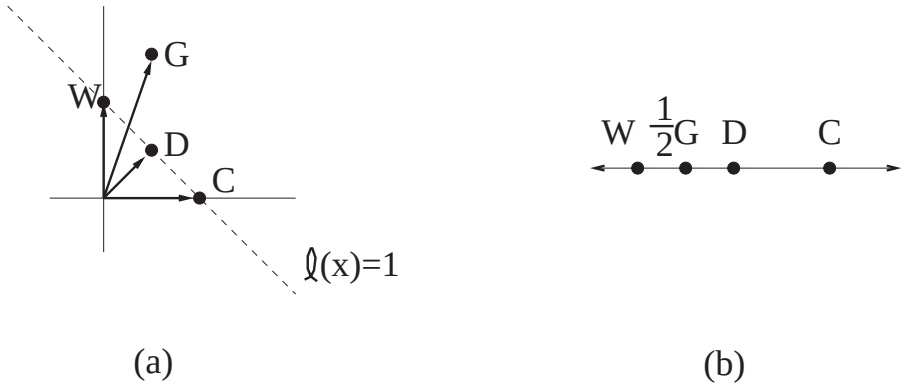


FIG. 2. (a) *Vector configuration* $\mathcal{V}$ *and* (b) *affine point configuration* $\mathcal{A}$ *for the chemical system with phases corundum (C), diaspore (D), gibbsite (G), and water (W).*

$\mathbb{R}^n$ *and specify each* v_i *by a column vector in* $\mathbb{R}^n$*. This allows us to identify* $\mathcal{V}$ *with an* $n \times m$ *matrix having full rank* n*, which we also call* $\mathcal{V}$ *by an abuse of notation.*

Example 3.2. Consider a chemical system of relevance to geology having $m = 4$ phases, which we denote by descriptive initials rather than by v_1, v_2, v_3, v_4:

$$\begin{array}{lll} C = & \text{corundum} & \mathrm{Al_2O_3}, \\ D = & \text{diaspore} & \mathrm{AlO(OH)}, \\ G = & \text{gibbsite} & \mathrm{Al(OH)_3}, \\ W = & \text{water} & \mathrm{H_2O}. \end{array} \tag{1}$$

Since the compounds in this system involve only the elements Al, O, H, this system can have n at most 3. However, one can check that these chemical formulae span a space of dimension $n = 2$. Choosing C and W to be the standard basis vectors in this space (that is, the *components* for this system), one can represent the configuration $\mathcal{V}$ as the columns of the matrix

$$\mathcal{V} = \begin{array}{c} \begin{matrix} C & D & G & W \end{matrix} \\ \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{3}{2} & 1 \end{bmatrix} \end{array}, \tag{2}$$

and the associated vector configuration is depicted in Figure 2(a).

Notice that, by choosing a basis that identifies this n-dimensional subspace with $\mathbb{R}^n$, we are already abstracting away from the actual chemical formulae of the phases and paying attention only to properties that are invariant under a simultaneous change-of-basis acting on the vectors, that is, properties invariant under $GL_n(\mathbb{R})$. One such set of properties is the oriented matroid associated to the vector configuration $\mathcal{V}$.

DEFINITION 3.3. *The* oriented matroid $\mathcal{M}$ *associated to* $\mathcal{V}$ *is a combinatorial abstraction of the vectors* $\mathcal{V}$*, which forgets their actual coordinates but retains data specifying the signs involved in linear dependences among the* v_i*. The way in which we choose to record this data is to list the* signed circuits *of* $\mathcal{M}$ *coming from each minimal (nontrivial) linear dependence* $\sum_i \lambda_i v_i = 0$*, that is, the* signed set (X^+, X^-)*, where*

$$X^{\pm} := \{i \in \{1, \ldots, m\} : \operatorname{sign}(\lambda_i) = \pm\}.$$

Here minimality for signed sets is interpreted with respect to the ordering of their support sets:

$$(X^+, X^-) < (Y^+, Y^-) \quad \textit{means} \quad X^+ \cup X^- \subseteq Y^+ \cup Y^-.$$

It is sometimes convenient to represent a signed set (X^+, X^-) *instead by its* sign vector *in* $\{+, 0, -\}^m$*, which has* $\pm$ *in the coordinates indexed by* $X^\pm$*, and* 0 *in all other coordinates. For example, a minimal dependence of the form* $+5v_1 - 3v_2 + \frac{7}{8}v_4$ *among* $m = 4$ *phases would be recorded by the circuit whose signed set is* $(\{1, 4\}, \{2\})$ *or by its sign vector* $(+ - 0 +)$.

It is possible to write down a small set of *circuit axioms* satisfied by the set $\mathcal{C}$ of signed circuits coming from any vector configuration, in such a way that collections of signed sets satisfying these axioms (*oriented matroids*) mimic many features of sets of vectors in a real vector space; see [2, p. 4]. Note that, since the negation of a linear dependence gives another linear dependence, one of these circuit axioms for an oriented matroid says that the set of sign vectors of circuits is closed under negation.

The linear dependences among the v_i have an obvious chemical interpretation: they are the coefficients in the mass-preserving chemical reactions possible among the phases.

Example 3.4. Continuing the previous example, there are three minimal linear dependences/reactions, giving rise to the following signed circuits (represented by two opposite sign vectors below):

Minimal reaction/dependence	Circuits as sign vectors
$G \rightleftharpoons 1D + 1W$	$0 + - +, \quad 0 - + -$
$2D \rightleftharpoons 1C + 1W$	$+ - 0 +, \quad - + 0 -$
$3D \rightleftharpoons 1C + 1G$	$+ - + 0, \quad - + - 0$

Note that another possible reaction among these phases is

$$1C + 2W \rightleftharpoons 1D + 1G,$$

which would give rise to dependences with sign vectors

$$+ - - +, \quad - + + -,$$

but these are not signed circuits because their support sets are not minimal under inclusion.

The fact that every chemical compound contains a nonnegative amount of each element implies that the vector configuration $\mathcal{V}$ will be *acyclic*, that is, there will be no signed circuits with $X^- = \{\varnothing\}$. Equivalently, there exists a linear functional $\ell(x)$ in $(\mathbb{R}^n)^*$ such that $\ell(v_i) > 0$ for all i. This allows one to replace each v_i by a rescaled vector a_i satisfying $\ell(a_i) = 1$, so that the a_i lie in the affine hyperplane $\ell(x) = 1$ inside $\mathbb{R}^n$, giving rise to an *affine point configuration*[2] $\mathcal{A}$ in $(n-1)$-dimensional affine space $\mathbb{R}^{n-1}$. In chemical terms, this replacement corresponds to simply changing conventions for writing down basic quantities of each phase: instead of considering one mole to be the basic unit of quantity for some phase, one can consider a half a mole or some other fraction to be its basic unit of quantity.[3]

[2] In [17], this affine point configuration $\mathcal{A}$ is called the *chemography* of the system.

[3] We have already pointed out that two phases can have the same chemical formula, giving rise to two copies of the same vector $v_i = v_j$ in $\mathcal{V}$; these give rise to what are called *parallel elements* of the oriented matroid $\mathcal{M}$. We should note, however, that parallel elements can also arise after doing the rescaling from $\mathcal{V}$ to $\mathcal{A}$ if two phases have chemical formulae which differ by a scalar multiple, e.g., if both oxygen O_2 and ozone O_3 were present as phases.

The effect of this rescaling is to turn linear dependences $\sum_i \lambda_i v_i = 0$ of the original vector configuration $\mathcal{V}$ into *affine dependences* of $\mathcal{A}$, that is, relations $\sum_i \bar{\lambda}_i a_i = 0$ with $\sum_i \bar{\lambda}_i = 0$. In terms of chemical reactions, this means that the reactions will not only achieve mass-balance for each atom, but also "coefficient balance," as in the following example.

Example 3.5. Continuing the previous example, we can choose the linear function $\ell(x) = x_1 + x_2$ in $\mathbb{R}^2$ and rescale the coordinates of C, D, G, W so that they have $\ell(x) = 1$, giving the new matrix

$$\mathcal{A} = \begin{array}{c} \begin{matrix} C & D & \frac{1}{2}G & W \end{matrix} \\ \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & \frac{3}{4} & 1 \end{bmatrix} \end{array}. \tag{3}$$

The affine point configuration in $\mathbb{R}^1$ represented by $\mathcal{A}$ is depicted in Figure 2(b).

The rescaling in this case required only replacing G by $\frac{1}{2}G$, so that, for example, the previous reaction/linear dependence

$$\begin{aligned} 1G &\rightleftharpoons 1D + 1W, \\ 1\mathrm{Al(OH)_3} &\rightleftharpoons 1\mathrm{AlO(OH)} + 1\mathrm{H_2O}, \end{aligned}$$

which achieves mass-balance for each atom but not coefficient balance ($1 \neq 1 + 1$), now gives rise to the affine dependence

$$2 \cdot \left(\frac{1}{2}G\right) \rightleftharpoons 1D + 1W,$$

achieving coefficient balance: $2 = 1 + 1$.

Switching from the vector configuration $\mathcal{V}$ to the affine point configuration has psychological advantages, in that it allows one to reduce the dimension by one for visualization purposes, and it makes it easier to think about our next topic: triangulations[4] of $\mathcal{A}$.

4. Triangulations and subdivisions. Our goal in this section is to explain a phenomenon well known to geochemists: by performing reactions that minimize the Gibbs free energy resulting in stable phase assemblages at a particular temperature and pressure, nature (generically) "computes" a triangulation of the point set $\mathcal{A}$.

Having fixed a temperature and pressure (T, P), each of the phases $a_1, \ldots, a_m$ of the chemical system will have a certain *Gibbs free energy* $g_i(T, P)$ per molar quantity (or per whatever basic quantity is being used after rescaling v_i to a_i).

DEFINITION 4.1. *These values $g_i(T, P)$ can be used as* heights *to "lift" the points a_i in $\mathbb{R}^{n-1}$ to points*

$$\hat{a}_i := \begin{bmatrix} a_i \\ g_i(T, P) \end{bmatrix} \in \mathbb{R}^n,$$

giving a new lifted *configuration of points $\hat{\mathcal{A}}$. In other words, we plot the points a_i together with their height along an extra* Gibbs energy axis*; see Figure* 3 *for two examples having $n = 2$.*

[4]We should point out that there is a well-defined notion of triangulations for general vector configurations (even when they are not acyclically oriented), as well as for oriented matroids that do not come from configurations of vectors. See [24] and the references contained therein.

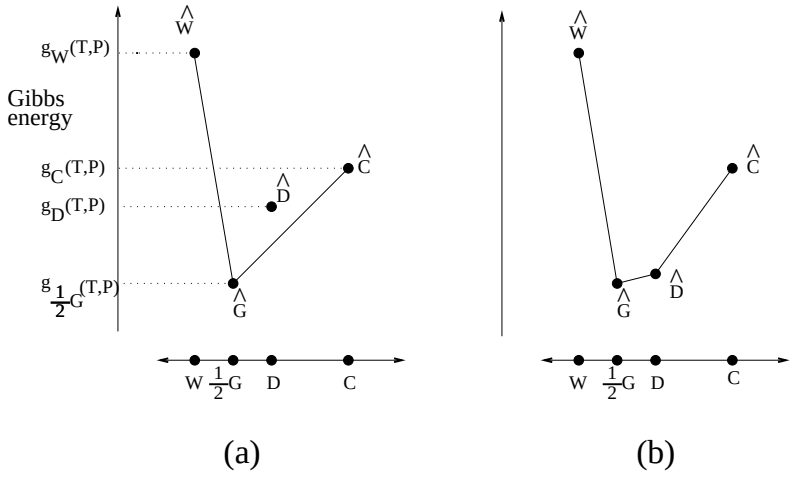


FIG. 3. *The chemography $\mathcal{A}$ from Figure 2, "lifted" to $\hat{\mathcal{A}}$ by the Gibbs energy values at two different values of temperature and pressure.*

The *convex hull*, that is, the set of all convex combinations, of this set $\hat{\mathcal{A}}$ of lifted points has the following physical interpretation. Suppose that we have an assemblage consisting of x_i units of the basic quantity of phase a_i for each $i = 1, \ldots, m$, and assume (without loss of generality) that $\sum_i x_i = 1$. Then this assemblage will have total Gibbs energy $\sum_i x_i g_i(T, P)$, which is the same as the height on the Gibbs energy axis of the point $\sum_i x_i \hat{a}_i$, which is a weighted average of the lifted points $\hat{\mathcal{A}}$ and therefore lies somewhere in their convex hull. If this point does not lie on the *lower convex hull* of these lifted points, then this is not a stable assemblage of phases: there exist some reaction(s) available which would alter the fractions of each phase a_i in a way that lowers the total Gibbs energy.

Example 4.2. We continue our previous example and assume that the Gibbs energies of the phases are as depicted in Figure 3(a). Consider the assemblage consisting of $\frac{1}{2}$ mole of C together with $\frac{1}{2}$ mole of D. It lifts to the point $\frac{1}{2}\hat{C} + \frac{1}{2}\hat{D}$ at the midpoint of the line segment $\hat{C}\hat{D}$ in Figure 3(a), whose height $\frac{1}{2}g_C(T_0, P_0) + \frac{1}{2}g_D(T_0, P_0)$ represents the total Gibbs energy of this assemblage. It is not stable because one can run the reaction

$$3D \rightleftharpoons 1C + 1G$$

in the forward direction to convert the $\frac{1}{2}$ mole of D into $\frac{1}{6}$ mole each of C and G. This creates an assemblage with lower total Gibbs energy consisting of $\frac{2}{3}$ mole of C together with $\frac{1}{6}$ mole of G (or, equivalently, $\frac{1}{3}$ mole of $\frac{1}{2}G$). The latter assemblage, however, is stable, as it lifts to a point on the segment $\hat{C}\hat{G}$ lying in the lower convex hull of $\hat{\mathcal{A}}$.

On the other hand, if the Gibbs energies of the phases look as they do in Figure 3(b), then the initial assemblage of $\frac{1}{2}$ mole of C together with $\frac{1}{2}$ mole of D would be stable, and no reactions would occur.

Note that even after the temperature and pressure (T, P) have been fixed, there can be more than one possible stable assemblage, and which stable assemblages appear depends upon the initial quantities of each phase present. In petrology, when one takes

various samples from different locations inside a stratum of rock formed under the same temperature and pressure conditions, one has a chance of sampling from all the different stable assemblages.

From the previous discussion, we conclude that the sets of phases which can form stable assemblages correspond to the sets of vertices that lie on a face of the lower convex hull of $\hat{\mathcal{A}}$. Note that projecting these faces of the lower hull in $\mathbb{R}^n$ down into $\mathbb{R}^{n-1}$ produces a set of convex polytopes that disjointly cover the convex hull of $\mathcal{A}$, forming what is usually called a *polytopal subdivision* of $\mathcal{A}$. If the vector

$$g = (g_1(T,P), \ldots, g_m(T,P)) \in \mathbb{R}^m$$

is sufficiently generic, then each of the faces of the lower convex hull will be an $(n-1)$-dimensional *simplex* (that is, the convex hull of n affinely independent points), and this polytopal subdivision is called a *triangulation* of $\mathcal{A}$; see [10, Chapter 7] for formal definitions.

DEFINITION 4.3. *Triangulations and polytopal subdivisions of $\mathcal{A}$ which are induced in this fashion from a vector of heights $g = (g_1, \ldots, g_m)$ in $\mathbb{R}^m$ are called* coherent *or* regular*, and we call $\Delta(g)$ the subdivision induced by g.*

We summarize here some of the conclusions of the preceding discussion.

PROPOSITION 4.4. *For each fixed temperature and pressure (T,P), the vector*

$$g = g(T,P) := (g_1(T,P), \ldots, g_m(T,P)) \in \mathbb{R}^m$$

of Gibbs energies for the phases $\mathcal{A} = \{a_1, \ldots, a_m\}$ induces a coherent polytopal subdivision $\Delta(g)$ of $\mathcal{A}$. The polytopes participating in this subdivision have vertex sets corresponding exactly to the stable assemblages of phases at that temperature and pressure (T,P).

It is perhaps surprising that in general not all triangulations of a point configuration $\mathcal{A}$ need be coherent. In Figure 4 we show an affine configuration with six points in $\mathbb{R}^2$ with two incoherent triangulations. This example is well known in the discrete geometry literature (see, e.g., [10, Chapter 7, Figure 27]), and is the "smallest" example due to the following result.

THEOREM 4.5 (see [15]). *When either $n \leq 2$ or $m - n \leq 2$, every triangulation of an affine point configuration of m points in $\mathbb{R}^{n-1}$ is coherent.*

Bearing in mind Proposition 4.4, in order to understand the topology of the phase diagram for a chemical system, one needs to understand how the coherent subdivision $\Delta(g)$ of a point configuration $\mathcal{A}$ varies with the height vector g in $\mathbb{R}^m$. This is our next goal.

5. Secondary fans. The goal of this section is to introduce the secondary fan $\mathcal{F}(\mathcal{A})$ and its close relative, the pointed secondary fan $\mathcal{F}'(\mathcal{A})$, which govern how the coherent subdivisions $\Delta(g)$ of $\mathcal{A}$ change as one varies the height vector g in $\mathbb{R}^m$. Some references for this material are [3, section 4] and [10, Chapter 7].

Having fixed a particular affine point configuration $\mathcal{A}$ of m points in $\mathbb{R}^{n-1}$, one can ask when two height vectors g and g' in $\mathbb{R}^m$ give rise to the same subdivision Δ. It should come as no surprise that the set of vectors g which give rise to a particular subdivision Δ forms a polyhedral cone $\mathcal{C}(\mathcal{A},\Delta) \subset \mathbb{R}^m$; that is, it is defined by a finite set of linear inequalities (which depend on the coordinates of the points $\mathcal{A}$ and on Δ). As one varies the subdivision Δ, these cones $\mathcal{C}(\mathcal{A},\Delta)$ fit together to disjointly cover $\mathbb{R}^m$.

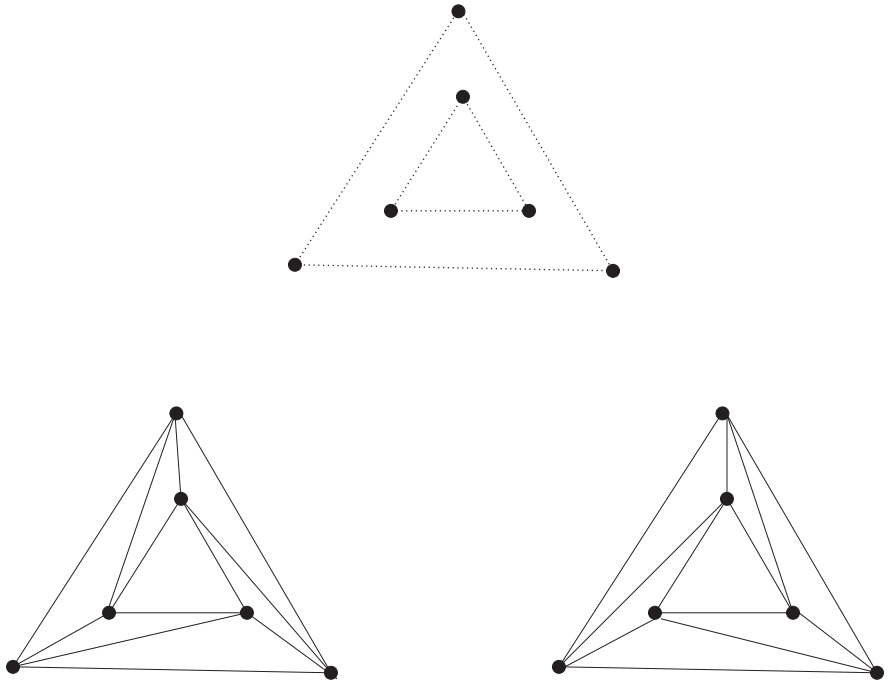


FIG. 4. *Incoherent triangulations of an affine point configuration $\mathcal{A}$: the standard, smallest example, along with its two incoherent triangulations.*

DEFINITION 5.1. *In the terminology of discrete geometry, these cones form a* (complete) fan *called the* secondary fan $\mathcal{F}(\mathcal{A})$*; see Figure* 5(a) *for the example of ice-water-steam from the Introduction.*[5]

We can now rephrase precisely what the phase diagram means in these terms.

DEFINITION 5.2. *Consider the (T, P)-plane in which the phase diagram is drawn as a copy of $\mathbb{R}^2$. Then the Gibbs energy functions $g_i(T, P)$ for the m phases can be viewed as specifying a* Gibbs energy map

$$\begin{array}{rccl} \gamma: & \mathbb{R}^2 & \to & \mathbb{R}^m, \\ & (T,P) & \mapsto & g(T,P) \quad = (g_1(T,P), \ldots, g_m(T,P)). \end{array}$$

The image $\gamma(\mathbb{R}^2)$ of this map will be some two-dimensional surface in $\mathbb{R}^m$. The decomposition of $\mathbb{R}^m$ into the cones of the secondary fan $\mathcal{F}(\mathcal{A})$ will restrict to a decomposition of this surface $\gamma(\mathbb{R}^2)$; see Figure 5(b) for the example of ice-water-steam. This decomposition of the surface then *pulls back* to induce a decomposition of the (T, P)-plane $\mathbb{R}^2$ into regions, which are the regions of simultaneous stability for various collections of phase assemblages (i.e., two pairs $(T, P), (T', P')$ lie in the same region of the phase diagram if and only if their images under γ lie in the same cone of $\mathcal{F}(\mathcal{A})$). In other words, we have the following statement, illustrated in Figure 5(b).

[5]Although we are not aware of a geochemical interpretation, we should point out a beautiful result of Gelfand, Kapranov, and Zelevinsky asserting that the secondary fan $\mathcal{F}(\mathcal{A})$ is actually the *normal fan* of a convex polytope, which they called the *secondary polytope* $\Sigma(\mathcal{A})$; see [10, Chapter 7]. One should also be aware of a slight difference in focus when referring to their results, which they mainly prove for $\Sigma(\mathcal{A})$, but which can then be translated into results about $\mathcal{F}(\mathcal{A})$.

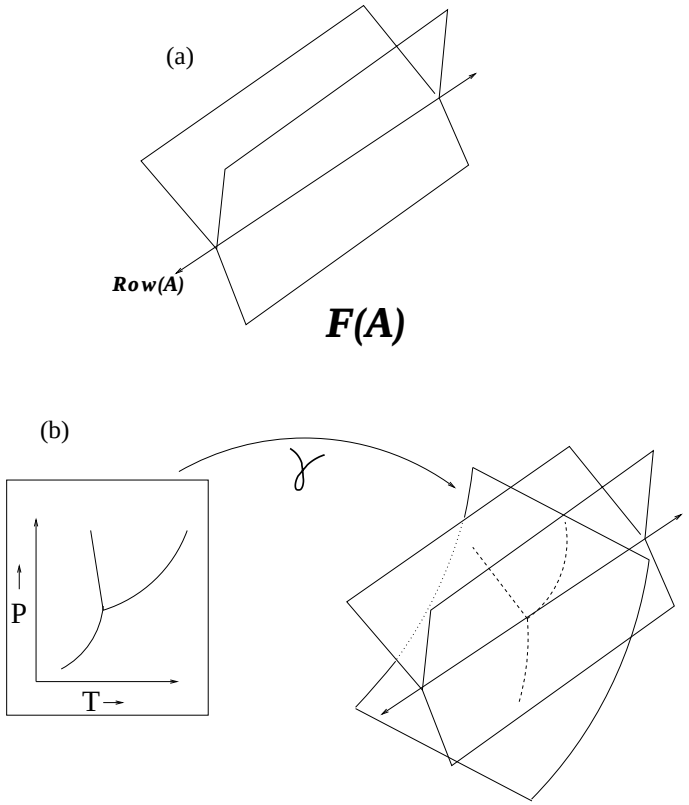


FIG. 5. *For the chemical system of ice, water, and steam with $n = 1$ and $m = 3$, we have* (a) *the secondary fan $\mathcal{F}(\mathcal{A})$,* (b) *the image surface $\gamma(\mathbb{R}^2)$ decomposed by $\mathcal{F}(\mathcal{A})$, and the phase diagram as the pull-back of this decomposition.*

PROPOSITION 5.3. *The phase diagram for a chemical system having chemography $\mathcal{A}$ is exactly the decomposition of the (T, P)-plane $\mathbb{R}^2$, which is the pull-back under γ^{-1} of the decomposition of the image surface $\gamma(\mathbb{R}^2)$ induced by the cones of the secondary fan $\mathcal{F}(\mathcal{A})$.*

Thus understanding possible phase diagram topologies amounts to understanding the structure of the secondary fan $\mathcal{F}(\mathcal{A})$ and the Gibbs energy map γ well enough to predict how the fan $\mathcal{F}(\mathcal{A})$ can decompose two-dimensional surfaces $\gamma(\mathbb{R}^2)$ in $\mathbb{R}^m$.

It turns out that there is a natural way to cut down the dimension of $\mathcal{F}(\mathcal{A})$ by the number of components n, without losing any information. Recall that $\mathcal{A}$ also denotes the $n \times m$ matrix whose columns are the n-vectors a_i (with each of these column vectors lying in the affine hyperplane $\ell(x) = 1$). It is not hard to see that two height vectors g and g', which differ by a vector lying in the row space $\mathrm{Row}(\mathcal{A})$ of this matrix $\mathcal{A}$, will induce the same coherent subdivision Δ: one can show that the two configurations of lifted points they produce will differ by an affine transformation of $\mathbb{R}^n$, and consequently their convex hulls will differ only by a "tilt" that does not affect the structure of their lower hulls. As a consequence, each of the cones $\mathcal{C}(\mathcal{A}, \Delta)$ in the secondary fan $\mathcal{F}(\mathcal{A})$ extends trivially in the n directions defined by $\mathrm{Row}(\mathcal{A})$; it is a Cartesian product

$$\mathcal{C}(\mathcal{A}, \Delta) = \mathcal{C}'(\mathcal{A}, \Delta) \times \mathrm{Row}(\mathcal{A}),$$

where $\mathcal{C}'(\mathcal{A}, \Delta)$ is a cone in an $(m-n)$-dimensional subspace of $\mathbb{R}^m$ complementary[6] to $\mathrm{Row}(\mathcal{A})$. As Δ varies over all coherent subdivisions, these cones $\mathcal{C}'(\mathcal{A}, \Delta)$ disjointly cover this complementary $(m-n)$-dimensional subspace, producing what is called the *pointed secondary fan* $\mathcal{F}'(\mathcal{A})$.

We next make explicit the simplifying assumption which is implicit in the geochemical literature on this subject.

Assumption 5.4 (geochemical assumption). Over the ranges of temperature and pressure $(T, P) \in \mathbb{R}^2$ relevant to most phase diagrams, the Gibbs energy map $\gamma : \mathbb{R}^2 \to \mathbb{R}^m$ is sufficiently close to linear that the image surface $\gamma(\mathbb{R}^2)$ behaves nearly like a two-dimensional affine plane in $\mathbb{R}^m$.

Furthermore, this two-dimensional affine plane is located generically with respect to the cones of the secondary fan $\mathcal{F}(\mathcal{A})$ in the following sense: it has transverse intersection with every cone $\mathcal{C}$ in the secondary fan (including the smallest face which is the row space $\mathrm{Row}(\mathcal{A})$). This means that the intersection is empty if the dimension of the cone $\mathcal{C}$ is less than $m-2$, and otherwise when $\mathcal{C}$ has dimension $m-2, m-1, m$, respectively, the intersection is either empty or it is of dimension $0, 1, 2$, respectively.

With these assumptions, and in particular the transversality assumption, the problem of enumerating the possible phase diagram topologies reduces to understanding the ways in which the pointed secondary fan $\mathcal{F}'(\mathcal{A})$ can decompose an affine two-dimensional plane inside the $(m-n)$-dimensional space where it lives. It turns out that Gale diagrams hold the key to this problem.

6. Gale diagrams and duality. In this section we introduce Gale diagrams of a vector configuration or affine point configuration, and explain their relationship to (pointed) secondary fans and oriented matroid duality.

DEFINITION 6.1. *Given the $n \times m$ matrix $\mathcal{A}$ of rank n whose columns give an affine point configuration, choose a* dual matrix $\mathcal{A}^*$ *to be any $(m-n) \times m$ matrix whose row space* $\mathrm{Row}(\mathcal{A}^*)$ *coincides with the nullspace (or kernel)* $\mathrm{Ker}(\mathcal{A})$. *By an abuse of notation similar to that in Definition* 3.1, *the configuration of column vectors $\{a_1^*, \ldots, a_m^*\}$ of this matrix will also be denoted $\mathcal{A}^*$ and is called a* Gale diagram *or* Gale transform [9] *for $\mathcal{A}$. Similarly, we could have started with any $n \times m$ matrix $\mathcal{V}$ of rank n corresponding to a vector configuration and defined a Gale diagram $\mathcal{V}^*$ for it in the same fashion.*

Example 6.2. Recall our running example with

$$\mathcal{A} = \begin{array}{c} \begin{matrix} C & D & \frac{1}{2}G & W \end{matrix} \\ \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & \frac{3}{4} & 1 \end{bmatrix} \end{array}. \tag{4}$$

An example of a valid Gale diagram for this is

$$\mathcal{A}^* = \begin{array}{c} \begin{matrix} C^* & D^* & G^* & W^* \end{matrix} \\ \begin{bmatrix} 1 & -3 & 2 & 0 \\ 2 & -4 & 0 & 2 \end{bmatrix} \end{array}, \tag{5}$$

pictured as a vector configuration in Figure 6.

[6] It would be more natural to think of the cones $\mathcal{C}'(\mathcal{A}, \Delta)$ as living in the *quotient space* $\mathbb{R}^m/\mathrm{Row}(\mathcal{A})$, but we won't quibble here.

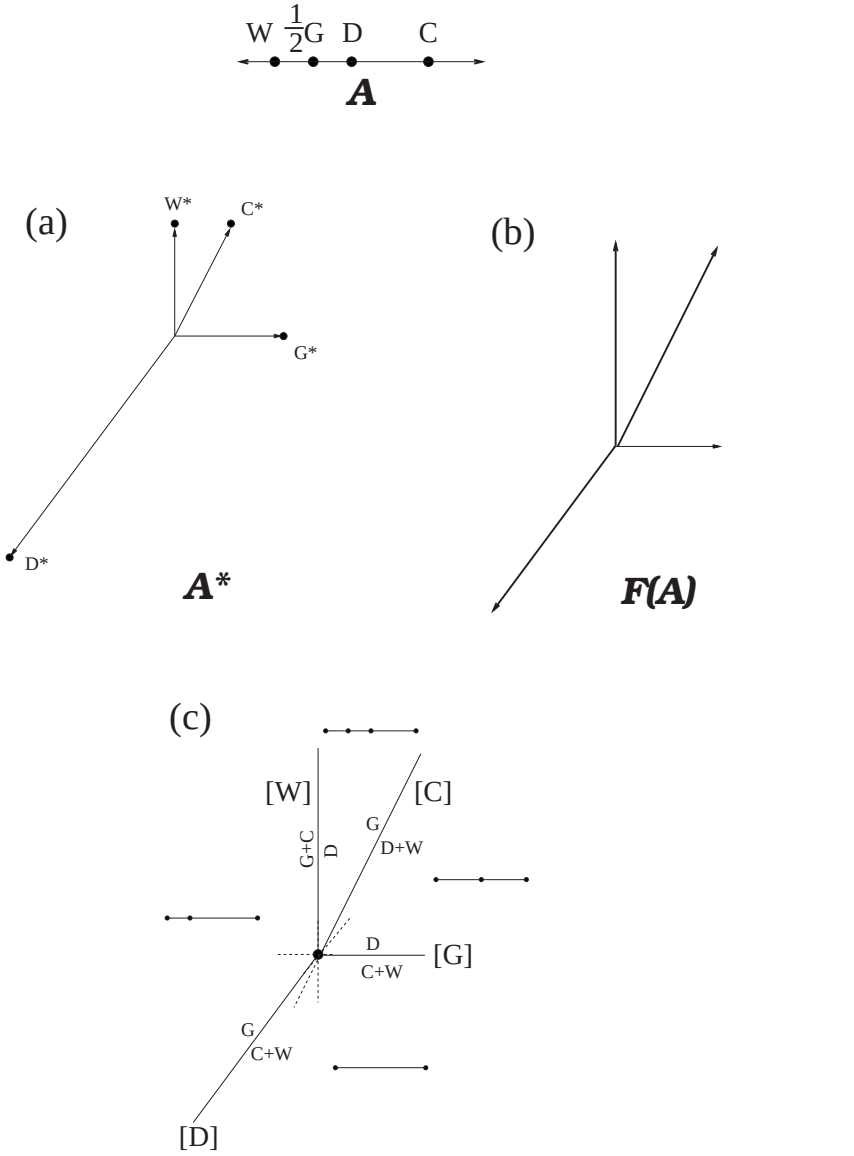


FIG. 6. (a) *Gale diagram* $\mathcal{A}^*$, (b) *secondary fan* $\mathcal{F}(\mathcal{A})$, *and* (c) *invariant point map, for the chemical system corundum-diaspore-gibbsite-water.*

Note the use of the term "a" Gale diagram, instead of "the" Gale diagram. This is because the rows of $\mathcal{A}^*$ are not uniquely defined: they can be altered by row operations, that is, by the action of $GL_{m-n}(\mathbb{R})$ on the left. This means that the Gale diagram vectors $\mathcal{A}^*$ are also well defined only up to the same $GL_{m-n}(\mathbb{R})$-action.[7]

However, the oriented matroid $\mathcal{M}^*$ associated to the Gale vectors $\mathcal{A}^*$ *is* uniquely defined by the oriented matroid $\mathcal{M}$ associated to $\mathcal{A}$: it is the *dual oriented matroid*

[7]Recall from section 3 that there was a similar ambiguity in the definition of the columns of $\mathcal{A}$ or $\mathcal{V}$, stemming from a choice of basis for the space that they span.

[2, section 3.4] of $\mathcal{M}$. One manifestation of this duality is that the circuits $\mathcal{C}$ for $\mathcal{A}$ (or $\mathcal{M}$) correspond to sets in $\mathcal{A}^*$ (or $\mathcal{M}^*$) with their own interesting geometric characterization. These sets in $\mathcal{A}^*$ are called cocircuits.

DEFINITION 6.3. *Given a configuration of vectors $\mathcal{V} = \{v_1, \ldots, v_m\}$ in $\mathbb{R}^n$, its* covectors *are all possible sign vectors in $\{+, 0, -\}^m$ that can be achieved by evaluating some nonzero linear functional $f \in (\mathbb{R}^n)^*$ on the vectors in $\mathcal{V}$:*

$$c = c(f) = (\text{sign}(f(v_1)), \ldots, \text{sign}(f(v_m))).$$

A covector c for $\mathcal{V}$ which is maximal with respect to its set of zeroes is called a cocircuit *of $\mathcal{V}$. Equivalently, a covector is a cocircuit if its corresponding signed subset has minimal support or, equivalently, if the subset of $\mathcal{V}$ on which it is 0 contains at least $n-1$ linearly independent vectors. Denote by $\mathcal{C}^*$ the set of cocircuits of $\mathcal{V}$.*

As with the circuits $\mathcal{C}$, it is possible to write down a list of *cocircuit axioms* that will be satisfied by the cocircuits $\mathcal{C}^*$ coming from any vector configuration $\mathcal{V}$, and in this way to axiomatize the definition of an oriented matroid $\mathcal{M}$ in terms of cocircuits. The observation from above that the circuits C of $\mathcal{A}, \mathcal{V}, \mathcal{M}$ are exactly the cocircuits $\mathcal{C}^*$ of $\mathcal{A}^*, \mathcal{V}^*, \mathcal{M}^*$ means that these cocircuit axioms will look exactly like the circuit axioms.

Example 6.4. In our previous example of $\mathcal{A}$, the first circuit of $\mathcal{A}$ as listed in Example 3.4 was $(0 + - +)$, coming from the reaction $1G \leftrightharpoons 1D + 1W$. In $\mathcal{A}^*$ this is a cocircuit representing the fact that the line spanned by C^* has G^* on one side and D^*, W^* on the opposite side; i.e., there exists a linear functional f for which

$$f(C^*) = 0, \quad f(D^*) > 0, \quad f(W^*) > 0, \quad \text{and } f(G^*) < 0.$$

How does the Gale diagram $\mathcal{A}^*$ relate to the pointed secondary fan $\mathcal{F}'(\mathcal{A})$? The relationship comes from looking at the positive cones spanned by the vectors of $\mathcal{A}^*$.

DEFINITION 6.5. *Given any set $W = \{w_1, \ldots, w_k\}$ of vectors in $\mathbb{R}^N$, define the* positive cone *spanned by W to be*

$$\text{pos}(W) := \left\{ \sum_{i=1}^{k} c_i w_i \in \mathbb{R}^N : c_i > 0 \quad \textit{for all } i \right\}.$$

If the set of vectors W happen to be linearly independent, then $\text{pos}(W)$ *is called a* (relatively open) simplicial cone.

Recall that the affine point configuration $\mathcal{A}$ corresponds to an acyclic vector configuration. It is a consequence of oriented matroid duality [2, Proposition 4.8.9] that $\mathcal{A}^*$ will be *totally cyclic*; that is, the origin 0 lies in the cone $\text{pos}(\mathcal{A}^*)$. As a consequence, the collection of positive cones spanned by subsets of $\mathcal{A}^*$ will cover the column space $\text{Col}(\mathcal{A}^*)$, and this covering is closely related to the pointed secondary fan $\mathcal{F}'(\mathcal{A})$, as shown in the following.

THEOREM 6.6 (see [3, section 4]). *The column space* $\text{Col}(\mathcal{A}^*)$ *has a natural identification with the $(m-n)$-dimensional subspace complementary to* $\text{Row}(\mathcal{A})$ *within $\mathbb{R}^m$ covered by the pointed secondary fan $\mathcal{F}'(\mathcal{A})$.*

Furthermore, under this identification, the cones of $\mathcal{F}'(\mathcal{A})$ are exactly the common refinement of all the open simplicial cones $\text{pos}(W)$ *spanned by linearly independent subsets W of the Gale (column) vectors $\mathcal{A}^*$.*

One can be more precise about this relationship, as follows.

THEOREM 6.7 (see [3, Lemma 4.3]). *Given a coherent triangulation Δ of $\mathcal{A}$, the corresponding $(m-n)$-dimensional cone in the secondary fan $\mathcal{F}'(\mathcal{A})$ is the intersection*

$$\bigcap_{\sigma \in \Delta} \operatorname{pos}(\mathcal{A}^* - \sigma^*),$$

where σ runs through the vertex sets of the $(n-1)$-dimensional simplices in the triangulation Δ and

$$\sigma^* := \{a_i^* : a_i \in \sigma\}.$$

Example 6.8. Because the corundum-diaspore-gibbsite-water example of $\mathcal{A}$ has $m = 4$ and $n = 2$, so that $m = n + 2$, the pointed secondary fan $\mathcal{F}'(\mathcal{A})$ is two-dimensional. Therefore its top-dimensional cones are simply the sectors between cyclically adjacent Gale vectors in $\mathcal{A}^*$. These cones are depicted in Figure 6(b). In (c) of the same figure, these regions are labelled (as part of the geochemists' invariant point map; see section 8 below) by their corresponding coherent triangulations.

For example, the sector lying between the Gale vectors W^* and D^* corresponds to a triangulation Δ having two segments $\{GW, CG\}$. This agrees with Theorem 6.7: the complementary sets $\{C^*D^*, D^*W^*\}$ are exactly the ones whose positive cones contain this sector. On the other hand, the sector between D^*, G^* lies in the positive cone of no other pairs of Gale vectors and hence corresponds to the triangulation having only one segment, namely, the one with vertices $\mathcal{A} - \{D, G\} = \{C, W\}$.

Sections 8 and 9 will closely explore the consequences of the geometry of the Gale diagram for phase diagrams when m is at most $n + 3$. But first we must further explore more general geometric questions.

7. Geometry of the phase diagram in general. In this section we will explain the relationship between Gibbs' phase rule (e.g., [21]) and the secondary fan, and how the phase rule predicts the dimension of various stability fields. We then look closely at the meaning of two-, one-, and zero-dimensional regions in the phase diagram, relating them to m-, $(m-1)$-, and $(m-2)$-dimensional cones in the secondary fan.

When one fixes particular molar fractions x_i of each of the (rescaled) phases a_i in $\mathcal{A}$ initially contained in a particular sample, the discussion of section 5 shows how to predict the stable assemblage of phases which will result after allowing the system to find chemical equilibrium. The initial molar fractions give a point $\sum_i x_i a_i = 1$ with $\sum_i x_i = 1$, which lies in the convex hull of $\mathcal{A}$ and lifts to a point $\sum_i x_i \hat{a}_i$, which lies in the convex hull of $\hat{\mathcal{A}}$ but may or may not lie in the *lower* convex hull. After performing reactions which affect the fractions x_i (but preserve the condition $\sum_i x_i = 1$ due to our rescaling) to reach chemical equilibrium, the lifted point $\sum_i x_i \hat{a}_i$ will eventually lie in a unique face $\hat{F}$ of the lower hull of $\hat{\mathcal{A}}$. The lifted points $\hat{\mathcal{A}}' \subset \mathcal{A}$ which happen to lie on this face $\hat{F}$ lie above a subset $\mathcal{A}' \subset \mathcal{A}$ of the original phases; that is, $\mathcal{A}'$ are those phases which appear with nonzero fraction in this chemical equilibrium, and $\mathcal{A}'$ labels a polytope F which is part of the corresponding coherent subdivision of $\mathcal{A}$.

Gibbs' phase rule relates three quantities relevant to this situation:

- the number of phases $m' = |\mathcal{A}'| (\leq m = |\mathcal{A}|)$ participating in this chemical equilibrium,
- the number of components $n' (\leq n)$ of the subsystem $\mathcal{A}'$, that is, the dimension of the subspace of chemical composition space spanned by $\mathcal{A}'$, and

- the number of degrees of freedom f in (T, P), which one can vary while maintaining these same phases in equilibrium, or in other words, the dimension of the *union* of all regions in the phase diagram which have $\mathcal{A}'$ labelling one of the polytopes F in their corresponding subdivision of $\mathcal{A}$.

PROPOSITION 7.1 (Gibbs' phase rule). *With the above notation,*

$$f = n' + 2 - m'.$$

In particular, one can have at most $n' + 2$ phases that involve n' components in chemical equilibrium.

Note that, in the geochemical literature, the phase rule is often stated as

$$f \leq n + 2 - m',$$

which is consistent with the fact that $n' \leq n$.

Example 7.2. In Figure 1, the triple point has an assemblage of $m' = 3$ phases in equilibria (ice, water, steam), with $n' = 1$ and $f = 0$, while the assemblage consisting of pure ice has $m' = n' = 1$ and $f = 2$.

In Figure 6(c), the line segment DW corresponds to a stable assemblage $\{D, W\}$ in two different divariant regions and the curve that separates them (all in the upper right), so $f = 2$, and it has $m' = n' = 2$. The assemblage $\{D, G, W\}$ is stable only along the univariant curve lying between the two aforementioned divariant regions, so $f = 1$, and it has $m' = 3, n' = 2$. The quadruple point in the middle ($f = 0$) of the diagram has all four phases in equilibrium; that is, $m'(= m) = 4$ and $n'(= n) = 2$.

We give here a proof of Gibbs' phase rule in terms of the secondary fan $\mathcal{F}(\mathcal{A})$ in $\mathbb{R}^m$.

Proof. There is a cone $\mathcal{C}$ in the secondary fan consisting of those vectors $g \in \mathbb{R}^m$ for which the lifted points $\hat{\mathcal{A}}$ have $\hat{\mathcal{A}}'$ lying on a face $\hat{F}$ of the lower hull: this cone is the intersection of the vector space V on which the lifted points $\hat{\mathcal{A}}'$ all lie on a single n'-dimensional affine subspace, with the half-spaces given by various inequalities that assert that all the other lifted points $\hat{a}_i$ in $\mathcal{A} - \mathcal{A}'$ lift *above* this affine subspace. The subspace V is defined by $m' - n'$ linear conditions: after choosing the height coordinates of g to lift n' of the elements $\mathcal{A}'$ which are affinely independent, the remaining $m' - n'$ coordinates must be lifted to heights which are linear functions of those first n' heights. Thus V has dimension $m - (m' - n')$. The fact that lifting all the points $\mathcal{A} - \mathcal{A}'$ to any sufficiently large heights will force $\hat{F}$ to be in the lower hull shows that the cone $\mathcal{C}$ obtained by intersecting V with the various half-space inequalities will have the same dimension as V, namely, $m - (m' - n')$.

By Proposition 5.3 and Assumption 5.4, the union of all regions in the phase diagram which have $\mathcal{A}'$ labelling one of the polytopes F in their corresponding subdivision of $\mathcal{A}$ comes from the transverse intersection of an affine 2-plane with the cone $\mathcal{C}$. If $m' - n' > 2$, there would be no intersection, due to Assumption 5.4. If $m' - n' \leq 2$, then, since we assumed that these phases $\mathcal{A}'$ could exist in stable equilibrium, there would be a nonempty intersection. Depending upon whether $m' - n' = 0, 1$, or 2, this transverse intersection of $\mathcal{C}$ with an affine 2-plane would have dimension $f = 2, 1$, or 0, respectively; i.e., $f = 2 - (m' - n') = n' + 2 - m'$. ∎

Understanding the two-, one-, and zero-dimensional regions in the phase diagram amounts to understanding the cones of dimensions $m, m-1$, and $m-2$ in the secondary fan $\mathcal{F}(\mathcal{A})$ or, equivalently, the cones of dimensions $m - n, m - n - 1, m - n - 2$ in the pointed secondary fan $\mathcal{F}'(\mathcal{A})$. The two-dimensional regions in the phase diagram correspond to the top-dimensional cones in $\mathcal{F}(\mathcal{A})$ (or $\mathcal{F}'(\mathcal{A})$), which are labelled by

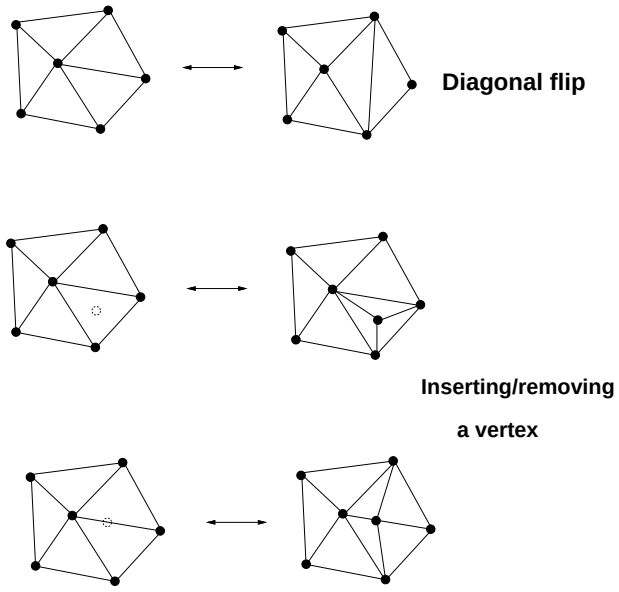


FIG. 7. *Three examples of pairs of bistellar operations for triangulations of affine point configurations $\mathcal{A}$ in $\mathbb{R}^2$.*

the coherent triangulations of $\mathcal{A}$ in the manner described in Theorem 6.7, and there is little to add to that description. The more interesting cases are those of the one- and zero-dimensional stability fields.

7.1. Bistellar operations and one-dimensional stability fields. The one-dimensional curves separating the regions in the phase diagrams correspond to natural transformations on triangulations of $\mathcal{A}$ called *bistellar operations*, which are closely related to the circuits of $\mathcal{A}$. We discuss these now somewhat informally; for a more formal treatment, see [10, Chapter 7, section 2C].

Figure 7 illustrates three examples of a pair of triangulations of affine point configurations $\mathcal{A}$ in $\mathbb{R}^2$ that are related by a bistellar operation. For each bistellar operation between two triangulations, there is a distinguished subset $C \subset \mathcal{A}$, which is the support set $C = X^+ \cup X^-$ of some signed circuit (X^+, X^-) of $\mathcal{A}$ and such that the convex hull of C is triangulated (differently!) in the two triangulations. In this case, we say that the bistellar operation is *supported on the circuit* C. Note that in the first two examples in Figure 7 this circuit C has a full two-dimensional convex hull, but as the third example illustrates, C can have a convex hull of lower dimension.

Recall that a circuit $C = X^+ \cup X^-$ of $\mathcal{A}$ corresponds to a cocircuit of $\mathcal{A}^*$, that is, there is an $(m-n-1)$-dimensional hyperplane H_C spanned by the Gale vectors indexed by $\mathcal{A} - C$, which separates the Gale vectors indexed by X^+ from those indexed by X^-. When C is the circuit supporting a bistellar operation between two triangulations, this reflects the following geometry of pointed secondary fans.

PROPOSITION 7.3 (see [10, section 7.2.C]). *Two triangulations Δ, Δ' differ by a bistellar operation supported on a signed circuit C if and only if their corresponding top-dimensional cones in $\mathcal{F}'(\mathcal{A})$ are adjacent along a wall whose linear span is the hyperplane H_C.*

This has an interpretation for the temperature-pressure phase diagram that is well known to geochemists: the segments of curves separating regions in the phase

diagram are always portions of a larger curve corresponding to some minimal reaction possible among the phases in the chemical system. Two regions will be adjacent and separated by such a curved segment if and only if their corresponding triangulations differ by retriangulating the convex hull of the phases involved in that reaction.

It is also useful to think of a bistellar operation as represented by the coherent polytopal subdivision that labels the wall between the two top-dimensional cones guaranteed by the previous proposition. In the language of Gibbs' phase rule, this subdivision contains a special polytopal face F labelled by a subset $\mathcal{A}' \subset \mathcal{A}$ having $m' = n' + 1$; namely, $\mathcal{A}'$ is the support set of the circuit C. Note that if $n' = n$, then F is a full $(n-1)$-dimensional polytope in the subdivision, and all the other full-dimensional polytopes in the subdivision are $(n-1)$-dimensional simplices.

A reasonable question at this point is, How well do the bistellar operations tie together the set of all triangulations of $\mathcal{A}$—is it possible to connect any two triangulations of a point set $\mathcal{A}$ by a sequence of bistellar operations? The answer has important consequences for calculating the set of triangulations of $\mathcal{A}$: the algorithms (e.g., [20]) that start with one triangulation and find the rest bistellarly connected to it by performing all possible bistellar operations are much faster than algorithms that find *all* triangulations by the currently available techniques [7], [20].

Unfortunately, the answer to the above question is in general no: Santos [22], [23] has recently produced examples of triangulations of affine point configurations that are connected to *no* other triangulations (!) by bistellar operations. Fortunately, however, there are positive results relevant for the geochemical applications:

- all triangulations are connected by bistellar operations when $n \leq 3$ (see [14]),
- the same holds when $m - n \leq 3$ (see [1]), and
- the subset of *coherent* triangulations are always connected by bistellar operations (see [10]).

In particular, this last result allows one to rely on the very fast bistellar flip algorithms of [20] (utilized in [18]) to find all of the coherent triangulations.

7.2. Invariant points and indifferent crossings. We conclude this section with an informal discussion of zero-dimensional regions in the phase diagram. These will correspond to cones of dimension $m-2$ in $\mathcal{F}(\mathcal{A})$ or cones of dimension $m-n-2$ in $\mathcal{F}'(\mathcal{A})$. These correspond to coherent polytopal subdivisions of $\mathcal{A}$ of two possible types, and therefore give rise to two distinct types of points in the phase diagram: invariant points and indifferent crossings.

DEFINITION 7.4. *If a coherent polytopal subdivision of $\mathcal{A}$ corresponds to a cone of dimension $m-2$ in $\mathcal{F}(\mathcal{A})$, then one of the polytopes F' in the subdivision might correspond to a stable assemblage $\mathcal{A}'$ having $m' = n'+2$ phases. When this occurs, the same holds for every polytope in the subdivision which contains F' as a face. However, the remaining polytopes which do not contain F' will all be simplices.*

It is in this situation that geochemists reserve the term invariant point *for the corresponding zero-dimensional region in the phase diagram. In this situation it is possible for all of the phases in $\mathcal{A}'$ to coexist in chemical equilibrium, but one cannot vary (T, P) at all while maintaining this. For example, the central points in Figures* 1 *and* 6(c) *are invariant points, as are the points labelled $[B], [C], [D], [G], [W]$ within the diagrammed regions of Figure* 12 *(below).*

Geochemists usually label the invariant point by the phases $\mathcal{B} := \mathcal{A} - \mathcal{A}'$ *not* involved in the invariant equilibrium. It is also well known to geochemists that the local structure of the phase diagram around an invariant point is similar to the corresponding phase diagram with $m' = n'+2$ for the chemical subsystem $\mathcal{A}'$. This corresponds to

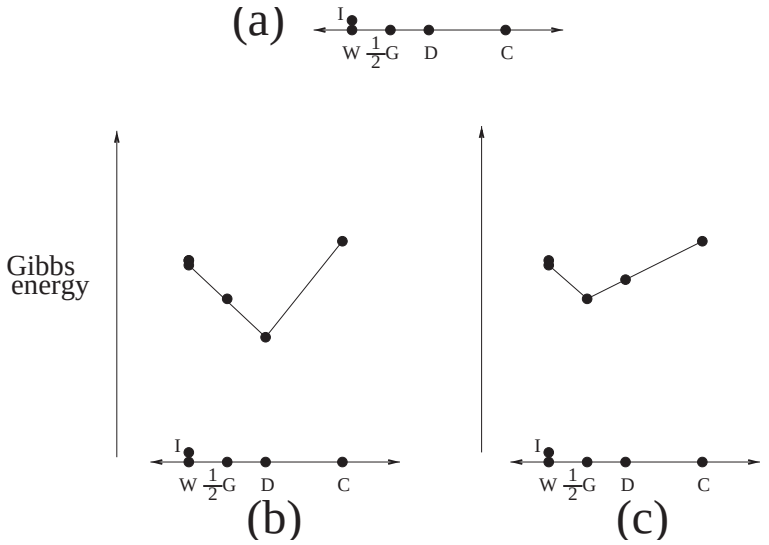


FIG. 8. *An $m = n+3$ system illustrating the distinction between invariant points and indifferent crossings.*

the known fact (see [4]) that the local structure of the pointed secondary fan $\mathcal{F}'(\mathcal{A})$ about the cone $\text{pos}(\mathcal{B})$ and the structure of the fan $\mathcal{F}'(\mathcal{A}') = \mathcal{F}'(\mathcal{A} - \mathcal{B})$ coincide, reflecting the fundamental duality between *deletion* and *contraction* in (oriented) matroid theory: the dual point configuration $(\mathcal{A}-\mathcal{B})^*$ to the deletion $\mathcal{A}-\mathcal{B}$ is isomorphic to the contraction $\mathcal{A}^*/\mathcal{B}^*$.

DEFINITION 7.5. *On the other hand, when a cone in $\mathcal{F}(\mathcal{A})$ has dimension $m-2$, there can also be two polytopes F' and F'' in the corresponding polytopal subdivision, neither contained as a face of the other, corresponding to stable assemblages $\mathcal{A}'$ and $\mathcal{A}''$ with $m' = n' + 1, m'' = n'' + 1$. The polytopes in the subdivision containing neither of F' or F'' will all be simplices. For each of $\mathcal{A}'$ or $\mathcal{A}''$ individually, the union of regions in the phase diagram where they occur as a stable assemblage corresponds to a cone of dimension $m-1$ in $\mathcal{F}(\mathcal{A})$ and a curve in the phase diagram coming from a minimal reaction possible in $\mathcal{A}$. These two curves intersect at what is called an* indifferent crossing*, where either $\mathcal{A}'$ or $\mathcal{A}''$ might exist in equilibrium (as might assemblages corresponding to other simplices in the subdivision), but the union $\mathcal{A}' \cup \mathcal{A}''$ cannot stably coexist: the faces F', F'' lift to two different faces in the lower hull of $\hat{\mathcal{A}}$, lying above disjoint possibilities for the molar fractions of the phases.*

Example 7.6. To illustrate the distinction between the two kinds of cones of dimension $m-2$ in $\mathcal{F}(\mathcal{A})$ that give rise to invariant points versus indifferent crossings, we augment our previous chemical system of corundum, diaspore, gibbsite, and water with a fifth phase: ice, abbreviated I, having chemical formula H_2O, the same as water. Rescaling this to a chemography as before gives a new chemography $\mathcal{A}$ with $n = 2$ as before, but now with $m = 5 = n+3$, depicted in Figure 8(a). Figure 8(b) and (c) depict lifted configurations $\hat{\mathcal{A}}$ that would correspond to cones of dimension $(m-2)$ in $\mathcal{F}(\mathcal{A})$, corresponding to an invariant point and indifferent crossing, respectively.

In (b), the interesting stable assemblage is $\mathcal{A}' = \{D, G, W, I\}$, having $m' = 4, n'(= n) = 2$, so $m' = n' + 2$, and the corresponding lifted points $\hat{\mathcal{A}}'$ lie on a single face F' in the lower hull of $\hat{\mathcal{A}}$. If the phase diagram were to contain a point corresponding to

this cone of $\mathcal{F}(\mathcal{A})$, it would be an invariant point, labelled $[C]$ for the missing phase corundum not present in $\mathcal{A}'$.

In (c), there are at least two interesting stable assemblages: $\mathcal{A}' = \{W, I\}$, with $m' = 2, n' = 1$, and $\mathcal{A}'' = \{C, D, G\}$, with $m'' = 3, n' = 2$, so that $m' = n' + 1, m'' = n'' + 1$, and their corresponding lifted points $\hat{\mathcal{A}}', \hat{\mathcal{A}}''$ span different faces F', F'' of the lower hull of $\hat{\mathcal{A}}$. If the phase diagram were to contain a point corresponding to this cone of $\mathcal{F}(\mathcal{A})$, it would be an indifferent crossing, lying at the intersection of two curves corresponding to the two circuits (reactions) involving the phases $\mathcal{A}'$ and $\mathcal{A}''$.

8. The case $m = n+2$: Phase diagram = Gale diagram. After dispensing quickly with the cases $m = n$ and $m = n + 1$, in this section we examine in detail the structure of the Gale diagram $\mathcal{A}^*$, the pointed secondary fan $\mathcal{F}'(\mathcal{A})$, and the phase diagram when $m = n + 2$. The conclusion is that they all look roughly the same in this case.

When $m = n$, the m phases cannot perform any reactions that preserve mass-balance, and so are mutually inert and nothing can happen.

When $m = n + 1$, not much interesting happens. There is exactly one reaction possible, corresponding to the unique signed circuit $C = (X^+, X^-)$ of $\mathcal{A}$. The Gale diagram $\mathcal{A}^*$ is a set of vectors lying on the real line $\mathbb{R}^1$ with their tails at the origin 0. Those a_i^* having $i \in X^+$ will point in the positive direction, those with $i \in X^-$ will point in the negative direction, and those $i \in \{1, \ldots, m\} - (X^+ \cup X^-)$ will be zero vectors pointing nowhere. The secondary fan $\mathcal{F}'(\mathcal{A})$ decomposes the $\mathbb{R}^1$ into two cones: the two rays emanating from the origin in the positive and negative directions. These rays correspond to the two triangulations of $\mathcal{A}$ which differ by a bistellar operation supported on C. At a particular temperature and pressure, the Gibbs energy of the ensemble of products/reactants, whichever is lower, will force the reaction to run in one direction or another, so that the stable assemblages will correspond to the simplices of one or the other triangulation. In this case, the phase diagram consists of two divariant regions separated by the univariant curve corresponding to the single reaction.

When $m = n+2$, things start to become interesting. First, we can assume without loss of generality that there are no *indifferent phases*,[8] that is, every phase participates in some possible reaction or phase change. By excluding indifferent phases, we know that the Gale diagram $\mathcal{A}^*$ has m nonzero Gale vectors $a_1^*, \ldots, a_m^*$, although it is possible that some differ by positive scalar multiples and hence point in the same direction.[9] The pointed secondary fan $\mathcal{F}'(\mathcal{A})$ will look very similar to the Gale diagram, having at most m rays emanating from the origin, pointing in the directions of the Gale vectors, and two-dimensional cones lying between cyclically adjacent Gale vectors. According to Proposition 5.3, the phase diagram should look roughly like a two-dimensional slice of this two-dimensional pointed secondary fan $\mathcal{F}'(\mathcal{A})$, that is, like $\mathcal{F}'(\mathcal{A})$ itself. Hence the phase diagram will closely resemble the Gale diagram $\mathcal{A}^*$.

Roughly speaking, geochemists have known some version of this, in the guise of a method for constructing their *invariant point maps* as schematic representations of the local picture around an invariant point, based on knowledge of the minimal reactions possible among the phases. Their method uses Schreinemakers' *fundamental axiom*

[8] An indifferent phase would give rise to an element a_i of the oriented matroid for $\mathcal{A}$, known as an *isthmus* or *coloop*, and also to a zero Gale vector $a_i^* = 0$.

[9] This will happen whenever there are affine hyperplanes in $\mathbb{R}^{n-1}$ that contain all but two of the points of $\mathcal{A}$.

[25], [30], which is a reformulation of the oriented matroid duality assertion that the circuits of $\mathcal{A}$ coincide with the cocircuits of $\mathcal{A}^*$. The axioms assert that

- each phase a_i should label some univariant reaction half-line emanating from the invariant point, corresponding to a minimal reaction among the remaining phases other than a_i, and
- the extension of this half-line to a line through the origin should separate the other univariant reaction half-lines into those corresponding to the two sides of the reaction in question.

In other words, each Gale vector a_i^* lies on a line through the origin corresponding to a cocircuit of $\mathcal{A}^*$, which corresponds to a circuit of $\mathcal{A}$.

Using this rule, one can sketch the invariant point map by proceeding through the list of minimal reactions among the phases and using the axiom to place the half-lines around each other in cyclic order. There is an initial choice of orientation one must make for the diagram using the first reaction (should the reactant/products/missing-phase-half-line go in *clockwise* or *counterclockwise* order around the invariant point?), but after that the picture is determined. To decide which orientation is consistent with the actual geochemical phase diagram (i.e., to determine the actual placement of the image surface $\gamma(\mathbb{R}^2)$ within the secondary fan $\mathcal{F}(\mathcal{A})$), some thermodynamic data is required.

Example 8.1. Figure 6(c) shows the invariant point map constructed for the corundum-diaspore-gibbsite-water example. Note that we have used the geochemical conventions of labelling the univariant reaction half-lines emanating from the invariant point by the phase(s) missing from the reaction, putting the product/reactants on either side of the line, and indicating with dashes the *metastable* extensions of these half-lines.

9. The case $m = n + 3$: Phase diagram = affine Gale diagram. We next examine in detail the structure of the Gale diagram $\mathcal{A}^*$, the pointed secondary fan $\mathcal{F}'(\mathcal{A})$, and the phase diagram when $m = n + 3$. The conclusion is that two methods used by geochemists to reduce an essentially three-dimensional picture to two dimensions have parallel constructions in discrete geometry, and the phase diagram bears a close resemblance to a two-dimensional affine Gale diagram.

When $m = n + 3$, we can again assume without loss of generality that there are no indifferent phases, and hence no zero Gale vectors a_i^*. However, we make no other genericity assumptions for the moment. The Gale diagram $\mathcal{A}^*$ is a vector configuration in $\mathbb{R}^3$. As before, some Gale vectors may differ by a positive scalar multiple and hence give rise to the same ray in the secondary fan $\mathcal{F}'(\mathcal{A})$, so we know there will be at most m such rays. Note that, unlike the case $m = n+2$, here the cones of the pointed secondary fan $\mathcal{F}'(\mathcal{A})$ can be more exotic in shape: they are intersections of the three-dimensional simplicial cones spanned by linearly independent subsets of $\mathcal{A}^*$, and hence can have arbitrary polygonal cross sections.

The one-dimensional cones (rays) in $\mathcal{F}'(\mathcal{A})$ will correspond to zero-dimensional (point) regions in the phase diagram when they intersect the image surface $\gamma(\mathbb{R}^2)$ of the Gibbs energy map. As discussed in subsection 7.2, these points will either be indifferent crossings or invariant points. Since the invariant points correspond to chemical subsystems $\mathcal{A}' \subset \mathcal{A}$ which have $m' = n' + 2$ if we have $n' = n$ (as happens generically), then $|\mathcal{A}'| = |\mathcal{A}| - 1$; that is, there is exactly one phase a_i missing from $\mathcal{A}'$, and the corresponding ray in $\mathcal{F}'(\mathcal{A})$ is spanned by the Gale vector a_i^*. This is the reason that invariant points in phase diagrams with $m = n + 3$ are generically labelled by the single phase missing from the invariant equilibrium at that point. As

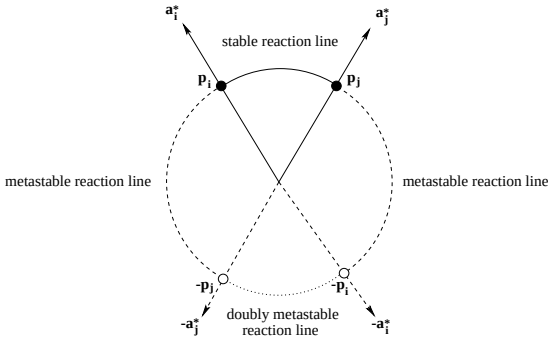


FIG. 9. *A typical reaction loop defined by two Gale vectors* $\mathbf{a}_i^*$ *and* $\mathbf{a}_j^*$.

also discussed in subsection 7.2, the local structure about the invariant point will look like the invariant point map for the $m' = n' + 2$ subsystem $\mathcal{A}'$.

Vector configurations in $\mathbb{R}^3$ (like $\mathcal{A}^*$) can be difficult to visualize. We discuss two methods that have been commonly used to cut down the dimension by one (and produce a picture closer in spirit to the phase diagram): the *spherical representation* and *affine Gale diagrams*.

9.1. The spherical representation: Closed nets. Intersecting the pointed secondary fan $\mathcal{F}'(\mathcal{A})$ in $\mathbb{R}^3$ with a unit sphere centered about the origin gives a useful spherical representation, similar to what has been called a *closed net* in [29]. In the conventions for the closed net, one includes not only the point of intersection with the sphere $p_i := a_i^*/|a_i^*|$ for each ray spanned by a Gale vector a_i^* (represented by a black dot labelled by the corresponding phase a_i), but also its negation $-p_i$ (represented by a white dot labelled similarly).[10] Furthermore, the arc representing the intersection curve on the sphere of a two-dimensional cone in $\mathcal{F}'(\mathcal{A})$ is augmented to be part of a great circle called a *reaction loop*, corresponding to the unique minimal reaction (circuit of $\mathcal{A}$, cocircuit of $\mathcal{A}^*$) to which it is associated. Typically such a reaction will involve all but two phases a_i, a_j (although this will not always be the case when $\mathcal{A}$ is not generic and therefore has some circuits of smaller support). As one traverses such a typical reaction loop, one passes through four arcs, as depicted in Figure 9.

The point of the closed net representation is that a hemispherical or planar projection of it from some angle should give a schematic picture of the actual phase diagram. Which projection occurs in nature will depend upon the location and orientation of the image surface $\gamma(\mathbb{R}^2)$ of the Gibbs energy map from section 5 inside the secondary fan $\mathcal{F}(\mathcal{A})$. Under our Assumption 5.4, one of each pair $\{p_i, -p_i\}$ will appear in the projection, and there are four possibilities for the portion of a typical reaction loop that will appear in the projection, depicted in Figure 10.

Example 9.1. We add a fifth phase (different from the ice added in Example 7.6) to our original example of corundum, diaspore, gibbsite, and water: the mineral *boehmite* (B), which is a *polymorph* of diaspore, that is, has the same chemical formula AlO(OH) but a different crystal structure. Thus B, D become parallel elements in the oriented matroid $\mathcal{M}$ for this new point configuration $\mathcal{A}$, having $m = 5$ and $n = 2$, so that $m = n + 3$.

[10] This supplementation of the Gale diagram $\mathcal{A}^*$ by adding in negations of all its vectors is reminiscent of the *Lawrence construction* [2, section 9.3] in oriented matroid theory.

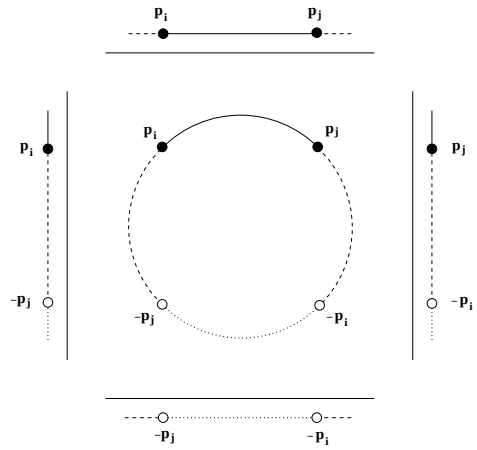


FIG. 10. *The four possible projections of a reaction loop onto an affine* 2*-plane.*

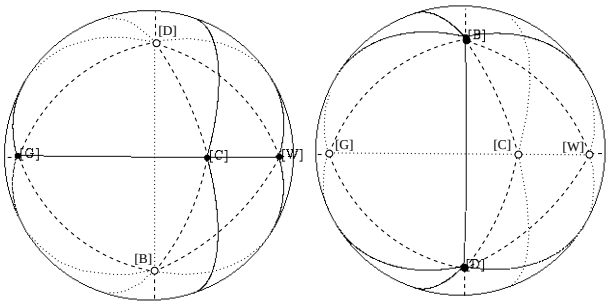


FIG. 11. *Two opposing hemispheric views of the closed net for the system with phases corundum, boehmite, diaspore, gibbsite, and water.*

We have

$$\mathcal{A} = \begin{array}{c} \begin{matrix} C & D & B & \frac{1}{2}G & W \end{matrix} \\ \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{3}{4} & 1 \end{bmatrix} \end{array}, \tag{6}$$

and a valid Gale transform is

$$\mathcal{A}^* = \begin{array}{c} \begin{matrix} C^* & D^* & B^* & G^* & W^* \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & -4 & 3 \\ 0 & 1 & 0 & -2 & 1 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix} \end{array}. \tag{7}$$

Two opposite hemispheric views of the closed net for this example are depicted in Figure 11. Note that the parallel elements B, D in $\mathcal{A}$ give rise to a circuit

$$\begin{matrix} C & D & B & G & W \\ 0 & + & - & 0 & 0 \end{matrix}$$

that corresponds to a cocircuit of $\mathcal{A}^*$: the Gale vectors C^*, G^*, W^* are coplanar, and

their corresponding points on the closed net lie on a great circle which separates D^* from B^*.

9.2. Two-dimensional affine Gale diagrams. The affine Gale diagram is simply an affine point configuration in $\mathbb{R}^2$ used to encode the three-dimensional vector configuration $\mathcal{A}^*$; see [33, Definition 6.17], [2, section 9.1]. Arbitrarily choose a two-dimensional affine plane Γ in $\mathbb{R}^3$ to "slice" the Gale vectors: if this plane Γ is defined by the equation $f(x) = c$ for some generic linear functional $f \in (\mathbb{R}^3)^*$ and some positive value c, then we replace each Gale vector a_i^* by the unique point $ca_i^*/f(a_i^*)$ in its span that lies in this plane Γ. Color these rescaled Gale points in Γ black or white, depending upon whether $f(a_i^*) > 0$ or $f(a_i^*) < 0$. Since $\mathcal{A}$ was an affine point configuration and hence $\mathcal{A}^*$ is a totally cyclic vector configuration, there will always be both black and white points in the affine Gale diagram, regardless of how the functional f is chosen.

We can further annotate the affine Gale diagram by drawing in line segments that correspond to the intersections of two-dimensional cones from $\mathcal{F}'(\mathcal{A})$ with the plane Γ that happen to connect the black points in the diagram. Bearing in mind Assumption 5.4, the choice of the functional f (equivalently, the choice of the plane Γ) corresponds to the choice of the location of the image surface $\gamma(\mathbb{R}^2)$ of the Gibbs energy map. It follows from Proposition 5.3 that this "decorated affine Gale diagram" is a schematic picture for one possible toplogy of the phase diagram. Such schematic pictures, when annotated further with more arcs of reaction loops using conventions similar to the closed nets discussed in subsection 9.1 above, have appeared in [17] and are called *potential solutions* for the phase diagram topology.

When are two such affine Gale diagrams/potential solutions considered "equivalent"? Fortunately, discrete geometers and geochemists agree on this answer: when the assignment of either a black or a white dot to each phase is the same. Equivalently, this means they have the same sign vector $(\text{sign}(f(a_1^*)), \ldots, \text{sign}(f(a_m^*))) \in \{+,-\}^m$ or, in oriented matroid terminology, that f achieves the same *acyclic (re)orientation* (or *tope*) of the vector configuration $\mathcal{A}^*$ (see [2, section 3.8]). This turns out to have the following geometric reinterpretation: if we regard the functional $f(x) = f_1x_1 + f_2x_2 + f_3x_3$ as its vector of coefficients (f_1, f_2, f_3), then the acyclic orientation achieved by f is determined by which side of each of the hyperplanes $(a_i^*)^\perp$ normal to the Gale vectors it lies on. Therefore intersecting this arrangement of hyperplanes $(a_i^*)^\perp$ with the unit sphere in $\mathbb{R}^3$ gives an arrangement of great circles on the sphere (called the *Euler sphere* in [16]), whose two-dimensional regions parametrize the different acyclic orientations/affine Gale diagrams/potential solutions.

The method developed in [16] of constructing potential solutions for systems with $m = n+3$ systems involved looking at each phase, using the method of Schreinemakers from section 8 to infer the local structure/invariant point map about the invariant point at which that phase is missing from the assemblage, and then "fitting together" these various invariant point maps to produce the straight line net.

Example 9.2. Figure 12 shows a view of the Euler sphere for the previous example with $m = n + 3$, along with two of its regions labelled by their corresponding affine Gale diagrams/potential solutions.

There is a good bit of theory to help one enumerate these acyclic reorientations (see [2, Theorem 4.6.1] and subsection 10.2 below) or to produce a list of them all algorithmically using a straightforward application of Farkas' lemma [26, section 7.3]. In the Java applet CHEMOGALE [18], such an algorithm is part of the implementation. For systems input by the user with $m = n+2$ or $n+3$, the program computes $\mathcal{A}^*$

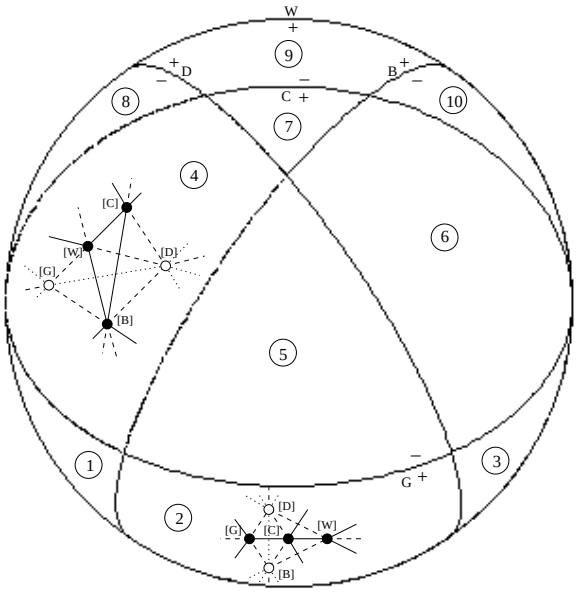


FIG. 12. *The hemisphere of the Euler sphere lying in one side of* $(W^*)^\perp$ *for the* $m = n + 3$ *system with phases corundum, boehmite, diaspore, gibbsite, and water. Two regions are shown labelled by the corresponding affine Gale diagrams/potential solutions to phase diagram topology.*

and uses data generated by [20] to obtain $\mathcal{F}'(\mathcal{A})$. When $m = n + 3$, the intersection of $\mathcal{F}'(\mathcal{A})$ with the unit sphere is depicted, allowing the user to select two-dimensional cones of $\mathcal{F}'(\mathcal{A})$ and receive the corresponding triangulation. For $m = n + 3$ systems, the user may also view the Euler sphere and see the potential solution to the phase diagram topology associated with each region. This work was fully described in [19, Chapter 3].

10. Further implications/applications. We collect here a few further implications/applications of some of the theory developed.

10.1. Slopes around invariant points. Let p be an invariant point in the phase diagram, and let $\{a_1, \ldots, a_k\}$ be the union of all sets of phases that can form stable assemblages at p. There will be at most k univariant reaction curves emanating from p corresponding to reactions that omit each of the phases a_i, and each has a limiting slope μ_i as it enters p. Doing the experiments to determine these slopes accurately is expensive and time-consuming, so it is helpful to be able to determine the slopes from as little data as possible.

PROPOSITION 10.1. *Knowing the formulae* $\{a_1, \ldots, a_k\}$ *of the phases and knowing three different limiting slopes* $\mu_{i_1}, \mu_{i_2}, \mu_{i_3}$ *determines all of the slopes* $\mu_1, \ldots, \mu_k$.

Proof. As discussed in section 7.2, the local structure of the phase diagram about p coincides with the phase diagram for only the subsystem of phases in $\mathcal{A}' := \{a_1, \ldots, a_k\}$, and this must be a system with $k = n+2$ phases. Let its Gale transform be $\mathcal{A}'^* := \{a_1^*, \ldots, a_k^*\}$, so that, up to an invertible linear change-of-basis in $\mathbb{R}^2$, these give the slopes of the rays emanating from the origin in the pointed secondary fan $\mathcal{F}'(\mathcal{A}')$. By Assumption 5.4, the slopes $\{\mu_1, \ldots, \mu_k\}$ are also related to the slopes of these rays in the pointed secondary fan by an invertible linear change-of-basis (namely,

the Jacobian matrix of the Gibbs energy map $\gamma : \mathbb{R}^2 \to \mathbb{R}^m$ evaluated at the point $p \in \mathbb{R}^2$, composed with the linear projection map from $\mathbb{R}^m \to \mathbb{R}^{m-n}$ that sends the secondary fan to the pointed secondary fan). Hence the (known) slopes of the Gale vectors in $\mathcal{A}'^*$ are related to the limiting slopes about p by an invertible linear change-of-basis. Therefore knowledge of three distinct slopes $\mu_{i_1}, \mu_{i_2}, \mu_{i_3}$ will determine any other slope μ_{i_r}, e.g., by invariance under invertible linear transformations of the *cross-ratio*

$$(\mu_{i_1}, \mu_{i_2} \mid \mu_{i_3}, \mu_{i_r}) := \frac{(\mu_{i_r} - \mu_{i_1})(\mu_{i_3} - \mu_{i_2})}{(\mu_{i_r} - \mu_{i_2})(\mu_{i_3} - \mu_{i_1})}. \qquad \square$$

Example 10.2. In the example of corundum, diaspore, gibbsite, and water, which had Gale diagram

$$\mathcal{A}^* = \begin{array}{c} \begin{matrix} C^* & D^* & G^* & W^* \end{matrix} \\ \begin{bmatrix} 1 & -3 & 2 & 0 \\ 2 & -4 & 0 & 2 \end{bmatrix} \end{array}, \tag{8}$$

we see that the slopes of C^*, D^*, G^*, W^* are $2, \frac{4}{3}, 0, \infty$, giving the cross-ratio

$$(\mu_{C^*}, \mu_{G^*} \mid \mu_{W^*}, \mu_{D^*}) = \frac{(\frac{4}{3} - 2)(\infty - 0)}{(\frac{4}{3} - 0)(\infty - 2)} = -\frac{1}{2}.$$

Thus if we have already determined (say, from thermodynamic data) that the phase diagram has limiting slopes $\mu_{[C]}, \mu_{[G]}, \mu_{[W]}$ for the three reaction curves labelled $[C], [G], [W]$ entering the invariant point, then the limiting slope $\mu_{[D]}$ of the fourth reaction curve labelled $[D]$ will satisfy

$$-\frac{1}{2} = (\mu_{[C]}, \mu_{[G]} \mid \mu_{[W]}, \mu_{[D]}) = \frac{(\mu_{[D]} - \mu_{[C]})(\mu_{[W]} - \mu_{[G]})}{(\mu_{[D]} - \mu_{[G]})(\mu_{[W]} - \mu_{[C]})},$$

which can be solved for $\mu_{[D]}$, giving the formula

$$\mu_{[D]} = \frac{3\mu_{[C]}\mu_{[G]} - 2\mu_{[C]}\mu_{[W]} - \mu_{[G]}\mu_{[W]}}{2\mu_{[G]} - 3\mu_{[W]} + \mu_{[C]}}.$$

10.2. Counting potential solutions. As mentioned in section 9.2, there is theory available for counting the acyclic orientations of a vector configuration (or oriented matroid) such as the Gale diagram $\mathcal{A}^*$. Here we elaborate on this and explain how to easily count potential solutions to phase diagram topology when $m = n + 3$.

As we saw in section 9.2, counting the potential solutions to phase diagram topology amounts to counting the three-dimensional cones cut out by an arrangement of planes through the origin in $\mathbb{R}^3$, or the two-dimensional regions cut out by an arrangement of great circles on a sphere, or the acyclic orientations of the oriented matroid $\mathcal{M}^*$ associated to the Gale diagram $\mathcal{A}^*$. The problem of counting the n-dimensional regions cut out by an arrangement of $(n-1)$-dimensional hyperplanes through the origin in $\mathbb{R}^n$ was treated first by Winder in 1966, and then later independently by both Las Vergnas and Zaslavsky around 1975. The basic idea is that even though the combinatorial structure of the regions cut out (e.g., how many faces of each dimension they have) depends on the associated *oriented* matroid, the number of regions only

depends on the coarser information recorded in the *matroid*. We review some of this material here; see [2, section 4.6] for a fuller treatment.

For example, one way to record the matroid data associated to $\mathcal{A}^*$ is to list all of its (unsigned) *circuits*, which are the support sets of minimal linear dependences (with no record of the signs of the coefficients in the dependence). A more useful way to encode the data for counting the regions (but equivalent data to specifying the circuits) is to write down the *lattice of flats* $L(\mathcal{A}^*)$, which is the partial ordering by inclusion of all subspaces spanned by subsets of $\mathcal{A}^*$. The bottom element of this partially ordered set, called $\hat{0}$, corresponds to the zero subspace (spanned by the empty set of Gale vectors).

The *Möbius function* $\mu(x, y)$ is an integer associated to each pair of elements $x \leq y$ which are related in the partial order L, defined recursively by these properties:

$$\mu(x, x) = +1,$$
$$\mu(x, y) = -\sum_{x \leq z < y} \mu(x, z) \quad \text{if } x < y.$$

One can use this to count regions via the following result.

THEOREM 10.3 (see [2, Theorem 4.6.1]). *The number of regions cut out by the hyperplanes normal to a collection of vectors with lattice of flats L is*

$$\sum_{x \in L} |\mu(\hat{0}, x)|.$$

Since we wish to apply this to geochemical systems with $m = n + 3$, we detail here explicitly (in more concrete terms) what happens in this case.

PROPOSITION 10.4. *Let $\mathcal{A}$ be a chemography with $m = n + 3$, so that its Gale diagram $\mathcal{A}^*$ is a configuration of vectors in $\mathbb{R}^3$. Then the number of potential solutions to phase diagram topology is*

$$2\left(1 + \sum_{P} (m_P - 1)\right),$$

where P runs through all two-dimensional planes spanned by pairs of the Gale vectors $\mathcal{A}^*$, and m_P is the number of distinct lines spanned by Gale vectors lying in the plane P (or, equivalently, the number of parallelism classes of Gale vectors within the plane P).

In particular, if $\mathcal{A}$ is in a general position (in the sense that every subset of n elements in $\mathcal{A}$ is affinely independent or, equivalently, every minimal reaction among the phases involves at least $n+1$ phases), the number of potential solutions is (cf. [17]*)*

$$2\left(1 + \binom{n}{2}\right).$$

Proof. Since $\mathcal{A}^*$ lives in $\mathbb{R}^3$, there are four kinds of elements x in the lattice of flats:

- $x = \hat{0}$, having $\mu(\hat{0}, \hat{0}) = +1$,
- $x = \ell$, a line spanned by a Gale vector a_i^*, having $\mu(\hat{0}, \ell) = -1$,
- $x = P$, a two-dimensional plane spanned by Gale vectors, having

$$\mu(\hat{0}, P) = -\left(\mu(\hat{0}, \hat{0}) + \sum_{\ell \subset P} \mu(\hat{0}, \ell)\right) = m_P - 1, \qquad \text{and}$$

- $x = \mathbb{R}^3$, having

$$\begin{aligned}\mu(\hat{0}, \mathbb{R}^3) &= -\left(\mu(\hat{0}, \hat{0}) + \sum_{\ell} \mu(\hat{0}, \ell) + \sum_{P} \mu(\hat{0}, P)\right) \\ &= -1 + \#\{\text{lines } \ell \text{ spanned by Gale vectors}\} \\ &\quad + \sum_{P} (m_P - 1).\end{aligned}$$

Adding the absolute values of all of these gives the result stated in the proposition.

In the generic case, since $\mathcal{A}$ has n points, there will be n distinct Gale vectors $\mathcal{A}^*$, no two of which span the same line ℓ, and there will be $\binom{n}{2}$ different planes P spanned by them. ($\mathcal{A}$ is generic if and only if $\mathcal{A}^*$ is generic by matroid duality.) Also, each of these planes will contain exactly two lines ℓ, so $m_P - 1 = 1$. The second assertion follows from plugging these values into the first equation. □

Example 10.5. In Figure 12, the Euler sphere shown has 20 regions total (10 visible on the hemisphere shown and 10 more on the "underside"). One can compare this with the formula predicted in Proposition 10.4, which can be evaluated with the aid of the closed net picture of Gale vectors in Figure 11. This figure shows (the intersection with the sphere of) eight planes P spanned by pairs of Gale vectors, namely,

$$BC, BG, BW, DG, DC, DW, BD, BGW,$$

of which the first seven have $m_P = 2$, and the last has $m_P = 3$. Thus the proposition would predict

$$2(1 + 7 \cdot (2 - 1) + 1 \cdot (3 - 1)) = 20$$

regions, as expected.

REFERENCES

[1] M. Azaola and F. Santos, *The graph of triangulations of a point configuration with $d + 4$ vertices is 3-connected*, Discrete Comput. Geom., 23 (2000), pp. 489–536.

[2] A. Björner, M. Las Vergnas, B. Sturmfels, N. White, and G. Ziegler, *Oriented Matroids*, Cambridge University Press, Cambridge, UK, 1993.

[3] L. J. Billera, P. Filliman, and B. Sturmfels, *Constructions and complexity of secondary polytopes*, Adv. Math., 83 (1990), pp. 155–179.

[4] L. J. Billera, I. M. Gelfand, and B. Sturmfels, *Duality and minors of secondary polyhedra*, J. Combin. Theory Ser. B, 57 (1993), pp. 258–268.

[5] R. G. Bland and M. Las Vergnas, *Orientability of matroids*, J. Combin. Theory, B 24 (1978), pp. 94–123.

[6] H. W. Day, *Geometric analysis of phase equilibria in ternary systems of six phases*, Amer. J. Sci., 272 (1972), pp. 711–734.

[7] J. A. de Loera, S. Hoşten, F. Santos, and B. Sturmfels, *The polytope of all triangulations of a point configuration*, Doc. Math., 1 (1996), pp. 103–119.

[8] J. Folkman and J. Lawrence, *Oriented matroids*, J. Combin. Theory, B 25 (1978), pp. 199–236.

[9] D. Gale, *Neighboring vertices on a convex polyhedron*, in Linear Inequalities and Related Systems, H. W. Kuhn and A. W. Tucker, eds., Ann. of Math. Stud. 38, Princeton University Press, Princeton, NJ, 1956, pp. 255–263.

[10] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky, *Discriminants, Resultants, and Multidimensional Determinants*, Birkhäuser Boston, Cambridge, MA, 1994.

[11] B. Grünbaum, *Convex Polytopes*, Wiley-Interscience, New York, 1967.

[12] B. Guy and J. M. Pla, *Structure of phase diagrams for n-component (n+k)-phase chemical systems: The concept of affigraphy*, C. R. Acad. Sci., 324 (1997), pp. 1–7.
[13] T. J. B. Holland and R. Powell, *An enlarged and updated internally consistent thermodynamic data set with uncertainties and correlations in the system* $K_2O - Na_2O - CaO - MgO - MnO - FeO - Fe_2O_3 - Al_2O_3 - TiO_2 - SiO_2 - C - H_2 - O_2$, J. Metamorphic Geol., 8 (1990), pp. 89–124.
[14] C. L. Lawson, *Transforming triangulations*, Discrete Math., 3 (1972), pp. 365–372.
[15] C. W. Lee, *Regular triangulations of convex polytopes*, in Applied Geometry and Discrete Mathematics, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 4, AMS, Providence, RI, 1991, pp. 443–456,
[16] G. Kletetschka and J. H. Stout, *Stability analysis of invariant points using Euler spheres, with an application to FMAS granulites*, J. Metamorphic Geol., 17 (1999), pp. 1–14.
[17] R. E. Mohr and J. H. Stout, *Multisystem nets for systems of* $n+3$ *phases*, American J. Sci., 280 (1980), pp. 143–172.
[18] S. W. Peterson, *CHEMOGALE: A Java applet for computing geochemical phase diagrams,* 2000, available online at www.math.umn.edu/∼reiner/CHEMOGALE.html.
[19] S. W. Peterson, *Oriented Matroid Analysis of Thermochemical Reaction Systems*, Masters thesis, Department of Mathematics, University of Minnesota, Minneapolis, 2000.
[20] J. Rambau, *TOPCOM: Triangulations of Point Configurations and Oriented Matroids,* Version 0.2.0, available online at www.zib.de/rambau/TOPCOM.html.
[21] J. E. Ricci, *The Phase Rule and Heterogeneous Equilibrium*, D. Van Nostrand Company, New York, 1951.
[22] F. Santos, *A point set whose space of triangulations is disconnected*, J. Amer. Math. Soc., 13 (2000), pp. 611–637.
[23] F. Santos, *Non-connected Toric Hilbert Schemes*, preprint, available online at http://arXiv.org/abs/math.CO/0204044.
[24] F. Santos, *Triangulations of oriented matroids*, Mem. Amer. Math. Soc., 156 (2002), paper 741.
[25] F. A. H. Schreinemakers, *In-, mono-, and di-variant equilibria*, Konink. Acad. Wetensch. Amsterdam, Vols. 18–28 (29 separate articles in a series), 1915–1925.
[26] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley-Interscience, New York, 1998.
[27] V. S. Sheplev, *Construction of chemographic diagrams*, Petrology, 4 (1996), pp. 97–102.
[28] S. I. Usdansky, *Some topological and combinatorial properties of c component* $(c+4)$ *phase multisystem nets*, Math. Geol., 19 (1987), pp. 793–805.
[29] E. Zen, *Some topological relationships in multisystems of* $n+3$ *phases* I. *General theory; unary and binary systems*, American J. Sci., 264 (1966), pp. 401–427.
[30] E. Zen, *Construction of pressure-temperature diagrams for multicomponent systems as the method of Schreinemakers—A geometric approach*, U.S. Geological Survey Bulletin 1225, 1966.
[31] E. Zen, *Some topological relationships in multisystems of* $n+3$ *phases* II. *Unary and binary metastable sequences*, American J. Sci., 265 (1967), pp. 871–897.
[32] E. Zen and E. H. Rosenboom, Jr., *Some topological relationships in multisystems of* $n+3$ *phases* III. *Ternary systems*, American J. Sci., 272 (1972), pp. 677–710.
[33] G. M. Ziegler, *Lectures on Polytopes*, Springer-Verlag, New York, 1995.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 260–276

GLOBAL RESULTS FOR AN EPIDEMIC MODEL WITH VACCINATION THAT EXHIBITS BACKWARD BIFURCATION*

JULIEN ARINO†, C. CONNELL MCCLUSKEY†, AND P. VAN DEN DRIESSCHE†

Abstract. Vaccination of both newborns and susceptibles is included in a transmission model for a disease that confers immunity. The interplay of the vaccination strategy together with the vaccine efficacy and waning is studied. In particular, it is shown that a backward bifurcation leading to bistability can occur. Under mild parameter constraints, compound matrices are used to show that each orbit limits to an equilibrium. In the case of bistability, this global result requires a novel approach since there is no compact absorbing set.

Key words. epidemic model, vaccination, backward bifurcation, compound matrices, global dynamics

AMS subject classifications. 92D30, 34D23

DOI. 10.1137/S0036139902413829

1. Introduction.

Vaccination is a commonly used method for controlling diseases, e.g., pertussis, measles, or influenza. Mathematical models including vaccination aid in deciding on a vaccination strategy and in determining changes in qualitative behavior that could result from such a control measure (see, e.g., [5, 6]). If the vaccine is not totally effective, then recent models show that a backward bifurcation is possible for some parameter values [9, 10]. In such a case, the basic reproduction number as modified by vaccination must be reduced below a certain threshold (that is less than one) in order to ensure that the disease dies out. Backward bifurcation has been observed in other disease transmission models, for example the HIV/AIDS models discussed in [2, 8] and the bovine respiratory syncytial virus model in [4].

Our model is a generalization of that of [10], allowing individuals recovering from the disease to go into a temporarily immune class rather than directly back into the susceptible class. A recent model [9] allows for a recovered class and considers vaccination for a disease that has acute and chronic infective stages as well as variable infectivity.

In section 2, we develop our model with general parameters, and illustrate its behavior in section 3 by using vaccination-related values appropriate for pertussis [1, 5]. In particular, we focus on the vaccination parameters and how changes in these may alter the qualitative behavior of the model by leading to subthreshold endemic states via backward bifurcation. Some local stability results are proved.

Previous investigations of the stability of subthreshold endemic states associated with backward bifurcations rely mainly on local results. We use compound matrices and geometric ideas to develop global results under mild parameter restrictions. These tools have been used for analyzing other models of disease transmission in which there is a unique endemic equilibrium; see, e.g., [11, 14, 16, 18]. In section 4, we present a brief summary of this geometric approach for studying the global dynamics of our model, concentrating on the novel features. This method is then used in section 5 to

*Received by the editors August 30, 2002; accepted for publication (in revised form) February 28, 2003; published electronically November 19, 2003. The research of the first and third authors was supported in part by an NSERC research grant and by MITACS.
http://www.siam.org/journals/siap/64-1/41382.html
†Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada V8W 3P4 (arino@math.mcmaster.ca, mccluskc@math.mcmaster.ca, pvdd@math.uvic.ca).

prove global results for the model. (Some technical details are placed in Appendices A and B.) Concluding remarks are given in section 6.

2. Formulation of an *SIRS* model with vaccination. Following [10], but with newborn vaccination and a recovered class, the model has the flow diagram given in Figure 2.1 with the following assumptions. Each of the N individuals can be in one of four states: susceptible, infective, recovered, and vaccinated; the numbers in these states are denoted by S, I, R, and V, respectively. Thus, $N = S + I + R + V$. Birth occurs in the system with rate constant $d > 0$. Of these newborns, a fraction $\alpha \in [0, 1]$ are vaccinated at birth. Death occurs with the same rate constant d as birth; thus the total population N is constant. The transmission coefficient β is the number of contacts made by one infective per unit time multiplied by the probability that a contact with a susceptible leads to infection. The disease is transmitted horizontally, with the transmission modeled using a standard incidence function; thus the rate at which susceptibles become infective is $\beta SI/N$. For contacts between infectives and vaccinated individuals this coefficient is multiplied by a factor $\sigma \in [0, 1]$. Thus $1 - \sigma$ is the vaccine efficacy. Susceptible individuals are vaccinated with rate constant ϕ, and the vaccine protection wanes with rate constant $\theta > 0$. Infective individuals recover with rate constant $\gamma > 0$ and then have temporary immunity. They leave the recovered state with rate constant ν. We assume $\alpha d + \phi > 0$ to ensure that there is a nonzero flow of individuals into class V.

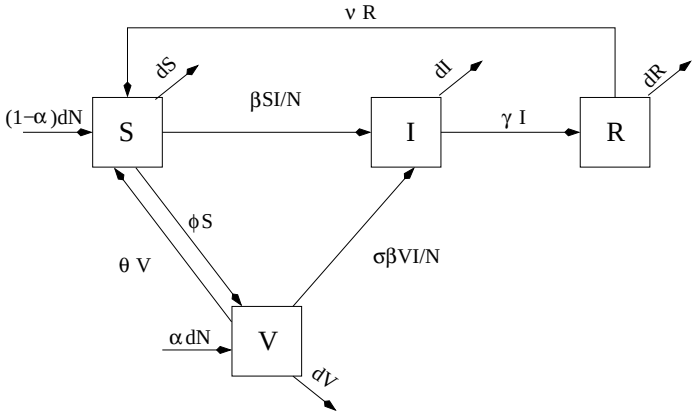


FIG. 2.1. *The flow diagram of the SIRV model.*

The model is formulated as the following system of ordinary differential equations:

$$\frac{dS}{dt} = (1-\alpha)dN - dS - \beta\frac{SI}{N} - \phi S + \theta V + \nu R, \tag{2.1a}$$

$$\frac{dI}{dt} = \beta\frac{SI}{N} + \sigma\beta\frac{VI}{N} - (d+\gamma)I, \tag{2.1b}$$

$$\frac{dR}{dt} = \gamma I - (d+\nu)R, \tag{2.1c}$$

$$\frac{dV}{dt} = \alpha dN + \phi S - (d+\theta)V - \sigma\beta\frac{VI}{N}, \tag{2.1d}$$

with nonnegative initial conditions and $N(0) > 0$.

System (2.1) is well posed: solutions remain nonnegative for nonnegative initial conditions. As the total population is constant, the system can be rewritten in terms

of proportions as

$$\frac{dS}{dt} = (1-\alpha)d - dS - \beta SI - \phi S + \theta(1-S-I-R) + \nu R, \tag{2.2a}$$

$$\frac{dI}{dt} = \beta SI + \sigma\beta(1-S-I-R)I - (d+\gamma)I, \tag{2.2b}$$

$$\frac{dR}{dt} = \gamma I - (d+\nu)R, \tag{2.2c}$$

$$V = 1-(S+I+R), \tag{2.2d}$$

where here S, I, R, V denote the proportions in the susceptible, infective, recovered, and vaccinated states, respectively. Conclusions about system (2.1) can be easily recovered from system (2.2), and we employ system (2.2) from now on. System (2.2a)–(2.2c) can be written as $dx/dt = f(x)$ with $x = (S, I, R)^T$.

In the case $\sigma = 1$, the vaccine is totally useless, and (2.2) reduces to an SIRS model without vaccination. The behavior is then determined by $\mathcal{R}_0 = \beta/(d+\gamma)$. This is the classical basic reproduction number in the SIRS model, namely, the average number of new infections caused by one infective (in a completely susceptible population) during the infective period. From now on we assume that $\sigma < 1$.

3. Equilibria and bifurcations. For system (2.2), there is always the disease-free equilibrium (DFE)

$$X_0 = (S_{DFE}, 0, 0, V_{DFE}) = \left(\frac{\theta + d(1-\alpha)}{d+\theta+\phi}, 0, 0, \frac{\phi + d\alpha}{d+\theta+\phi}\right). \tag{3.1}$$

Now consider endemic equilibria with $I = I^* > 0$. From (2.2b) at an endemic equilibrium, $\beta(S+\sigma V) = d+\gamma$. Since $S + \sigma V < 1$, this can be true only for $\beta > d+\gamma$; hence, for $\mathcal{R}_0 \le 1$ there exists no endemic equilibrium. For $\mathcal{R}_0 > 1$, the existence of endemic equilibria is determined by the presence in $(0, 1]$ of positive real solutions of the quadratic

$$P(I) = AI^2 + BI + C = 0,$$

with

$$A = -\sigma\beta^2\frac{d+\nu+\gamma}{d+\nu},$$

$$B = \sigma\beta^2 - \beta(d+\theta+\sigma(d+\gamma+\phi)) - \frac{\beta\gamma}{d+\nu}(d+\theta+\sigma\phi),$$

$$C = (d+\theta+\sigma\phi - d\alpha(1-\sigma))\beta - (d+\gamma)(d+\theta+\phi).$$

Thus, depending on parameter values, the number of endemic equilibria is zero, one, or two. For $\sigma = 0$ (the vaccine is totally effective), at most one endemic equilibrium is possible. From now on we make the realistic assumption that the vaccine is not totally effective, and thus $0 < \sigma < 1$. From (2.2a)–(2.2d), it can be shown that if I^* is a positive solution of $P(I) = 0$, then S^*, R^*, and V^* are positive; thus the equilibrium is biologically relevant. For a positive real solution I^* to $P(I) = 0$, the endemic equilibrium point (EEP) in system (2.2) is given by

$$(S^*, I^*, R^*, V^*) = \left(\frac{(1-\alpha)d + \frac{(\nu-\theta)\gamma I^*}{d+\nu} + (1-I^*)\theta}{d+\beta I^*+\phi+\theta}, I^*, \frac{\gamma I^*}{d+\nu}, 1-S^*-I^*-R^*\right).$$

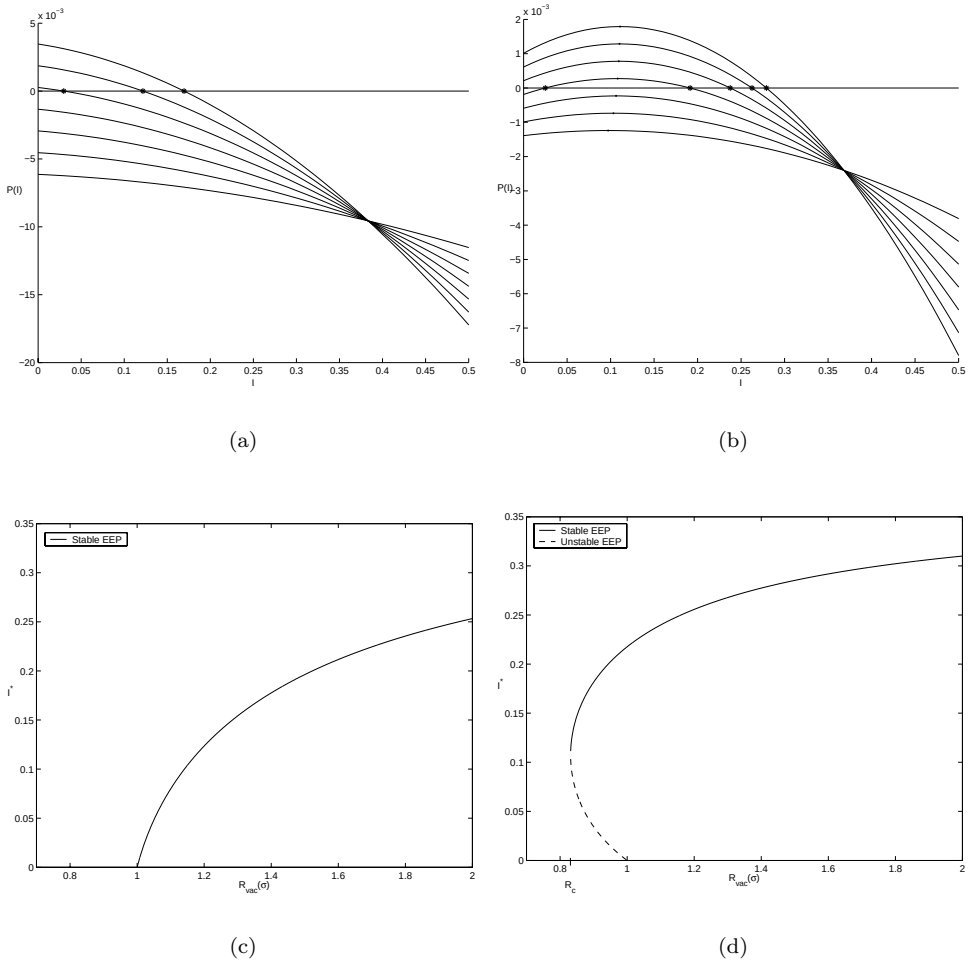


FIG. 3.1. (a) *Plot of the quadratic $P(I)$, with increasing values of σ (at left, from bottom to top, $\sigma = 0.04, 0.06, \ldots, 0.16$) in the forward bifurcation case, $\phi = 0.2$.* (b) *As* (a) *but in the backward bifurcation case, $\phi = 0.05$.* (c) *Bifurcation in the $(\mathcal{R}_{vac}(\sigma), I^*)$-plane, $\phi = 0.2$.* (d) *Bifurcation in the $(\mathcal{R}_{vac}(\sigma), I^*)$-plane, $\phi = 0.05$.*

In Figure 3.1, $P(I)$ is plotted versus I for increasing values of σ and $\phi = 0.2$ or 0.05 (all other parameters being fixed at the values indicated in Table 3.1). The values of γ, d and the vaccination parameters of Table 3.1 are appropriate for pertussis [1, 5], whereas β and ν are estimated to illustrate our bifurcation results. Figure 3.1(a) shows the situation that prevails when the bifurcation is a classical forward one. In this case, an increase in σ through some critical value σ_c (which depends on the other parameter values) leads through a transcritical bifurcation to a unique endemic equilibrium. Figure 3.1(b) shows the occurrence of the backward bifurcation. In this case, an increase of σ leads to the curve $P(I)$ becoming tangent to the horizontal axis defining a critical value σ_c at a saddle-node bifurcation. As σ becomes larger than σ_c, two equilibria exist. We expect bistability with the DFE and the equilibrium with the larger I value being stable. As σ increases further, the equilibrium with the smaller

TABLE 3.1
Parameter values used in simulations.

Parameter	Typical value or range	Meaning
β	0.4 /day	Transmission coefficient
γ	1/(21 days)	Average infectious period 21 days
d	1/(75 years)	Average lifespan 75 years
ν	1/(31 days)	Average period of immunity 31 days
α	0.9	Proportion of vaccinated newborns
ϕ	0.05 to 0.2/day	Vaccination rate constant
σ	0.04 to 0.2	Vaccine is between 80% and 96% effective
θ	1/(5 years)	Average vaccine waning time 5 years

I value moves to the left. When this equilibrium leaves the positive orthant through a transcritical bifurcation with the DFE, there is only one endemic equilibrium.

Since the concavity of the quadratic $P(I)$ is fixed (as $A < 0$), observation of Figure 3.1(b) gives necessary conditions for the existence of two equilibria: $P'(0) = B > 0$ and $P(0) = C < 0$. Together with the fact that the roots of $P(I)$ are real, this gives the bistability region $B > 0$, $C < 0$ and $\Delta = B^2 - 4AC > 0$. Figure 3.2 shows this region as a function of σ and ϕ, with all other parameters fixed as in Table 3.1.

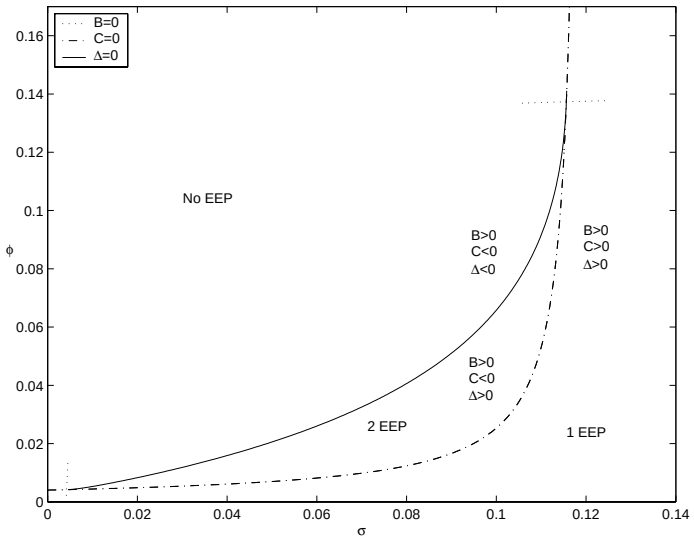


FIG. 3.2. *Bifurcation diagram in the (σ, ϕ)-plane.*

Using for example the method of [20], the basic reproduction number as modified by vaccination is

$$\mathcal{R}_{vac} = \frac{\beta}{d+\gamma}\left(S_{DFE} + \sigma V_{DFE}\right),$$

which from (3.1) gives

$$\mathcal{R}_{vac} = \mathcal{R}_0 \frac{d+\theta+\sigma\phi-d\alpha(1-\sigma)}{d+\theta+\phi}. \tag{3.2}$$

We write $\mathcal{R}_{vac}(\sigma)$ to indicate σ as the bifurcation parameter when all other parameters are fixed. Note that $\mathcal{R}_0(d(1-\alpha)+\theta)/(d+\theta+\phi) < \mathcal{R}_{vac}(\sigma) < \mathcal{R}_0$ (equalities are achieved at $\sigma = 0$ and $\sigma = 1$, respectively). The constant term C in the polynomial $P(I)$ can be written as $(d+\gamma)(d+\theta+\phi)(\mathcal{R}_{vac}-1)$; thus $P(0)$ has the same sign as $\mathcal{R}_{vac}-1$. Define $\mathcal{R}_c = \mathcal{R}_{vac}(\sigma_c)$. For a forward bifurcation, this gives $\mathcal{R}_c = 1$; see Figure 3.1(c). For a backward bifurcation, $\mathcal{R}_c < 1$; see Figure 3.1(d). The existence of endemic equilibria is summarized as follows.

PROPOSITION 3.1. *For model* (2.2)*, if* $\mathcal{R}_{vac} < \mathcal{R}_c$ *or* $\mathcal{R}_{vac} = \mathcal{R}_c = 1$*, there is no endemic equilibrium; if* $\mathcal{R}_c < \mathcal{R}_{vac} < 1$*, then there are two distinct endemic equilibria; if* $\mathcal{R}_c = \mathcal{R}_{vac} < 1$*,* $\mathcal{R}_c < \mathcal{R}_{vac} = 1$*, or* $\mathcal{R}_{vac} > 1$*, there is a unique endemic equilibrium.*

When two endemic equilibria are present, let X^* and X_* be the endemic equilibria with the larger and smaller value of I^*, respectively; when $\mathcal{R}_{vac} \neq \mathcal{R}_c$ and a unique endemic equilibrium exists, it is denoted by X^*; when $\mathcal{R}_{vac} = \mathcal{R}_c$ and a unique endemic equilibrium exists, it is denoted by X_c. A global result (for $\mathcal{R}_0 < 1$) and local stability of the equilibria are summarized in the following theorem, which justifies the stability of equilibria as shown in Figures 3.1(c) and 3.1(d).

THEOREM 3.2. *If* $\mathcal{R}_0 < 1$*, then the DFE* X_0 *is the only equilibrium for system* (2.2a)–(2.2c)*, and it is globally asymptotically stable;* X_0 *is locally asymptotically stable for* $\mathcal{R}_{vac} < 1$ *and unstable for* $\mathcal{R}_{vac} > 1$*. When present, the endemic equilibrium* X_* *is unstable, and if* $\theta \le \nu$*, then* X^* *is locally asymptotically stable.*

Proof. As remarked earlier, for $\mathcal{R}_0 \le 1$ there exists no endemic equilibrium. Further, if $\mathcal{R}_0 < 1$, then I can be used as a Lyapunov function to show that the DFE is globally asymptotically stable.

From [20, Theorem 2], $\mathcal{R}_{vac}$ is a threshold value, with X_0 being locally asymptotically stable if $\mathcal{R}_{vac} < 1$ and unstable if $\mathcal{R}_{vac} > 1$. Linearizing (2.2a)–(2.2c) about an endemic equilibrium gives the Jacobian matrix

$$\frac{\partial f}{\partial x}(S^*, I^*, R^*) = \begin{bmatrix} -d-\beta I^*-\phi-\theta & -\beta S^*-\theta & \nu-\theta \\ (1-\sigma)\beta I^* & -\sigma\beta I^* & -\sigma\beta I^* \\ 0 & \gamma & -(d+\nu) \end{bmatrix}.$$

In the case in which two endemic equilibria exist, $\det\big(\frac{\partial f}{\partial x}(X_*)\big) > 0$ and $\operatorname{tr}\big(\frac{\partial f}{\partial x}(X_*)\big) < 0$. Thus $\frac{\partial f}{\partial x}(X_*)$ has a positive eigenvalue and two eigenvalues with negative real part, making X_* unstable hyperbolic.

Let λ_j, $j = 1, 2, 3$, be the eigenvalues of $\frac{\partial f}{\partial x}(X^*)$ with $\Re(\lambda_1) \le \Re(\lambda_2) \le \Re(\lambda_3)$. It can be shown that $\det\big(\frac{\partial f}{\partial x}(X^*)\big) < 0$ and so $\lambda_1\lambda_2\lambda_3 < 0$. This means that either $\Re(\lambda_j) < 0$ for $j = 1, 2, 3$ or $\Re(\lambda_1) < 0 \le \Re(\lambda_2) \le \Re(\lambda_3)$. Since $\operatorname{tr}\big(\frac{\partial f}{\partial x}(X^*)\big) < 0$, it follows that $\lambda_1+\lambda_2+\lambda_3 < 0$, which implies that $\Re(\lambda_1+\lambda_2) < 0$ and $\Re(\lambda_1+\lambda_3) < 0$.

Assume now that $\theta \le \nu$, and consider the second additive compound [12] of the Jacobian matrix

$$\frac{\partial f}{\partial x}^{[2]}(X^*) = \begin{bmatrix} -\begin{pmatrix}(1+\sigma)\beta I^* \\ +d+\phi+\theta\end{pmatrix} & -\sigma\beta I^* & \theta-\nu \\ \gamma & -\begin{pmatrix}\beta I^*+2d \\ +\phi+\theta+\nu\end{pmatrix} & -\beta S^*-\theta \\ 0 & (1-\sigma)\beta I^* & -\big(\sigma\beta I^*+d+\nu\big) \end{bmatrix}.$$

Using the signs of the matrix elements, it is easily shown that $\det\big(\frac{\partial f}{\partial x}^{[2]}(X^*)\big) < 0$.

The eigenvalues of $\frac{\partial f}{\partial x}^{[2]}(X^*)$ are $\lambda_i + \lambda_j$, $1 \le i < j \le 3$, and so

$$\begin{aligned} -1 &= \operatorname{sgn}\left(\det\left(\frac{\partial f}{\partial x}^{[2]}(X^*)\right)\right) \\ &= \operatorname{sgn}\left(\Re(\lambda_1+\lambda_2)\,\Re(\lambda_1+\lambda_3)\Re(\lambda_2+\lambda_3)\right) \\ &= \operatorname{sgn}\left(\Re(\lambda_2+\lambda_3)\right). \end{aligned}$$

Thus, $\Re(\lambda_j) < 0$ for $j = 1, 2, 3$, and therefore X^* is locally asymptotically stable. ∎

Remark 3.3. For all of the numerical simulations performed here, the parameters satisfy $\theta \le \nu$; i.e., the average period of immunity is no longer than the average vaccine waning time. If $\theta \ge \nu$, then there is no endemic equilibrium for $\mathcal{R}_{vac} \le 1$, since each of the coefficients in $P(I)$ is nonpositive; thus there can be no bistability.

More general techniques are needed to determine the global dynamics for the case $\mathcal{R}_0 > 1$.

4. A geometric approach to global dynamics. In this section, a brief outline of a general mathematical framework for studying global dynamics is given. This approach to global dynamics is developed in the papers of Smith [17] and Li and Muldowney [12, 13, 15]. While this method is usually applied to demonstrate the global stability of a unique equilibrium [11, 14], here it is used to demonstrate bistability for a system that exhibits a backward bifurcation. In [11, 14], compound matrix techniques together with the existence of a compact absorbing set are used to prove global asymptotic stability of the endemic equilibrium point. For cases in which our model exhibits bistability, no such compact absorbing set exists; thus, a sequence of surfaces that exists for time $\epsilon > 0$ and minimizes the functional measuring surface area must be considered.

Let B be the Euclidean ball in $\mathbb{R}^2$, and let $\bar{B}$ and ∂B be its closure and boundary, respectively. Letting $\operatorname{Lip}(X \to Y)$ denote the set of Lipschitzian functions from X to Y, a function $\varphi \in \operatorname{Lip}(\bar{B} \to \mathcal{D})$ is a (simply connected rectifiable) surface in $\mathcal{D}$. A function $\psi \in \operatorname{Lip}(\partial B \to \mathcal{D})$ is a closed rectifiable curve in $\mathcal{D}$ and is called simple if it is one-to-one. Let $\Sigma(\psi, \mathcal{D}) = \{\varphi \in \operatorname{Lip}(\bar{B} \to \mathcal{D}) : \varphi|_{\partial B} = \psi\}$. In [15], it is shown that if ψ is contained in a simply connected open subset of $\mathcal{D}$, then $\Sigma(\psi, \mathcal{D})$ is nonempty.

Let $\|\cdot\|$ be a norm on $\mathbb{R}^{\binom{n}{2}}$. Consider a functional $\mathcal{S}$ on surfaces in $\mathcal{D}$ defined by

$$\mathcal{S}\varphi = \int_{\bar{B}} \left\| P \cdot \left(\frac{\partial \varphi}{\partial u_1} \wedge \frac{\partial \varphi}{\partial u_2} \right) \right\| du, \tag{4.1}$$

where $u = (u_1, u_2)$, $u \mapsto \varphi(u)$ is Lipschitzian on $\bar{B}$, the wedge product $\frac{\partial \varphi}{\partial u_1} \wedge \frac{\partial \varphi}{\partial u_2}$ is a vector in $\mathbb{R}^{\binom{n}{2}}$ (see [19]), and P is an $\binom{n}{2} \times \binom{n}{2}$ matrix such that $\|P^{-1}\|$ is bounded on $\varphi(\bar{B})$. The following result follows from the development in [12] and [15].

PROPOSITION 4.1. *Suppose that ψ is a simple closed rectifiable curve in $\mathbb{R}^n$. Then there exists $\delta > 0$ such that*

$$\mathcal{S}\varphi \ge \delta$$

for all $\varphi \in \Sigma(\psi, \mathbb{R}^n)$.

Functionals of the form (4.1) give a measure of the surface area of the surface φ. In this context, Proposition 4.1 can be interpreted as stating that, given a curve $\psi \subset \mathbb{R}^n$ and a measure of surface area, all surfaces with boundary ψ have surface area uniformly bounded away from zero.

Let $x \mapsto f(x) \in \mathbb{R}^n$ be a C^1 function for x in a set $\mathcal{D} \subset \mathbb{R}^n$. Consider the differential equation

$$\frac{dx}{dt} = f(x). \tag{4.2}$$

(This is used in section 5 with $x = (S, I, R)^T$ for the system (2.2a)–(2.2c).) For any surface φ, the new surface φ_t is defined by $\varphi_t(u) = x(t, \varphi(u))$. Note that when viewed as a function of t, $\varphi_t(u)$ gives the solution to (4.2) that passes through the point $\varphi(u)$ at $t = 0$.

It is shown in [15] that $D_+\mathcal{S}\varphi_t$, the right-hand derivative of $\mathcal{S}\varphi_t$, is given by

$$D_+\mathcal{S}\varphi_t = \int_{\bar{B}} \lim_{h\to 0^+} \frac{1}{h}\Big[\big\|z + hQ\big(\varphi_t(u)\big)z\big\| - \|z\|\Big]\, du, \tag{4.3}$$

where the matrix $Q = P_f P^{-1} + P\frac{\partial f}{\partial x}^{[2]} P^{-1}$. Here P_f is the directional derivative of P in the direction of the vector field f, $\frac{\partial f}{\partial x}^{[2]}$ is the second additive compound [12] of $\frac{\partial f}{\partial x}$, and $z = P \cdot (\frac{\partial \varphi}{\partial u_1} \wedge \frac{\partial \varphi}{\partial u_2})$ is a solution to the differential equation

$$\frac{dz}{dt} = Q\big(\varphi_t(u)\big)z. \tag{4.4}$$

Thus, (4.3) can be rewritten as

$$D_+\mathcal{S}\varphi_t = \int_{\bar{B}} D_+\big\|z\big\|\, du.$$

If there exists $\eta > 0$ such that $D_+\|z\| \le -\eta\|z\|$ for all $z \in \mathbb{R}^{\binom{n}{2}}$ and all $x \in \mathcal{D}$, then $D_+\mathcal{S}\varphi_t \le \int_{\bar{B}} -\eta\|z\|\, du = -\eta\mathcal{S}\varphi_t$, and so $\mathcal{S}\varphi_t \le \mathcal{S}\varphi\, \mathrm{e}^{-\eta t}$ as long as φ_t remains in $\mathcal{D}$. If $\varphi_t \subset \mathcal{D}$ for all t, then $\lim_{t\to\infty} \mathcal{S}\varphi_t = 0$.

Suppose that ψ is the trace of a periodic solution of (4.2). Then ψ is invariant under the flow described by (4.2). Let $\varphi \in \Sigma(\psi, \mathcal{D})$. Then $\varphi_t(\partial B) = x(t, \varphi(\partial B)) = x(t, \psi(\partial B)) = \psi(\partial B)$. Thus, $\varphi_t \in \Sigma(\psi, \mathcal{D})$ as long as $\varphi_t \subset \mathcal{D}$. If $\mathcal{D}$ is positively invariant, then $\varphi_t \in \Sigma(\psi, \mathcal{D})$ for all $t \ge 0$, and therefore, by Proposition 4.1, $\mathcal{S}\varphi_t \ge \delta$ for all $t \ge 0$. Thus, by the remarks of the previous paragraph, the condition that $D_+\|z\| \le -\eta\|z\|$ for all z and x precludes the existence of periodic solutions to (4.2).

In the absence of a compact absorbing set, a surface may not remain in $\mathcal{D}$ for all time. Thus, we consider a sequence of surfaces $\{\varphi^k\}$ in $\Sigma(\psi, \mathcal{D})$ such that $\lim_{k\to\infty} \mathcal{S}\varphi^k = \delta$, where $\delta = \inf\{\mathcal{S}\varphi : \varphi \in \Sigma(\psi, \mathcal{D})\}$ and for which there exists $\epsilon > 0$ such that $\varphi^k_t(\bar{B}) \subset \mathcal{D}$ for $t \in [0, \epsilon]$ and $k = 1, 2, \ldots$. If $D_+\|z\| \le -\eta\|z\|$ for all $z \in \mathbb{R}^{\binom{n}{2}}$ and all $x \in \mathcal{D}$, then $\mathcal{S}\varphi^k_\epsilon \le \mathcal{S}\varphi^k \mathrm{e}^{-\eta\epsilon}$, and therefore there exists l such that $\mathcal{S}\varphi^l_\epsilon < \delta$. This implies that the boundary of φ^l_ϵ is not ψ, and therefore ψ is not invariant under (4.2). Thus, if for every simple closed curve ψ in $\mathcal{D}$ there is a sequence of surfaces $\{\varphi^k\}$ in $\Sigma(\psi, \mathcal{D})$ that all remain in $\mathcal{D}$ for some time $\epsilon > 0$, and there is a surface functional $\mathcal{S}$ of the form given in (4.1), then the condition $D_+\|z\| \le -\eta\|z\|$ precludes the existence of invariant closed curves, including periodic orbits, homoclinic orbits, and heteroclinic cycles.

The above conditions are robust under local C^1 perturbations to the original differential equation (4.2). Thus, if (4.2) satisfies the above hypotheses, then so do all systems that are sufficiently C^1-close to (4.2). Therefore, Pugh's closing lemma [7] leads to the following result in the spirit of Criterion 3.1 in [15], giving conditions that preclude the existence of nonconstant nonwandering points.

THEOREM 4.2. *Suppose there exists a norm $\|\cdot\|$ on $\mathbb{R}^{\binom{n}{2}}$ and $\eta > 0$ such that $D_+\|z\| \le -\eta\|z\|$ for all $z \in \mathbb{R}^{\binom{n}{2}}$ satisfying (4.4) and all $x \in \mathcal{D}$ for $\mathcal{D}$ simply connected. Further, suppose that for any simple closed curve ψ in $\mathcal{D}$ there exists a sequence of surfaces $\{\varphi^k\}$ that minimizes $\mathcal{S}$ relative to $\Sigma(\psi, \mathcal{D})$ and there exists $\epsilon > 0$ such that $\varphi_t^k \subset \mathcal{D}$ for $t \in [0, \epsilon]$ and $k = 1, 2, \ldots$. Then any omega limit point of (4.2) in the interior of $\mathcal{D}$ is an equilibrium.*

In order to apply the theorem to a particular system, it is necessary to find a norm $\|\cdot\|$ and a matrix P (which then determines the matrix Q) such that $D_+\|z\| \le -\eta\|z\|$ and to show that an appropriate sequence of surfaces exists.

5. Global analysis of the *SIRS* model with vaccination. Recalling that for $\mathcal{R}_0 < 1$ the DFE is globally asymptotically stable (Theorem 3.2), we now apply the theory outlined in the previous section to system (2.2a)–(2.2c) for $\mathcal{R}_0 > 1$. Let $\mathcal{D} = \{(S, I, R) : S, R \ge 0, I > 0, S + I + R \le 1\}$. The Jacobian matrix at a general point $x = (S, I, R)^T$ is given by

$$\frac{\partial f}{\partial x} = \begin{bmatrix} -d - \beta I - \phi - \theta & -\beta S - \theta & \nu - \theta \\ (1-\sigma)\beta I & \beta(S + \sigma V - \sigma I) - (d + \gamma) & -\sigma\beta I \\ 0 & \gamma & -(d + \nu) \end{bmatrix}, \tag{5.1}$$

where $V = 1 - S - I - R$ from (2.2d). The second additive compound [12] of the Jacobian matrix is the 3×3 matrix given by

$$\frac{\partial f}{\partial x}^{[2]} = \begin{bmatrix} \begin{pmatrix} \beta(S + \sigma V - (1+\sigma)I) \\ -[2d + \phi + \theta + \gamma] \end{pmatrix} & -\sigma\beta I & \theta - \nu \\ \gamma & -(\beta I + 2d + \phi + \theta + \nu) & -(\beta S + \theta) \\ 0 & (1-\sigma)\beta I & \begin{pmatrix} \beta(S + \sigma V - \sigma I) \\ -[2d + \gamma + \nu] \end{pmatrix} \end{bmatrix}.$$

Let $P = \frac{1}{I} I_3$, where I_3 is the 3×3 identity matrix. Then $P_f P^{-1} = -\frac{1}{I}\frac{dI}{dt} I_3$ with $\frac{dI}{dt}$ given by (2.2b), and

$$\begin{aligned} Q &= P_f P^{-1} + P \frac{\partial f}{\partial x}^{[2]} P^{-1} \\ &= \begin{bmatrix} -[(1+\sigma)\beta I + d + \phi + \theta] & -\sigma\beta I & \theta - \nu \\ \gamma & \begin{pmatrix} \gamma - [\beta(S + \sigma V + I) \\ + d + \phi + \theta + \nu] \end{pmatrix} & -(\beta S + \theta) \\ 0 & (1-\sigma)\beta I & -(\sigma\beta I + d + \nu) \end{bmatrix}. \end{aligned} \tag{5.2}$$

For $z = (z_1, z_2, z_3)^T$, let $\|z\|$ be given by

$$\|z\| = \begin{cases} \max\{|z_1| + |z_3|, |z_2| + |z_3|\} & \text{if} \quad 0 \le z_2 z_3, \\ \max\{|z_1| + |z_3|, |z_2|\} & \text{if} \quad z_2 z_3 \le 0. \end{cases} \tag{5.3}$$

This norm is used as a Lyapunov function for system (4.4). The following two propositions, with proofs given in the appendices, lead to our main result.

PROPOSITION 5.1. *Assume that in system* (2.2a)–(2.2c) *the parameters satisfy the following inequalities:*

$$\begin{aligned} \theta &< d + 2\nu, \\ 2\gamma &< d + \phi + \theta + \nu, \\ \gamma &< d + \phi + \nu. \end{aligned} \tag{5.4}$$

Then there exists $\eta > 0$ *such that* $D_+\|z\| \le -\eta\|z\|$ *for all* $z \in \mathbb{R}^3$ *and all* $S, I, R, V \ge 0$, $I \ne 0$, *where* z *is a solution of* (4.4) *with* Q *and* $\|\cdot\|$ *given by* (5.2) *and* (5.3), *respectively.*

Note that inequalities (5.4) are independent of the transmission coefficient β, the proportion of newborns vaccinated α, and the vaccine efficacy σ but depend on the other model parameters.

PROPOSITION 5.2. *Let* ψ *be a simple closed curve in* $\mathcal{D}$. *There exist* $\epsilon > 0$ *and a sequence of surfaces* $\{\varphi^k\}$ *that minimizes* $\mathcal{S}$ *given by* (4.1) *relative to* $\Sigma(\psi, \mathcal{D})$ *such that* $\varphi_t^k \subset \mathcal{D}$ *for all* $k = 1, 2, \ldots$ *and all* $t \in [0, \epsilon]$.

THEOREM 5.3. *If inequalities* (5.4) *hold, then each positive semitrajectory of* (2.2a)–(2.2c) *in* $\bar{\mathcal{D}}$ *limits to a single equilibrium.*

Proof. Let Γ be a positive semitrajectory in $\bar{\mathcal{D}}$ with omega limit set Ω. Suppose that Ω intersects the interior of $\mathcal{D}$. Propositions 5.1 and 5.2 ensure that Theorem 4.2 can be applied to system (2.2a)–(2.2c). Theorem 4.2 implies that every omega limit point of (2.2a)–(2.2c) in the interior of $\mathcal{D}$ is an equilibrium. Since the system has a finite number of equilibria, there are only a finite number of points in the interior of $\mathcal{D}$ which can be in Ω. As Γ is bounded, Ω must be connected. Thus, Ω must consist of a single equilibrium.

Suppose, on the other hand, that Ω is contained in the boundary $\partial\mathcal{D}$ of $\mathcal{D}$. Since omega limit sets are invariant, Ω must be contained in the largest invariant subset of $\partial\mathcal{D}$. By considering (2.2a)–(2.2c) with the assumption that θ, γ, and $\alpha d + \phi$ are positive, it is easily shown that $\{X_0\}$ is the only invariant subset of $\partial\mathcal{D}$ and therefore $\Omega = \{X_0\}$. □

For the parameters given in Table 3.1 and used in Figure 3.2, inequalities (5.4) are satisfied for $\phi > 0.063$, which contains part of the bistability region. With $\theta = 1/(1$ year) and $\nu = 1/(14$ days), and all other parameters fixed as in Table 3.1, Figure 5.1 shows a two-dimensional bifurcation diagram in the (σ, ϕ)-plane. Inequalities (5.4) are satisfied for $\phi > 0.021$, including the entire region for which bistability occurs.

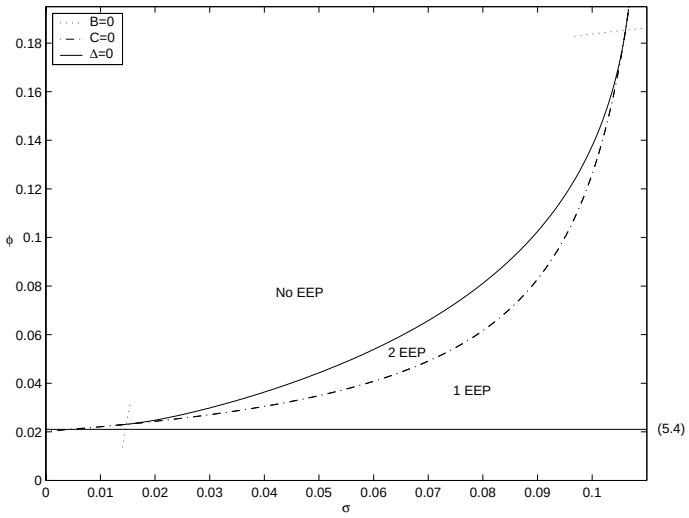


FIG. 5.1. *Bifurcation diagram in the* (σ, ϕ)*-plane, for* $1/\theta$ *equal to one year and* $1/\nu$ *equal to two weeks. The second inequality of* (5.4) *holds for* ϕ *above the horizontal line labeled* (5.4). *The other two inequalities of* (5.4) *hold in the whole region.*

COROLLARY 5.4. *Assume that inequalities* (5.4) *hold for system* (2.2a)–(2.2c).

(i) *If there is no endemic equilibrium, then all solutions limit to the DFE* X_0.

(ii) *If* $\mathcal{R}_{vac} > 1$, *then* (S, I, R) *tends to the unique endemic equilibrium* X^*, *provided that* $I(0) > 0$.

(iii) *If there are two endemic equilibria, or if* X_c *is the only endemic equilibrium, then depending on the initial values, the disease dies out or limits to a constant endemic value.*

Proof. Statements (i) and (iii) follow directly from Theorem 5.3. For (ii), $\mathcal{R}_{vac} > 1$, and so the DFE is unstable with a two-dimensional stable manifold (see (5.1)), which is $\{I = 0\}$. Since from Theorem 5.3 every solution limits to an equilibrium, and since solutions with $I(0) > 0$ do not go to the DFE, they must limit to X^*. □

COROLLARY 5.5. *Suppose that* $\theta \leq \nu$ *and* $2\gamma < d+\phi+\theta+\nu$. *Then the conclusions of Corollary* 5.4 *hold. Furthermore, if system* (2.2a)–(2.2c) *has two endemic equilibria, then the basins of attraction of* X^* *and* X_0 *have positive measure and the basin of attraction of* X_* *has zero measure.*

Proof. If $\theta \leq \nu$ and $2\gamma < d + \phi + \theta + \nu$, then the inequalities (5.4) are satisfied and so the conclusions of Corollary 5.4 follow. From Theorem 3.2, X^* is locally asymptotically stable. Thus it has a basin of attraction with positive measure. It is shown in section 3 that X_* is unstable hyperbolic. Thus, the set of points which limit to X_* has Lebesgue measure zero. When there are two endemic equilibria, $\mathcal{R}_{vac} < 1$, and thus X_0 is locally asymptotically stable and has a basin of attraction with positive measure. □

Remark 5.6. For ν sufficiently large (i.e., a sufficiently short period of immunity), inequalities (5.4) hold and the conclusion of Corollary 5.5 holds. In the limiting case as ν tends to infinity, upon recovery infective individuals progress directly to the susceptible class; thus the model reduces to an SIS model with vaccination, similar to that considered in [10].

Remark 5.7. By using norms other than the norm given by (5.3), inequalities (5.4) can be replaced with other conditions which lead to the same conclusions. For example, if $\|z\|$ is given by

$$\|z\| = \max_{j=1,2,3} |z_j|,$$

then (5.4) can be replaced with

$$2\gamma \leq d + \nu + \theta. \tag{5.5}$$

Similarly, if $\|z\|$ is given by

$$\|z\| = \begin{cases} |z_1| + |z_2| + |z_3| & \text{if} \quad \operatorname{sgn}(z_1) = \operatorname{sgn}(z_2) = \operatorname{sgn}(z_3), \\ \max\{|z_1| + |z_2|, |z_1| + |z_3|\} & \text{if} \quad \operatorname{sgn}(z_1) = \operatorname{sgn}(z_2) = -\operatorname{sgn}(z_3), \\ \max\{|z_1| + |z_3|, |z_2|\} & \text{if} \quad \operatorname{sgn}(z_1) = -\operatorname{sgn}(z_2) = \operatorname{sgn}(z_3), \\ \max\{|z_1| + |z_3|, |z_2| + |z_3|\} & \text{if} \quad -\operatorname{sgn}(z_1) = \operatorname{sgn}(z_2) = \operatorname{sgn}(z_3), \end{cases}$$

then (5.4) can be replaced with

$$\begin{aligned} &\theta < d + 2\nu, \\ &\gamma < d + \phi + \min\{\theta, \nu\}. \end{aligned} \tag{5.6}$$

For parameter values given in Table 3.1, the inequalities found in (5.4) are less restrictive than those given in (5.5) or (5.6).

6. Concluding remarks. The model formulated in section 2 incorporates vaccination for a disease in a simple manner, with vaccinated individuals in a class distinct from that of the individuals who have recovered from the disease. By contrast, in some models (e.g., [3]) these two classes are combined. The basic reproduction number as modified by vaccination, namely $\mathcal{R}_{vac}$ as given by (3.2), is a key parameter in our model. To eradicate the disease, it may not be sufficient to reduce $\mathcal{R}_{vac}$ below one. In the case of bistability, $\mathcal{R}_{vac}$ must be further reduced; see Figure 3.1(d). Increasing vaccination of either newborns or the population at risk has the effect of reducing $\mathcal{R}_{vac}$. An important parameter in $\mathcal{R}_{vac}$ is the efficacy of the vaccine, namely, $1-\sigma$. Bistability may occur for a range of σ values. This range depends on the values of the other parameters in the model. The occurrence of a backward bifurcation is illustrated for some parameter values in Figures 3.1(b), 3.1(d), 3.2, 5.1. In the case of bistability, the asymptotic behavior of the proportion of infectives depends on the initial conditions. In such a situation, global analysis is more complicated than in a situation with a unique endemic equilibrium and a compact absorbing set. An appropriate sequence of surfaces that minimizes the functional measuring surface area must be considered. This novel approach is outlined in section 4 and then applied to our model in section 5. Global results are proved under mild parameter restrictions, and it is indicated that alternative restrictions arise from alternative choices of norms. The rate of vaccination of susceptibles and the vaccine waning rate play a role in these restrictions, whereas the vaccine efficacy and the proportion of newborns vaccinated do not. Theorem 5.3 rules out any complicated behavior (e.g., limit cycles) under mild parameter restrictions, and numerical simulations have found no such behavior for any parameter values. A more realistic model including vaccination should incorporate age structure and demographics (see, e.g., [5]). However, global results are then not available, and simulations must be performed to gain some insight into the model behavior and to determine vaccination strategy.

The analysis of our model can be regarded as the first application of using a minimizing sequence of surfaces in this context. This method may also be useful in other models for which there exist solutions that limit to boundary equilibria.

Appendix A. Proof of Proposition 5.1.

Proof. We demonstrate the existence of some $\eta > 0$ such that $D_+\|z\| \le -\eta\|z\|$, where z is a solution of (4.4). By linearity, if this inequality is true for some z, then it is also true for $-z$. The proof is subdivided into eight cases based on the octant and the definition of the norm in (5.3).

Case 1. If $0 < z_1, z_2, z_3$ and $|z_1| + |z_3| > |z_2| + |z_3|$, then $\|z\| = |z_1| + |z_3|$ and

$$\begin{aligned} D_+\|z\| &= D_+(|z_1| + |z_3|) \\ &= D_+(z_1 + z_3) \\ &= \frac{dz_1}{dt} + \frac{dz_3}{dt} \\ &= -\big((1+\sigma)\beta I + d + \phi + \theta\big)z_1 + (1-2\sigma)\beta I z_2 + \big(\theta - (\sigma\beta I + d + 2\nu)\big)z_3. \end{aligned}$$

Noting that $(1-2\sigma)\beta I z_2 \le (1-\sigma)\beta I z_2 \le (1-\sigma)\beta I z_1$,

$$\begin{aligned} D_+\|z\| &< -(2\sigma\beta I + d + \phi + \theta)z_1 + (\theta - (\sigma\beta I + d + 2\nu))z_3 \\ &\le \max\{-(2\sigma\beta I + d + \phi + \theta),\ \theta - (\sigma\beta I + d + 2\nu)\}\,\|z\|. \end{aligned}$$

Thus, in order that $D_+\|z\|$ be bounded away from zero on the negative side for all z

and all $I > 0$, we require that

$$\theta < d + 2\nu. \tag{A.1}$$

Case 2. If $0 < z_1, z_2, z_3$ and $|z_1| + |z_3| < |z_2| + |z_3|$, then $\|z\| = |z_2| + |z_3|$ and

$$\begin{aligned} D_+\|z\| &= D_+(|z_2| + |z_3|) \\ &= \frac{dz_2}{dt} + \frac{dz_3}{dt} \\ &= \gamma z_1 + \big(\gamma - \big(\beta(S + \sigma V + \sigma I) + d + \phi + \theta + \nu\big)\big) z_2 - \big(\beta(S + \sigma I) + d + \theta + \nu\big) z_3. \end{aligned}$$

Since $z_1 < z_2$, this becomes

$$D_+\|z\| < \big(2\gamma - \big(\beta(S + \sigma V + \sigma I) + d + \phi + \theta + \nu\big)\big) z_2 - \big(\beta(S + \sigma I) + d + \theta + \nu\big) z_3.$$

In order that $D_+\|z\|$ be bounded away from zero on the negative side for all $S, I, V > 0$, we require that

$$2\gamma < d + \phi + \theta + \nu. \tag{A.2}$$

Case 3. If $z_1 < 0 < z_2, z_3$ and $|z_1| + |z_3| > |z_2| + |z_3|$, then $\|z\| = |z_1| + |z_3|$ and

$$\begin{aligned} D_+\|z\| &= D_+(|z_1| + |z_3|) \\ &= D_+(-z_1 + z_3) \\ &= -\frac{dz_1}{dt} + \frac{dz_3}{dt} \\ &= \big((1 + \sigma)\beta I + d + \phi + \theta\big) z_1 + \beta I z_2 - (\sigma\beta I + d + \theta) z_3 \\ &= -\big((1 + \sigma)\beta I + d + \phi + \theta\big)|z_1| + \beta I |z_2| - (\sigma\beta I + d + \theta)|z_3|. \end{aligned}$$

Since $|z_2| < |z_1|$,

$$\begin{aligned} D_+\|z\| &< -(\sigma\beta I + d + \phi + \theta)|z_1| - (\sigma\beta I + d + \theta)|z_3| \\ &\le -(d + \theta)\|z\|. \end{aligned}$$

Thus, in this case, $D_+\|z\|$ is automatically bounded away from zero on the negative side.

Case 4. If $z_1 < 0 < z_2, z_3$ and $|z_1| + |z_3| < |z_2| + |z_3|$, then $\|z\| = |z_2| + |z_3|$ and

$$\begin{aligned} D_+\|z\| &= D_+(|z_2| + |z_3|) \\ &= \frac{dz_2}{dt} + \frac{dz_3}{dt} \\ &= \gamma z_1 + \big(\gamma - \big(\beta(S + \sigma V + \sigma I) + d + \phi + \theta + \nu\big)\big) z_2 - \big(\beta(S + \sigma I) + d + \theta + \nu\big) z_3 \\ &= -\gamma|z_1| + \big(\gamma - \big(\beta(S + \sigma V + \sigma I) + d + \phi + \theta + \nu\big)\big)|z_2| \\ &\qquad - \big(\beta(S + \sigma I) + d + \theta + \nu\big)|z_3| \\ &< \big(\gamma - \big(\beta(S + \sigma V + \sigma I) + d + \phi + \theta + \nu\big)\big)|z_2| - \big(\beta(S + \sigma I) + d + \theta + \nu\big)|z_3| \\ &\le \max\big\{\gamma - \big(\beta(S + \sigma V + \sigma I) + d + \phi + \theta + \nu\big), -\big(\beta(S + \sigma I) + d + \theta + \nu\big)\big\}\,\|z\|. \end{aligned}$$

Thus we require that

$$\gamma < d + \phi + \theta + \nu. \tag{A.3}$$

Case 5. If $z_2 < 0 < z_1, z_3$ and $|z_1| + |z_3| > |z_2|$, then $\|z\| = |z_1| + |z_3|$ and

$$\begin{aligned} D_+\|z\| &= D_+(|z_1| + |z_3|) \\ &= \frac{dz_1}{dt} + \frac{dz_3}{dt} \\ &= -\big((1+\sigma)\beta I + d + \phi + \theta\big)z_1 + (1-2\sigma)\beta I z_2 + \big(\theta - (\sigma\beta I + d + 2\nu)\big)z_3 \\ &= -\big((1+\sigma)\beta I + d + \phi + \theta\big)|z_1| + (2\sigma-1)\beta I|z_2| + \big(\theta - (\sigma\beta I + d + 2\nu)\big)|z_3|. \end{aligned}$$

Since $(2\sigma - 1)\beta I|z_2| \le \sigma\beta I|z_2| < \sigma\beta I(|z_1| + |z_3|)$,

$$\begin{aligned} D_+\|z\| &< -(\beta I + d + \phi + \theta)|z_1| + \big(\theta - (d + 2\nu)\big)|z_3| \\ &\le \max\{-(\beta I + d + \phi + \theta),\ \theta - (d + 2\nu)\}\|z\|. \end{aligned}$$

Thus, we require that (A.1) hold.

Case 6. If $z_2 < 0 < z_1, z_3$ and $|z_1| + |z_3| < |z_2|$, then $\|z\| = |z_2|$ and

$$\begin{aligned} D_+\|z\| &= D_+(|z_2|) \\ &= -\frac{dz_2}{dt} \\ &= -\gamma z_1 - \big(\gamma - \big(\beta(S + \sigma V + I) + d + \phi + \theta + \nu\big)\big)z_2 + (\beta S + \theta)z_3 \\ &= -\gamma|z_1| + \big(\gamma - \big(\beta(S + \sigma V + I) + d + \phi + \theta + \nu\big)\big)|z_2| + (\beta S + \theta)|z_3|. \end{aligned}$$

Since $-\gamma|z_1| < 0$ and $|z_3| < |z_2|$,

$$D_+\|z\| < \big(\gamma - \big(\beta(\sigma V + I) + d + \phi + \nu\big)\big)|z_2|.$$

Thus, we require that

$$\gamma < d + \phi + \nu. \tag{A.4}$$

Case 7. If $z_3 < 0 < z_1, z_2$ and $|z_1| + |z_3| > |z_2|$, then $\|z\| = |z_1| + |z_3|$ and

$$\begin{aligned} D_+\|z\| &= D_+(|z_1| + |z_3|) \\ &= \frac{dz_1}{dt} - \frac{dz_3}{dt} \\ &= -\big((1+\sigma)\beta I + d + \phi + \theta\big)z_1 - \beta I z_2 + (\theta + \sigma\beta I + d)z_3 \\ &= -\big((1+\sigma)\beta I + d + \phi + \theta\big)|z_1| - \beta I|z_2| - (\theta + \sigma\beta I + d)|z_3| \\ &\le -(d + \theta)\|z\|. \end{aligned}$$

Thus, in this case, $D_+\|z\|$ is automatically bounded away from zero on the negative side.

Case 8. If $z_3 < 0 < z_1, z_2$ and $|z_1| + |z_3| < |z_2|$, then $\|z\| = |z_2|$ and

$$\begin{aligned} D_+\|z\| &= D_+(|z_2|) \\ &= \frac{dz_2}{dt} \\ &= \gamma z_1 + \big(\gamma - \big(\beta(S + \sigma V + I) + d + \phi + \theta + \nu\big)\big)z_2 - (\beta S + \theta)z_3 \\ &= \gamma|z_1| + \big(\gamma - \big(\beta(S + \sigma V + I) + d + \phi + \theta + \nu\big)\big)|z_2| + (\beta S + \theta)|z_3|. \end{aligned}$$

Noting that $|z_1| + |z_3| < |z_2|$,

$$D_+\|z\| < \big(\gamma - \big(\beta(S + \sigma V + I) + d + \phi + \theta + \nu\big) + \max\{\gamma, \beta S + \theta\}\big)\|z\|.$$

Thus, we require that (A.2) and (A.4) hold.

Note that (A.2) and (A.4) each imply (A.3). Thus, if inequalities (A.1), (A.2), and (A.4) hold, then there exists $\eta > 0$ such that $D_+\|z\| \le -\eta\|z\|$ for almost every $z \in \mathbb{R}^3$ and all nonnegative S, I, R, and V. The boundary between the different cases, including, for example, $z_j = 0$ for some j, is resolved by continuity. Thus, (A.1), (A.2), and (A.4) (equivalently (5.4)) imply that $D_+\|z\| \le -\eta\|z\|$ for all $z \in \mathbb{R}^3$. □

Appendix B. Proof of Proposition 5.2.

Proof. Let $\xi = \frac{1}{2}\min\{I : (S, I, R) \in \psi\}$ and let $\epsilon > 0$. Note that the model is well posed so solutions remain in the nonnegative orthant. In $\mathcal{D}$, $\frac{dI}{dt} \ge -(d + \nu)I$. Thus, if a solution satisfies $I(0) \ge \xi$, then the solution remains in $\mathcal{D}$ for time ϵ.

Therefore, it suffices to show that there exists a sequence of surfaces $\{\varphi^k\}$ that minimizes $\mathcal{S}$ relative to $\Sigma(\psi, \tilde{\mathcal{D}})$, where $\tilde{\mathcal{D}} = \{(S, I, R) \in \mathcal{D} : I \ge \xi\}$. Let $\varphi = (S(u), I(u), R(u)) \in \Sigma(\psi, \mathcal{D})$. Define a new surface $\tilde{\varphi} = (\tilde{S}, \tilde{I}, \tilde{R})$ by

(B.1)

$$\tilde{\varphi}(u) = \begin{cases} \varphi(u) & \text{if } I(u) \ge \xi, \\ (S, \xi, R) & \text{if } I(u) < \xi \text{ and } S(u) + \xi + R(u) \le 1, \\ \big(\frac{S}{S+R}(1-\xi),\ \xi,\ \frac{R}{S+R}(1-\xi)\big) & \text{if } I(u) < \xi \text{ and } S(u) + \xi + R(u) > 1. \end{cases}$$

Note that $\tilde{\varphi} \in \Sigma(\psi, \tilde{\mathcal{D}})$. We now demonstrate that $\mathcal{S}\tilde{\varphi} \le \mathcal{S}\varphi$.

Since φ is Lipschitzian, the partial derivatives $\frac{\partial \varphi}{\partial u_j}$, $j = 1, 2$, exist almost everywhere. Thus,

$$\frac{\partial \varphi}{\partial u_1} \wedge \frac{\partial \varphi}{\partial u_2} = \begin{bmatrix} \frac{\partial S}{\partial u_1} \\ \frac{\partial I}{\partial u_1} \\ \frac{\partial R}{\partial u_1} \end{bmatrix} \wedge \begin{bmatrix} \frac{\partial S}{\partial u_2} \\ \frac{\partial I}{\partial u_2} \\ \frac{\partial R}{\partial u_2} \end{bmatrix} = \begin{bmatrix} \det \begin{pmatrix} \frac{\partial S}{\partial u_1} & \frac{\partial S}{\partial u_2} \\ \frac{\partial I}{\partial u_1} & \frac{\partial I}{\partial u_2} \end{pmatrix} \\ \det \begin{pmatrix} \frac{\partial S}{\partial u_1} & \frac{\partial S}{\partial u_2} \\ \frac{\partial R}{\partial u_1} & \frac{\partial R}{\partial u_2} \end{pmatrix} \\ \det \begin{pmatrix} \frac{\partial I}{\partial u_1} & \frac{\partial I}{\partial u_2} \\ \frac{\partial R}{\partial u_1} & \frac{\partial R}{\partial u_2} \end{pmatrix} \end{bmatrix}$$

is a vector in $\mathbb{R}^3$ for almost every $u \in B$. To examine $\big\|\frac{\partial \tilde{\varphi}}{\partial u_1} \wedge \frac{\partial \tilde{\varphi}}{\partial u_2}\big\|$, we do a case analysis based on the definition of $\tilde{\varphi}$ given in (B.1).

Case 1. If $I(u) \ge \xi$, then $\tilde{\varphi} = \varphi$ and therefore $\big\|\frac{\partial \tilde{\varphi}}{\partial u_1} \wedge \frac{\partial \tilde{\varphi}}{\partial u_2}\big\| = \big\|\frac{\partial \varphi}{\partial u_1} \wedge \frac{\partial \varphi}{\partial u_2}\big\|$ when all of the relevant partial derivatives exist.

Case 2. If $I(u) < \xi$ and $S(u) + \xi + R(u) \le 1$, then $\tilde{\varphi}(v) = (S(v), \xi, R(v))$. Therefore

$$\frac{\partial \tilde{\varphi}}{\partial u_1} \wedge \frac{\partial \tilde{\varphi}}{\partial u_2} = \begin{bmatrix} \det \begin{pmatrix} \frac{\partial S}{\partial u_1} & \frac{\partial S}{\partial u_2} \\ 0 & 0 \end{pmatrix} \\ \det \begin{pmatrix} \frac{\partial S}{\partial u_1} & \frac{\partial S}{\partial u_2} \\ \frac{\partial R}{\partial u_1} & \frac{\partial R}{\partial u_2} \end{pmatrix} \\ \det \begin{pmatrix} 0 & 0 \\ \frac{\partial R}{\partial u_1} & \frac{\partial R}{\partial u_2} \end{pmatrix} \end{bmatrix} = \begin{bmatrix} 0 \\ \det \begin{pmatrix} \frac{\partial S}{\partial u_1} & \frac{\partial S}{\partial u_2} \\ \frac{\partial R}{\partial u_1} & \frac{\partial R}{\partial u_2} \end{pmatrix} \\ 0 \end{bmatrix}$$

almost everywhere. If y_j is equal to either z_j or zero for $j=1,2,3$, then $\|(y_1,y_2,y_3)^T\| \le \|(z_1,z_2,z_3)^T\|$ for the given norm. It follows that $\left\|\frac{\partial\tilde{\varphi}}{\partial u_1}\wedge\frac{\partial\tilde{\varphi}}{\partial u_2}\right\| \le \left\|\frac{\partial\varphi}{\partial u_1}\wedge\frac{\partial\varphi}{\partial u_2}\right\|$.

Case 3. If $I(u)<\xi$ and $S(u)+\xi+R(u)>1$, then $\tilde{\varphi}(v)=\left(\frac{S}{S+R}(1-\xi),\ \xi,\ \frac{R}{S+R}(1-\xi)\right)$. Therefore

$$\frac{\partial\tilde{\varphi}}{\partial u_j}=(1-\xi)\frac{R\frac{\partial S}{\partial u_j}-S\frac{\partial R}{\partial u_j}}{(S+R)^2}\begin{bmatrix}1\\0\\-1\end{bmatrix}$$

for $j=1,2$. Thus, $\frac{\partial\tilde{\varphi}}{\partial u_1}$ and $\frac{\partial\tilde{\varphi}}{\partial u_2}$ are linearly dependent, and so their wedge product is zero [19]. Therefore $\left\|\frac{\partial\tilde{\varphi}}{\partial u_1}\wedge\frac{\partial\tilde{\varphi}}{\partial u_2}\right\|=0\le\left\|\frac{\partial\varphi}{\partial u_1}\wedge\frac{\partial\varphi}{\partial u_2}\right\|$.

The above three cases show that $\left\|\frac{\partial\tilde{\varphi}}{\partial u_1}\wedge\frac{\partial\tilde{\varphi}}{\partial u_2}\right\| \le \left\|\frac{\partial\varphi}{\partial u_1}\wedge\frac{\partial\varphi}{\partial u_2}\right\|$ for almost all $u\in\bar{B}$. We also note that $\tilde{I}(u)=\max\{I(u),\xi\}$ and thus $1/\tilde{I}\le 1/I$. Therefore from (4.1),

$$\begin{aligned}\mathcal{S}\tilde{\varphi}&=\int_{\bar{B}}\frac{1}{\tilde{I}}\left\|\frac{\partial\tilde{\varphi}}{\partial u_1}\wedge\frac{\partial\tilde{\varphi}}{\partial u_2}\right\|du\\&\le\int_{\bar{B}}\frac{1}{I}\left\|\frac{\partial\varphi}{\partial u_1}\wedge\frac{\partial\varphi}{\partial u_2}\right\|du\\&=\mathcal{S}\varphi.\end{aligned}$$

Let $\{\varphi^k\}$ be a sequence of surfaces that minimizes $\mathcal{S}$ relative to $\Sigma(\psi,\mathcal{D})$. Let $\{\tilde{\varphi}^k\}$ be a sequence of surfaces in $\Sigma(\psi,\tilde{\mathcal{D}})$ defined by the above construction. Since $\mathcal{S}\tilde{\varphi}^k\le\mathcal{S}\varphi^k$ for each k, and $\Sigma(\psi,\tilde{\mathcal{D}})$ is a subset of $\Sigma(\psi,\mathcal{D})$, it follows that $\{\tilde{\varphi}^k\}$ minimizes $\mathcal{S}$ relative to $\Sigma(\psi,\tilde{\mathcal{D}})$. □

REFERENCES

[1] J. Chin, ed., *Control of Communicable Diseases Manual*, 17th ed., American Public Health Association, Washington, DC, 2000.

[2] J. Dushoff, W. Huang, and C. Castillo-Chavez, *Backwards bifurcations and catastrophe in simple models of fatal diseases*, J. Math. Biol., 36 (1998), pp. 227–248.

[3] D. J. D. Earn, P. Rohani, B. M. Bolker, and B. T. Grenfell, *A simple model for complex dynamical transitions in epidemics*, Science, 287 (2000), pp. 667–670.

[4] D. Greenhalgh, O. Diekmann, and M. C. M. de Jong, *Subcritical endemic steady states in mathematical models for animal infections with incomplete immunity*, Math. Biosci., 165 (2000), pp. 1–25.

[5] H. W. Hethcote, *Oscillations in an endemic model for pertussis*, Canadian Appl. Math. Quart., 6 (1998), pp. 61–88.

[6] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., 42 (2000), pp. 599–653.

[7] M. W. Hirsch, *Systems of differential equations that are competitive or cooperative.* VI: *A local C^r closing lemma for* 3*-dimensional systems*, Ergodic Theory Dynam. Systems, 11 (1991), pp. 443–454.

[8] W. Huang, K. L. Cooke, and C. Castillo-Chavez, *Stability and bifurcation for a multiple-group model for the dynamics of HIV/AIDS transmission*, SIAM J. Appl. Math., 52 (1992), pp. 835–854.

[9] C. Kribs-Zaleta and M. Martcheva, *Vaccination strategies and backward bifurcation in an age-since-infection structured model*, Math. Biosci., 177/178 (2002), pp. 317–332.

[10] C. Kribs-Zaleta and J. Velasco-Hernández, *A simple vaccination model with multiple endemic states*, Math. Biosci., 164 (2000), pp. 183–201.

[11] M. Y. Li and J. S. Muldowney, *Global stability for the SEIR model in epidemiology*, Math. Biosci., 125 (1995), pp. 155–164.

[12] M. Y. Li and J. S. Muldowney, *On R. A. Smith's autonomous convergence theorem*, Rocky Mountain J. Math., 25 (1995), pp. 365–379.

[13] M. Y. Li and J. S. Muldowney, *A geometric approach to global-stability problems*, SIAM J. Math. Anal., 27 (1996), pp. 1070–1083.

[14] M. Y. Li, H. L. Smith, and L. Wang, *Global dynamics of an SEIR epidemic model with vertical transmission*, SIAM J. Appl. Math., 62 (2001), pp. 58–69.
[15] Y. Li and J. S. Muldowney, *On Bendixson's criterion*, J. Differential Equations, 106 (1993), pp. 27–39.
[16] C. C. McCluskey, *A model of HIV/AIDS with staged progression and amelioration*, Math. Biosci., 181 (2003), pp. 1–16.
[17] R. A. Smith, *Some applications of Hausdorff dimension inequalities for ordinary differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 104 (1986), pp. 235–259.
[18] B. Song, C. Castillo-Chavez, and J. P. Aparicio, *Global dynamics of tuberculosis models with density dependent demography*, in Mathematical Approaches for Emerging and Reemerging Infectious Diseases: Models, Methods, and Theory, C. Castillo-Chavez, with S. Blower, P. van den Driessche, D. Kirschner, and A.-A. Yakubu, eds., IMA Vol. Math. Appl. 126, Springer-Verlag, New York, 2002, pp. 275–294.
[19] M. Spivak, *Calculus on Manifolds*, W. A. Benjamin, New York, 1965.
[20] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180 (2002), pp. 29–48.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 277–296

ANALYSIS OF A VARIATIONAL APPROACH TO PROGRESSIVE LENS DESIGN*

JING WANG†, ROBERT GULLIVER‡, AND FADIL SANTOSA‡

Abstract. We consider a variational approach to the progressive lens design problem. The corresponding Euler–Lagrange equation is a fourth-order nonlinear elliptic partial differential equation. We analyze two linearizations of the equation and show the existence and uniqueness as well as the regularity of the solutions for various boundary conditions. We end with an example of a progressive lens designed by solving the elliptic partial differential equation.

Key words. progressive lens, fourth-order boundary value problems, well-posedness

AMS subject classifications. 35J35, 35J40

DOI. 10.1137/S0036139902408941

1. Introduction. *Presbyopia* is very common in people over age forty, and is caused by the loss of the accommodative power of the human eye. Single-vision reading glasses can be used to correct this problem; however, they enable good vision only for nearby objects and have to be taken off for distance vision. To avoid this inconvenience, more complicated lenses such as the bifocal lens, which consists of two single-vision lenses with different powers, and the trifocal lens, which adds one more lens for intermediate vision, have been designed. A major drawback of these kinds of lenses is the jump in the image as the eye moves from seeing far-distance to near-distance objects.

One of the best solutions to presbyopia is the progressive addition lens (PAL, for short). A PAL comprises a large distance-view zone with low power on the upper part of the lens and a small near-view zone with higher power on the lower part; between these two zones, the power increases progressively and smoothly.

In progressive lenses, the progression of the power arises from a local variation in the curvatures of the surface. In ophthalmic optics, the power at each point is given by the formula

$$Pow = (1-n)P^b + \frac{(n-1)P^f}{1 - d(1-\frac{1}{n})P^f},$$

where n is the refractive index of the material, d is the thickness of the lens, and P^f and P^b are the mean curvatures of the front and back surfaces, respectively. For a thin lens ($d \ll 1$), the formula simplifies to

$$Pow = (n-1)(P^f - P^b).$$

In typical progressive lenses, the back surface is usually a surface of constant mean curvature. Thus P^f is the only factor determining the progressive power of the lens. To simplify notation, we write it as P.

*Received by the editors May 29, 2002; accepted for publication (in revised form) March 14, 2003; published electronically November 19, 2003.
http://www.siam.org/journals/siap/64-1/40894.html

†Institute for Mathematics and Its Applications, University of Minnesota, 400 Lind Hall, 207 Church Street S.E., Minneapolis, MN 55455 (jwang@ima.umn.edu).

‡School of Mathematics, University of Minnesota-Twin Cities, 127 Vincent Hall, 206 Church Street S.E., Minneapolis, MN 55455 (gulliver@math.umn.edu, santosa@math.umn.edu).

Suppose that k_1, k_2 are the two local principal curvatures of the front surface; then $P = (k_1 + k_2)/2$. The quantity

$$A = (n-1)|k_1 - k_2|$$

is called the local cylinder (or astigmatism).

An ideal progressive surface is one with the prescribed smooth progressive power and with zero astigmatism everywhere. But in order for the astigmatism to be zero everywhere, the surface has to be either a sphere or a plane, which will not provide progressive power. Alternatively, one would wish the surface to have smooth progressive power and as small an astigmatism as possible. In fact, in design of ophthalmic lenses, certain regions on the surface are required to have very small astigmatism; for the rest of the surface, the astigmatism should be smooth and as small as possible. The regions of very small astigmatism are areas of the lenses critical to the eyes for far, intermediate, and near vision.

There have been some direct methods for designing progressive lenses, e.g., those of Winthrop [12] and Baudart, Ahsbahs, and Miege [4]. In a direct method, power is assigned along a curve on the lens. Elsewhere on the lens, power is distributed smoothly away from this curve. The construction is done in such a way that the cylinder is as small as possible in the critical areas of the lens. However, the performance of such lenses are often less than satisfactory because there is no real control over the distribution of the cylinder. Indirect methods based on optimization have been proposed by other lens designers. In such a method, a cost function that attempts to balance the power distribution with the unavoidable cylinder is created. The goal is to minimize the cost function. An example of such a cost function, and one which we will analyze in this work, is that proposed by Loos, Greiner, and Seidel [9]. In their approach, the desired power function over the whole surface is specified. The task then is to construct the surface with power close to this function and with a weighted total cylinder as small as possible.

The current paper addresses theoretical issues arising in lens design by the variational approach. In section 2, we give a brief review of principal curvatures and the compatibility conditions for the existence of surfaces. Section 3 introduces the variational approach for the lens design problem. The Euler–Lagrange equation, which is a fourth-order nonlinear elliptic equation, is also derived there.

A similar variational problem arises in the study of surfaces governed by free energy functionals. In [10], Nitsche analyzed such a problem with constant weight functions and power function. He showed that if the constants satisfy the so-called structural conditions, then for both Dirichlet boundary conditions and natural boundary conditions the solution to the problem exists, is unique, and is in the regular class $C^{4,\lambda}(\overline{\Omega})$, where the Hölder exponent λ is a number in the interval $0 < \lambda < 1$.

For progressive lenses, the power varies among different areas. Also, due to the fact that there is importance associated with critical areas of the lens, the weight functions will not be constants. Thus the Euler–Lagrange equation is more complicated. To simplify the analysis of the partial differential equation, we consider two linearizations.

In section 4, we linearize the equation about a flat surface. The linearized PDE is very similar to the equation of elastic plate theory (see Ciarlet [7]). We will solve the problem on a square region. Three types of boundary conditions are considered: (1) a clamped boundary condition, (2) a partially clamped boundary condition, (3) a natural boundary condition. For all three boundary conditions, we prove the existence

and uniqueness of the solutions with reasonable choices of weight functions α and β. For the third boundary condition, a solution that is unique up to a linear ambiguity exists. Regularity of the solutions follows from Friedrichs's work [8]. In section 5, we consider a linearization near a spherical surface. Similar to the linearization from a plane, we consider the same boundary conditions as the above linearization about the flat surface. We show that, with the first boundary condition, the equation has a unique solution when the size of the square region over which the problem is solved is less than the radius of the base spherical surface. With the second and the third boundary conditions, the solution is unique when the size of the square region is less than 1.15 times the radius of the base spherical surface. Again, for the natural boundary condition case, the solution is unique up to an addition of a linear function.

A numerical approach for this problem is addressed in a separate work [11], where we propose a finite element method to solve the resulting PDEs.

2. Compatibility conditions. It is clear that the power and the cylinder of the front surface are determined only by the principal curvatures. Thus, if we could control the distribution of the principal curvatures, then we could control the quality of the lens. Unfortunately, there are complicated relations governing the two principal curvatures, as shown below.

Given a surface $z = u(x, y)$, we can write it in a parametric form as $U(x_1, x_2) = (u_1(x_1, x_2), u_2(x_1, x_2), u_3(x_1, x_2))$ with $u_1(x_1, x_2) = x_1 = x$, $u_2(x_1, x_2) = x_2 = y$, and $u_3(x_1, x_2) = u(x, y)$. Let $g_{ij} = \langle \frac{\partial U}{\partial x_i}, \frac{\partial U}{\partial x_j} \rangle$. Then the *first fundamental form* is defined as

$$\mathrm{I} = g_{11}dx_1^2 + 2g_{12}dx_1dx_2 + g_{22}dx_2^2.$$

Let $n = \frac{\partial U}{\partial x_1} \times \frac{\partial U}{\partial x_2} / |\frac{\partial U}{\partial x_1} \times \frac{\partial U}{\partial x_2}|$ be the unit outer normal vector to the surface, and let $L_{ij} = \langle \frac{\partial^2 U}{\partial x_i \partial x_j}, n \rangle$. The *second fundamental form* is defined as

$$\mathrm{II} = L_{11}dx_1^2 + 2L_{12}dx_1dx_2 + L_{22}dx_2^2.$$

Suppose that the principal curvatures are k_1 and k_2; then the *Gauss curvature* K and the *mean curvature* H are

$$K = k_1 \cdot k_2 = \frac{L_{11}L_{22} - L_{12}^2}{g_{11}g_{22} - g_{12}^2}, \quad H = \frac{k_1 + k_2}{2} = \frac{L_{11}g_{22} - 2L_{12}g_{12} + L_{22}g_{11}}{2(g_{11}g_{22} - g_{12}^2)^{3/2}}.$$

In terms of the original surface $z = u(x, y)$, H and K can be written explicitly as

$$H = \frac{(1 + u_x^2)u_{yy} - 2u_xu_yu_{xy} + (1 + u_y^2)u_{xx}}{2(1 + u_x^2 + u_y^2)^{3/2}}, \quad K = \frac{u_{xx}u_{yy} - u_{xy}^2}{(1 + u_x^2 + u_y^2)^2}.$$

In general, for any arbitrary specified principal curvatures, the surface may not exist. Certain conditions must be satisfied for the existence of the surface. These conditions are called the *Gauss equations* and *Codazzi equations* and are given by

$$\frac{\partial \Gamma_{jk}^l}{\partial x_i} - \frac{\partial \Gamma_{ik}^l}{\partial x_j} + \Gamma_{jk}^s\Gamma_{is}^l - \Gamma_{ik}^s\Gamma_{js}^l = L_{jk}g^{lm}L_{mi} - L_{ik}g^{lm}L_{mj}, \tag{2.1}$$

$$\Gamma_{ij}^l L_{lk} + \frac{\partial L_{ij}}{\partial x_k} = \Gamma_{kj}^l L_{li} + \frac{\partial L_{kj}}{\partial x_i}, \tag{2.2}$$

where $\Gamma_{ij}^k = \frac{1}{2}g^{lk}(\frac{\partial g_{jl}}{\partial x_i} + \frac{\partial g_{li}}{\partial x_j} - \frac{\partial g_{ij}}{\partial x_l})$ are the *Christoffel symbols* and (g^{ij}) is the inverse matrix of (g_{ij}).

These two conditions are exactly the compatibility conditions for the existence of the surface, as shown in the next theorem (see [3]).

THEOREM 2.1 (fundamental theorem of surface theory). *Let Ω be an open, simply connected subset of $\mathbb{R}^2$. Assume*

$$\mathrm{I} = g_{11}dx_1^2 + 2g_{12}dx_1dx_2 + g_{22}dx_2^2, \quad \mathrm{II} = L_{11}dx_1^2 + 2L_{12}dx_1dx_2 + L_{22}dx_2^2,$$

where g_{ij} and L_{ij} are differentiable functions of x for all $x \in \Omega$. If I *is positive definite and the Gauss equations* (2.1) *and Codazzi equations* (2.2) *are satisfied, then*

(i) *there exists a surface $U : \Omega \to \mathbb{R}^3$ whose first and second fundamental forms are* I *and* II*;*

(ii) *any two surfaces U and V defined on Ω, which have the same first and second fundamental forms, differ by an isometry of $\mathbb{R}^3$.*

3. Variational approach to progressive surface design. As we see from the previous section, certain relationships for the principal curvatures must be met in order that they correspond to a surface. Unfortunately, the relationships are too complicated to be of direct use in lens design. An indirect method, in which the lens design problem is posed as a variational problem of attempting to fit the prescribed power on the surface while minimizing the total weighted cylinder, has been proposed by Loos, Greiner, and Seidel [9]. We will describe such an approach next.

3.1. Design of an objective functional. Let us consider a surface given by $z = u(x, y) : \Omega \to \mathbb{R}$, where the domain $\Omega \in \mathbb{R}^2$. Suppose that the desired distribution of the mean curvature on the graph is specified by a function $P(x, y) : \Omega \to \mathbb{R}$. Then the quality of the surface could be measured by

$$\mathrm{I}(u) = \int_\Omega \left\{ \alpha(x,y) \left(\frac{k_1 - k_2}{2} \right)^2 + \beta(x,y) \left(\frac{k_1 + k_2}{2} - P(x,y) \right)^2 \right\} dx\, dy, \tag{3.1}$$

where $\alpha(x, y), \beta(x, y) : \Omega \to \mathbb{R}$ are two positive weight functions for the surface astigmatism and power, respectively. The weight α will be large in areas where the lens is to have minimum astigmatism, while the weight β will be large in areas where the lens is to have the correct prescribed power. In typical designs, these areas overlap.

In terms of the Gauss curvature $K = k_1 \cdot k_2$ and the mean curvature $H = (k_1 + k_2)/2$, the above functional can be rewritten as

$$\mathrm{I}(u) = \int_\Omega \left\{ \alpha \left(H(x,y)^2 - K(x,y) \right) + \beta \left(H(x,y) - P(x,y) \right)^2 \right\} dx\, dy. \tag{3.2}$$

The task of this approach is to minimize $\mathrm{I}(u)$. At a minimum, we would have constructed a lens with nearly correct powers in the appropriate regions and minimized the astigmatism in these regions.

Numerical optimization methods could be used to minimize the functional. In [9], Loos, Greiner, and Seidel first approximated the functional by a quadratic functional, and then Newton's method was used to locate a minimum. For the weight functions α and β, piecewise constant functions were used. Allione, Ahsbahs, and Saux [2] discussed other types of numerical methods for lens optimization problems.

3.2. Euler–Lagrange equations. For the surface $z = u(x, y)$, H and K can be expressed explicitly as

$$H = \frac{(1+u_x^2)u_{yy} - 2u_x u_y u_{xy} + (1+u_y^2)u_{xx}}{2(1+u_x^2+u_y^2)^{3/2}} \quad \text{and} \quad K = \frac{u_{xx}u_{yy} - u_{xy}^2}{(1+u_x^2+u_y^2)^2}. \tag{3.3}$$

Thus

$$\begin{aligned}
\mathrm{I}(u) &= \mathrm{I}(u_x, u_y, u_{xx}, u_{xy}, u_{yy}) \\
&= \int_\Omega \Big\{ \alpha(x,y) \left(H(u_x, u_y, u_{xx}, u_{xy}, u_{yy})^2 - K(u_x, u_y, u_{xx}, u_{xy}, u_{yy}) \right) \\
&\qquad + \beta(x,y) \left(H(u_x, u_y, u_{xx}, u_{xy}, u_{yy}) - P(x,y) \right)^2 \Big\} \, dxdy.
\end{aligned}$$

The necessary condition for $\mathrm{I}(u)$ to be the minimum is

$$\left. \frac{d}{dt} \right|_{t=0} \mathrm{I}(u + t\delta) = 0 \quad \forall \delta \in C_0^\infty(\Omega),$$

i.e.,

$$\begin{aligned}
0 &= \left. \frac{d}{dt} \right|_{t=0} \mathrm{I}(u + t\delta) \\
&= \int_\Omega \Big\{ 2 \left((\alpha+\beta)H - \beta P \right) \left(H_{u_x}\delta_x + H_{u_y}\delta_y + H_{u_{xx}}\delta_{xx} + H_{u_{xy}}\delta_{xy} + H_{u_{yy}}\delta_{yy} \right) \\
&\qquad - \alpha \left(K_{u_x}\delta_x + K_{u_y}\delta_y + K_{u_{xx}}\delta_{xx} + K_{u_{xy}}\delta_{xy} + K_{u_{yy}}\delta_{yy} \right) \Big\} \, dxdy.
\end{aligned} \tag{3.4}$$

Suppose that α, β, and P are at least twice differentiable; then integration by parts twice yields the following fourth-order PDE:

$$\begin{aligned}
&(\alpha K_{u_x})_x + (\alpha K_{u_y})_y - (\alpha K_{u_{xx}})_{xx} - (\alpha K_{u_{xy}})_{xy} - (\alpha K_{u_{yy}})_{yy} \\
&\quad - 2\left((\alpha H + \beta H - \beta P) H_{u_x} \right)_x - 2\left((\alpha H + \beta H - \beta P) H_{u_y} \right)_y \\
&\quad + 2\left((\alpha H + \beta H - \beta P) H_{u_{xx}} \right)_{xx} + 2\left((\alpha H + \beta H - \beta P) H_{u_{xy}} \right)_{xy} \\
&\quad + 2\left((\alpha H + \beta H - \beta P) H_{u_{yy}} \right)_{yy} = 0.
\end{aligned} \tag{3.5}$$

This is the *Euler–Lagrange equation* of the problem.

For the nonlinear fourth-order PDE (3.5), it is easy to check that the fourth-order terms contributed by K are zero, and thus all the fourth-order terms are from H. The fourth-order leading terms are

$$\begin{aligned}
P_4(D)u = \frac{(\alpha+\beta)}{(1+u_x^2+u_y^2)^3} \Big\{ &(1+u_y^2)[(1+u_x^2)u_{yyxx} - 2u_x\, u_y\, u_{xyxx} + (1+u_y^2)u_{xxxx}] \\
&- 2u_x\, u_y[(1+u_x^2)u_{yyxy} - 2u_x\, u_y\, u_{xyxy} + (1+u_y^2)u_{xxxy}] \\
&+ (1+u_x^2)[(1+u_x^2)u_{yyyy} - 2u_x\, u_y\, u_{xyyy} + (1+u_y^2)u_{xxyy}] \Big\}.
\end{aligned}$$

Since

$$P_4(\xi,\eta) = \frac{(\alpha+\beta)}{(1+u_x^2+u_y^2)^3}[(1+u_x^2)\eta^2 - 2u_x\,u_y\xi\,\eta + (1+u_y^2)\xi^2]^2$$

$$= \frac{(\alpha+\beta)}{(1+u_x^2+u_y^2)^3}[\xi^2+\eta^2+(u_x\eta - u_y\xi)^2]^2$$

$$\geq \frac{(\alpha+\beta)}{(1+u_x^2+u_y^2)^3}[\xi^2+\eta^2]^2,$$

we can conclude that the PDE (3.5) is elliptic if $\alpha+\beta > 0$.

To simplify the analysis of (3.5), we are going to linearize it. Two types of linearizations will be considered. The first linearization is about a plane, under the assumption that $|\nabla u| \ll 1$. The second linearization is about a sphere; namely, we will assume that the surface is a slight perturbation from a spherical surface and suppose that the gradient of the perturbation is small. Due to the fact that a lens is closer to a sphere than to a plane, the second linearization is of more practical value for lens design. The techniques used in the analysis of the linearized equation about a plane will be useful in studying the case of linearization about a sphere.

4. Linearization about the flat surface. If $|\nabla u| \ll 1$, then approximately $H \approx \frac{1}{2}\Delta u$ and $K \approx u_{xx}u_{yy} - u_{xy}^2$. In this case, the variational equation will be

$$\int_\Omega \{(\alpha+\beta)\Delta u\Delta v - 2\alpha(u_{xx}v_{yy} + u_{yy}v_{xx} - 2u_{xy}v_{xy})\}\,dx\,dy = \int_\Omega 2\beta P\Delta v\,dx\,dy \tag{4.1}$$

for all $v \in C_0^\infty(\Omega)$. The corresponding PDE is (again, if all the coefficients are at least twice differentiable)

$$\Delta((\alpha+\beta)\Delta u) - 2[(\alpha u_{yy})_{xx} + (\alpha u_{xx})_{yy} - 2(\alpha u_{xy})_{xy}] = \Delta(2\beta P). \tag{4.2}$$

Equation (4.1) is very similar to the equation of Kirchoff plate theory [7]. Borrowing terms used in plate theory, the boundary condition (BC) can be one of the following types:

(i) clamped,
(ii) partially clamped,
(iii) natural.

Define the bilinear functional

$$B(u,v) = \int_\Omega (\alpha+\beta)\Delta u\Delta v - 2\alpha(u_{xx}v_{yy} + u_{yy}v_{xx} - 2u_{xy}v_{xy})dx\,dy \tag{4.3}$$

and the linear functional

$$L(v) = \int_\Omega 2\beta P\Delta v\,dx\,dy. \tag{4.4}$$

Then the problem is to solve

$$B(u,v) = L(v) \quad \forall v \in C_0^\infty(\Omega). \tag{4.5}$$

The basic tool we use to establish the results in the subsequent analysis is the Lax–Milgram lemma [7], which we restate here for completeness.

LEMMA 4.1 (Lax–Milgram lemma). *Let V be a Banach space with norm $\|\cdot\|$. Let $L : V \to \mathbb{R}$ be a continuous linear form, and let $B : V \times V \to \mathbb{R}$ be a symmetric and continuous bilinear form that is V-elliptic, in the sense that there exists a constant $c > 0$ such that*

$$B(v, v) \geq c\|v\|^2 \quad \forall v \in V.$$

Then the problem: Find $u \in V$ such that

$$B(u, v) = L(v) \quad \forall v \in V$$

has a unique solution.

The space we will use for establishing the results is $H^2(\Omega)$, for which the norm is defined as (see [1])

$$\|v\|_{2,\Omega} = \left\{ |v|_{0,\Omega}^2 + \sum_{i=1}^{2} |\partial_i v|_{0,\Omega}^2 + \sum_{i,j=1}^{2} |\partial_{ij} v|_{0,\Omega}^2 \right\}^{1/2},$$

where $|v|_{0,\Omega} = \left\{ \int_\Omega |v|^2 dx\, dy \right\}^{1/2}$, $\partial_1 = \partial/\partial x$, $\partial_2 = \partial/\partial y$, and $\partial_{ij} = \partial_i \partial_j$. Define the seminorm $|\cdot|_{2,\Omega}$ by

$$|v|_{2,\Omega} = \left\{ \sum_{i,j=1}^{2} |\partial_{ij} v|_{0,\Omega}^2 \right\}^{1/2}.$$

4.1. BC (i): Clamped boundary condition. We suppose that u is clamped on the boundary, i.e., $u = 0, \nabla u = 0$ on $\partial\Omega$.

An ingredient critical for this boundary condition is the Poincaré–Friedrichs inequality [5].

LEMMA 4.2 (Poincaré–Friedrichs inequality). *Let Ω be a domain in $\mathbb{R}^2$; then there exists a constant $c > 0$ such that*

$$c^{-1}\|v\|_{2,\Omega} \leq |v|_{2,\Omega} \quad \forall v \in H_0^2(\Omega). \tag{4.6}$$

THEOREM 4.3. *Let Ω be a domain in $\mathbb{R}^2$. Assume that $P(x, y), \alpha(x, y), \beta(x, y) \leq M$ in Ω, where $M > 0$ is a constant; also $\alpha(x, y) \geq 0$ and there exists a constant $\beta_0 > 0$ such that $\beta(x, y) \geq \beta_0$ for all $(x, y) \in \Omega$. Then the variational equation (4.1) has a unique solution in $H_0^2(\Omega)$. In addition, if $\alpha(x, y), \beta(x, y) \in C^k(\Omega)$ and $P(x, y) \in H^k(\Omega)$, then for every compact subdomain Ω_1 of Ω, the solution $u \in H^{k+2}(\Omega_1)$.*

Proof. For all $v \in C_0^\infty(\Omega)$, by Green's formula,

$$\int_\Omega \partial_{ij} v\, \partial_{ij} v\, dx\, dy = \int_\Omega \partial_{iij} v\, \partial_j v\, dx\, dy = \int_\Omega \partial_{ii} v\, \partial_{jj} v\, dx\, dy.$$

Since $C_0^\infty(\Omega)$ is dense in $H_0^2(\Omega)$, the above relation remains true for all $v \in H_0^2(\Omega)$; i.e., for all $v \in H_0^2(\Omega)$, $|v|_{2,\Omega} = |\Delta v|_{0,\Omega}$. Thus by the Poincaré–Friedrichs inequality (Lemma 4.2),

$$|\Delta v|_{0,\Omega} = |v|_{2,\Omega} \geq c^{-1}\|v\|_{2,\Omega}.$$

Since $\alpha((\Delta v)^2 - 4v_{xx}v_{yy} + 4v_{xy}^2)) = \alpha((v_{xx} - v_{yy})^2 + 4v_{xy}^2) \geq 0$, this implies

$$B(v, v) \geq \beta|\Delta v|_{0,\Omega}^2 \geq c^{-2}\beta_0\|v\|_{2,\Omega}^2,$$

and thus $B(v, v)$ is $H_0^2(\Omega)$-elliptic. Thus, by the Lax–Milgram lemma, the variational problem has a unique solution in $H_0^2(\Omega)$.

The regularity of the solution follows from the result in Friedrichs [8]. ∎

Remark. Note that the conditions here for α and β resemble the so-called structural conditions in Nitsche [10].

4.2. BC (ii): Partially clamped boundary conditions. Let $H^2_\Gamma(\Omega)$ be the subspace of $H^2(\Omega)$ such that $f|_\Gamma = 0$, $\frac{\partial f}{\partial \nu}|_\Gamma = 0$, where Γ is part of the boundary of Ω with positive measure. A Poincaré-type inequality is still valid, as shown in [7].

LEMMA 4.4. *Let Ω be a domain in $\mathbb{R}^2$, and let Γ be a measurable subset of $\partial\Omega$ with length $(\Gamma) > 0$. Then there exists a constant $c > 0$ such that*

$$c^{-1}\|v\|_{2,\Omega} \le |v|_{2,\Omega} \quad \forall v \in H^2_\Gamma(\Omega).$$

THEOREM 4.5. *Let Ω be a domain in $\mathbb{R}^2$; assume that $P(x,y)$, $\alpha(x,y)$, and $\beta(x,y)$ are bounded above in Ω, and that $\beta(x,y) \ge \alpha(x,y) \ge \alpha_0$ for all $(x,y) \in \Omega$, where $\alpha_0 > 0$ is a constant. Then the variational equation (4.1) has a unique solution in $H^2_\Gamma(\Omega)$. In addition, if $\alpha(x,y), \beta(x,y) \in C^k(\Omega)$ and $P(x,y) \in H^k(\Omega)$, then for every compact subdomain Ω_1 of Ω, the solution $u \in H^{k+2}(\Omega_1)$.*

Proof. When $\beta(x,y) \ge \alpha(x,y) \ge \alpha_0$, we have $(\alpha+\beta)\Delta u \Delta u \ge 2\alpha(\Delta u)^2$, and so

$$\begin{aligned} B(u,u) &\ge \int_\Omega 2\alpha((\Delta u)^2 - 2u_{xx}u_{yy} + 2u_{xy}^2)\,dx\,dy \\ &= \int_\Omega 2\alpha(|u_{xx}|^2 + |u_{yy}|^2 + 2|u_{xy}|^2)\,dx\,dy \\ &\ge 2\alpha_0 |u|^2_{2,\Omega}. \end{aligned}$$

By Lemma 4.4,

$$B(u,u) \ge 2\alpha_0 c^{-2}\|u\|^2_{2,\Omega}.$$

Then the Lax–Milgram lemma implies that the problem has a unique solution in $H^2_\Gamma(\Omega)$. The regularity of the solution follows from the result in Friedrichs [8]. ∎

This means that if the solution is fixed on a part of the boundary and free on the rest of the boundary, then the PDE has a unique solution.

4.3. BC (iii): Natural boundary conditions. To derive the natural boundary condition, we return to the variational equation (4.5). For the first term of $B(u,v)$ in (4.3), by Green's formula, we have

$$\begin{aligned} \int_\Omega (\alpha+\beta)\Delta u\,\Delta v\,dx\,dy &= \int_\Omega v\Delta((\alpha+\beta)\Delta u)\,dx\,dy \\ &\quad + \int_{\partial\Omega}\left[(\alpha+\beta)\Delta u\frac{\partial v}{\partial \nu} - \frac{\partial((\alpha+\beta)\Delta u)}{\partial \nu}v\right] dS. \end{aligned}$$

For the second term of $B(u,v)$ in (4.3), notice that (with Einstein's convention)

$$u_{xx}v_{yy} + u_{yy}v_{xx} - 2u_{xy}v_{xy} = u_{ii}v_{jj} - u_{ij}v_{ij}.$$

Using Green's formula again, we arrive at

$$\int_\Omega \alpha u_{ii}\,v_{jj}\,dx\,dy = \int_\Omega (\alpha u_{ii})_{jj}\,v\,dx\,dy + \int_{\partial\Omega}\alpha u_{ii}\,v_j\,\nu_j\,dS - \int_{\partial\Omega}(\alpha u_{ii})_j\,v\,\nu_j\,dS,$$

$$\int_\Omega \alpha u_{ij}\,v_{ij}\,dx\,dy = \int_\Omega (\alpha u_{ij})_{ij}\,v\,dx\,dy + \int_{\partial\Omega}\alpha u_{ij}\,v_j\,\nu_i\,dS - \int_{\partial\Omega}(\alpha u_{ij})_i\,v\,\nu_j\,dS.$$

Thus the second term of $B(u,v)$ becomes

$$\begin{aligned}
&\int_\Omega \alpha(u_{xx}v_{yy}+u_{yy}v_{xx}-2u_{xy}v_{xy})\,dx\,dy\\
&\quad=\int_\Omega[(\alpha u_{ii})_{jj}-(\alpha u_{ij})_{ij}]\,v\,dx\,dy\\
&\qquad+\int_{\partial\Omega}[(\alpha_i u_{ij}\,\nu_j-\alpha_j u_{ii}\,\nu_j)\,v+\alpha u_{ii}\,v_j\,\nu_j-\alpha u_{ij}\,v_j\,\nu_i]\,dS.
\end{aligned}$$

To simplify this further, let τ be the unit tangent vector along $\partial\Omega$ and ν be the unit outward normal vector to $\partial\Omega$; then we have

$$\begin{cases} v_\nu = v_x\nu_x + v_y\nu_y,\\ v_\tau = -v_x\nu_y + v_y\nu_x \end{cases}$$

and

$$\begin{cases} v_x = v_\nu\nu_x - v_\tau\nu_y,\\ v_y = v_\nu\nu_y + v_\tau\nu_x. \end{cases}$$

Using these, we obtain that

$$\int_{\partial\Omega}(\alpha_i u_{ij}\nu_j-\alpha_j u_{ii}\nu_j)\,v\,dS=\int_{\partial\Omega}(-\alpha_x u_{y\tau}+\alpha_y u_{x\tau})\,v\,dS.$$

Moreover,

$$\begin{aligned}
&\int_{\partial\Omega}(\alpha u_{ii}v_j\nu_j-\alpha u_{ij}v_j\nu_i)\,dS\\
&\quad=-\int_{\partial\Omega}\alpha(u_{yx}\nu_y-u_{yy}\nu_x)v_x+\alpha(u_{xy}\nu_x-u_{xx}\nu_y)v_y\,dS\\
&\quad=\int_{\partial\Omega}(\alpha u_{y\tau}v_x-\alpha u_{x\tau}v_y)\,dS\\
&\quad=\int_{\partial\Omega}-\alpha(u_{x\tau}\nu_x+u_{y\tau}\nu_y)v_\tau+\alpha(u_{y\tau}\nu_x-u_{x\tau}\nu_y)v_\nu\,dS\\
&\quad=\int_{\partial\Omega}(\alpha(u_{x\tau}\nu_x+u_{y\tau}\nu_y))_\tau v+\alpha(u_{y\tau}\nu_x-u_{x\tau}\nu_y)v_\nu\,dS.
\end{aligned}$$

The last equality is obtained by integration by parts along the boundary.

Also

$$\int_\Omega \beta P\Delta v\,dx\,dy=\int_\Omega\Delta(\beta P)v\,dx\,dy+\int_{\partial\Omega}\left[\beta P\frac{\partial v}{\partial\nu}-\frac{\partial(\beta P)}{\partial\nu}v\right]dS.$$

Thus the natural boundary conditions are

(4.7)
$$\begin{cases} 2(\alpha_x u_{y\tau}-\alpha_y u_{x\tau})-\dfrac{\partial((\alpha+\beta)\Delta u)}{\partial\nu}-2(\alpha(u_{x\tau}\nu_x+u_{y\tau}\nu_y))_\tau+2\dfrac{\partial(\beta P)}{\partial\nu}=0,\\ (\alpha+\beta)\Delta u-2\alpha(u_{y\tau}\nu_x-u_{x\tau}\nu_y)-2\beta P=0. \end{cases}$$

Notice that the differential equation (4.5) and the natural boundary conditions (4.7) involve only the second and higher derivatives of u. Therefore if $u(x, y)$ is a solution of the problem, then for any linear function $l(x, y)$ in $\mathbb{R}^2$, $u(x, y)+l(x, y)$ will also be a solution of the problem. That is, the solution is not unique. We will show that the kernel is exactly the space of all the linear functions. We can then apply Fredholm's alternative after eliminating the finite dimensional kernel and establish the well-posedness of the problem.

4.3.1. Kernel of the equation. To show that the kernel is exactly the three dimensional space of linear functions, we first impose some constraints on u.

Since $C^2(\Omega)$ is dense in $H^2(\Omega)$, we first assume $u \in C^2(\Omega)$ with the following constraints:

$$\int_\Omega u \, dx\, dy = 0, \quad \int_\Omega u_x \, dx\, dy = 0, \quad \int_\Omega u_y \, dx\, dy = 0. \tag{4.8}$$

As we mentioned above, with the natural boundary condition, the PDE is equivalent to the variational equation

$$B(u, v) = L(v), \quad u, v \in H^2(\Omega).$$

Here, we will need a more general version of the Poincaré inequality [6].

LEMMA 4.6 (generalized Poincaré inequality). *Let Ω be a domain in $\mathbb{R}^2$, and let $1 \le p < \infty$; then there exists a constant c_0 such that*

$$\int_\Omega |v|^p dx\, dy \le c_0 \left\{ \int_\Omega |\nabla v|^p dx\, dy + \left| \int_\Omega v \, dx\, dy \right|^p \right\} \tag{4.9}$$

for all $v \in W^{1,p}(\Omega)$, where $c_0 = c_0(\Omega)$.

COROLLARY 4.7. *Let Ω be a domain in $\mathbb{R}^2$ and $u(x, y) \in H^2(\Omega)$ with the constraints* (4.8)*; then there exists a constant c such that*

$$\|u\|_{2,\Omega} \le c|u|_{2,\Omega}.$$

Proof. First assume $u \in C^2(\Omega)$; then, since $\int_\Omega u \, dx\, dy = 0$, we have immediately from (4.9)

$$\int_\Omega |u|^2 dx\, dy \le c_0 \int_\Omega (|u_x|^2 + |u_y|^2) dx\, dy.$$

Similarly, due to the constraints that $\int_\Omega u_x \, dx\, dy = 0$ and $\int_\Omega u_y \, dx\, dy = 0$, we have

$$\int_\Omega |u_x|^2 dx\, dy \le c_0 \int_\Omega (|u_{xx}|^2 + |u_{xy}|^2) dx\, dy,$$

$$\int_\Omega |u_y|^2 dx\, dy \le c_0 \int_\Omega (|u_{xy}|^2 + |u_{yy}|^2) dx\, dy.$$

Combining the results, we get

$$\int_\Omega |u|^2 dx\, dy \le c_0^2 \int_\Omega (|u_{xx}|^2 + 2|u_{xy}|^2 + |u_{yy}|^2) dx\, dy,$$

$$\int_\Omega (|u_x|^2 + |u_y|^2) dx\, dy \le c_0 \int_\Omega (|u_{xx}|^2 + 2|u_{xy}|^2 + |u_{yy}|^2) dx\, dy,$$

which implies

$$\|u\|_{2,\Omega} \le \sqrt{1 + c_0 + c_0^2}\ |u|_{2,\Omega}.$$

Because $C^2(\Omega)$ is dense in $H^2(\Omega)$, the above inequality is also true for all $u \in H^2(\Omega)$. □

With Corollary 4.7, the result for the partially clamped boundary condition can be applied to the natural boundary condition.

THEOREM 4.8. *Let Ω be a domain in $\mathbb{R}^2$; assume that $P(x,y)$, $\alpha(x,y)$, and $\beta(x,y)$ are bounded above in Ω; and take $\beta(x,y) \ge \alpha(x,y) \ge \alpha_0$ for all $(x,y) \in \Omega$, where $\alpha_0 > 0$ is a constant. Then, with the constraints* (4.8)*, the variational equation* (4.5) *has a unique solution in $H^2(\Omega)$. In addition, if $\alpha(x,y), \beta(x,y) \in C^k(\Omega)$ and $P(x,y) \in H^k(\Omega)$, then for every compact subdomain Ω_1 of Ω, the solution $u \in H^{k+2}(\Omega_1)$.*

We denote the solution to (4.5) with natural boundary condition (4.7) and constraints (4.8) by u_0. Now let u be an arbitrary solution of the variational equation without the constraints, and let

$$b = \frac{1}{|\Omega|}\int_\Omega u_x\,dx\,dy, \quad c = \frac{1}{|\Omega|}\int_\Omega u_y\,dx\,dy,$$

and

$$a = \frac{1}{|\Omega|}\int_\Omega (u - bx - cy)dx\,dy.$$

Define $\tilde{u} = u - (a + bx + cy)$; then $\tilde{u}$ is also a solution of the variational equation and

$$\int_\Omega \tilde{u}\,dx\,dy = 0, \quad \int_\Omega \tilde{u}_x\,dx\,dy = 0, \quad \int_\Omega \tilde{u}_y\,dx\,dy = 0.$$

By Theorem 4.8, $\tilde{u} \equiv u_0$ in $H^2(\Omega)$. This means that any solution to the variational equation must differ from u_0 by a linear function. We summarize the result in the following.

THEOREM 4.9. *Let Ω be a domain in $\mathbb{R}^2$; assume that $P(x,y)$, $\alpha(x,y)$, and $\beta(x,y)$ are bounded above in Ω; and assume $\beta(x,y) \ge \alpha(x,y) \ge \alpha_0$ for all $(x,y) \in \Omega$, where $\alpha_0 > 0$ is a constant. Then the solution to the variational equation* (4.1) *is unique in $H^2(\Omega)$ up to the addition of a linear function. In addition, if $\alpha(x,y)$, $\beta(x,y) \in C^k(\Omega)$ and $P(x,y) \in H^k(\Omega)$, then for every compact subdomain Ω_1 of Ω, the solution $u \in H^{k+2}(\Omega_1)$.*

5. Linearization about spherical surfaces. Since a progressive surface is in practice a slight perturbation from a spherical surface, it is certainly reasonable to suppose

$$u = u_0 + v,$$

where u_0 is a prescribed spherical surface of radius R and v is the deviation from it. Then a straightforward calculation from (3.3) leads to

$$H(u) = \frac{1}{R} + \frac{(1 + u_{0x}^2)v_{yy} - 2u_{0x}u_{0y}v_{xy} + (1 + u_{0y}^2)v_{xx}}{2g^3} + E_1$$

and

$$K(u) = \frac{(u_{0xx} + v_{xx})(u_{0yy} + v_{yy}) - (u_{0xy} + v_{xy})^2}{g^4} + E_2,$$

where $g = \sqrt{1 + u_{0x}^2 + u_{0y}^2}$ and E_1, E_2 are the error terms which satisfy

$$|E_i| \le C_i |\nabla v|, \quad i = 1, 2.$$

Here $C_i, i = 1, 2$, are two constants which depend only on $|\nabla v|$, $|D^2(v)|$, $|\nabla u_0|$, and $|D^2(u_0)|$.

Suppose that the domain satisfies $x^2 + y^2 \le cR^2$ with constant $c < 1$. In addition, suppose $|\nabla v| \ll 1$ and that $|D^2(v)|$ is bounded above in the domain. Then, $|E_i| \ll 1, i = 1, 2$, and thus

$$H(u) \approx \frac{1}{R} + \frac{(1 + u_{0x}^2)v_{yy} - 2u_{0x}u_{0y}v_{xy} + (1 + u_{0y}^2)v_{xx}}{2g^3},$$

$$K(u) \approx \frac{(u_{0xx} + v_{xx})(u_{0yy} + v_{yy}) - (u_{0xy} + v_{xy})^2}{g^4}.$$

Let

$$H_{u_0}(v) = \frac{(1 + u_{0x}^2)v_{yy} - 2u_{0x}u_{0y}v_{xy} + (1 + u_{0y}^2)v_{xx}}{2g^3}, \tag{5.1}$$

$$K_{u_0}(w, v) = \frac{w_{xx}v_{yy} + w_{yy}v_{xx} - 2w_{xy}v_{xy}}{g^4}. \tag{5.2}$$

Then $H_{u_0}(v)$ is linear and $K_{u_0}(w, v)$ is bilinear and symmetric. Consider again the functional (3.2):

$$\mathrm{I}(u) = \int_\Omega [\alpha(x, y)(H(u)^2 - K(u)) + \beta(x, y)(H(u) - P(x, y))^2] dx\, dy.$$

Assume that u minimizes the above functional, and let $\delta \in C_0^\infty(\Omega)$ be the test function. By computing the first variation of $\mathrm{I}(u)$ at u and using the notation (5.1) and (5.2), we will get

$$\int_\Omega [2(\alpha + \beta)H_{u_0}(v)H_{u_0}(\delta) - \alpha K_{u_0}(\delta, v)]\, dx\, dy$$

$$= \int_\Omega \left[\alpha K_{u_0}(\delta, u_0) + 2\left(\beta P - \frac{\alpha + \beta}{R}\right) H_{u_0}(\delta)\right] dx\, dy.$$

Let

$$B(w, v) = \int_\Omega [2(\alpha + \beta)H_{u_0}(w)H_{u_0}(v) - \alpha K_{u_0}(w, v)]\, dx\, dy \tag{5.3}$$

and

$$L(w) = \int_\Omega \left[\alpha K_{u_0}(w, u_0) + 2\left(\beta P - \frac{\alpha + \beta}{R}\right) H_{u_0}(w)\right] dx\, dy. \tag{5.4}$$

Then, it is clear that $B(w, v)$ is bilinear and $L(w)$ is linear. Thus the problem is to find $v \in V$ such that

$$B(w, v) = L(w) \quad \forall w \in V, \tag{5.5}$$

where V is a subspace of $H^2(\Omega)$.

Similar to the linearization about the flat surface, three types of BCs will be considered:

(i) clamped,
(ii) partially clamped,
(iii) natural.

5.1. BC (i): Clamped boundary condition.

THEOREM 5.1. *Let Ω be a subdomain of a square region of size d centered at $(0,0)$ in $\mathbb{R}^2$, and assume that $P(x,y)$, $\alpha(x,y)$, and $\beta(x,y)$ are bounded above in Ω. Assume further that $\alpha(x,y) \geq 0$ and that there exists a constant $\beta_0 > 0$ such that $\beta(x,y) \geq \beta_0$ for all $(x,y) \in \Omega$. If $d < 0.8165\,R$, then the variational equation (5.5) has a unique solution in $H_0^2(\Omega)$. In addition, if $\alpha(x,y), \beta(x,y) \in C^k(\Omega)$ and $P(x,y) \in H^k(\Omega)$, then for every compact subdomain Ω_1 of Ω, the solution $u \in H^{k+2}(\Omega_1)$.*

Proof. Since $u_0(x,y)$ is a spherical surface, then $u_0(x,y)$ could be expressed as $u_0(x,y) = -\sqrt{R^2 - x^2 - y^2}$. By direct calculation, we will have

$$H_{u_0}(v) = \frac{\sqrt{R^2 - x^2 - y^2}[(R^2 - y^2)v_{yy} - 2xyv_{xy} + (R^2 - x^2)v_{xx}]}{2R^3}, \tag{5.6}$$

$$K_{u_0}(v,v) = \frac{(R^2 - x^2 - y^2)^2(2v_{xx}v_{yy} - 2v_{xy}^2)}{R^4}. \tag{5.7}$$

We can rewrite (5.3) as

$$\begin{aligned} B(v,v) &= \int_\Omega \left\{2(\alpha+\beta)H_{u_0}^2(v) - \alpha K_{u_0}(v,v)\right\} dx\,dy \\ &= \int_\Omega \alpha\left(2H_{u_0}^2(v) - K_{u_0}(v,v)\right) dx\,dy + 2\int_\Omega \beta H_{u_0}^2(v) dx\,dy. \end{aligned} \tag{5.8}$$

The first integrand of $B(v,v)$ is

$$\begin{aligned} &\alpha\left(2H_{u_0}^2(v) - K_{u_0}(v,v)\right) \\ &\quad = \frac{\alpha(R^2 - x^2 - y^2)}{2R^6}\Big[\left((R^2 - y^2)v_{yy} - 2xyv_{xy} + (R^2 - x^2)v_{xx}\right)^2 \\ &\qquad\qquad - 2\,R^2(R^2 - x^2 - y^2)(2v_{xx}v_{yy} - 2v_{xy}^2)\Big]. \end{aligned}$$

Now, consider the matrix

$$A = \begin{pmatrix} R^2 - x^2 & -xy \\ -xy & R^2 - y^2 \end{pmatrix}\begin{pmatrix} v_{xx} & v_{xy} \\ v_{xy} & v_{yy} \end{pmatrix}.$$

The trace of A is

$$\mathrm{tr}(A) = \left[(R^2 - y^2)v_{yy} - 2xyv_{xy} + (R^2 - x^2)v_{xx}\right],$$

and the determinant of A is

$$\det(A) = R^2(R^2 - x^2 - y^2)(v_{xx}v_{yy} - v_{xy}^2).$$

By the fact that $\mathrm{tr}(A)^2 - 4\det(A) \geq 0$ for any 2×2 matrix A, we have

$$\alpha\,(2H_{u_0}(v)H_{u_0}(v) - K_{u_0}(v,v)) \geq 0.$$

It remains to analyze the second part of $B(v,v)$ in (5.8):

$$\begin{aligned}
&[(R^2-y^2)v_{yy} - 2xyv_{xy} + (R^2-x^2)v_{xx}]^2 \\
&\quad = [R^2\Delta v - (y^2v_{yy} + x^2v_{xx} + 2xyv_{xy})]^2 \\
&\quad = R^4(\Delta v)^2 - 2R^2\Delta v(y^2v_{yy} + x^2v_{xx} + 2xyv_{xy}) \\
&\quad + (y^2v_{yy} + x^2v_{xx} + 2xyv_{xy})^2.
\end{aligned}$$

Assume $|x|, |y| \leq r = \delta R$, where $\delta \in [0,1]$ is to be chosen later. Then

$$|y^2v_{yy} + x^2v_{xx} + 2xyv_{xy}| \leq \delta^2R^2(|v_{xx}| + |v_{yy}| + 2|v_{xy}|).$$

By the Schwarz inequality,

$$\begin{aligned}
&|2R^2\Delta v(y^2v_{yy} + x^2v_{xx} + 2xyv_{xy})| \\
&\qquad \leq 2R^4\delta^2(|v_{xx}| + |v_{yy}|)(|v_{xx}| + |v_{yy}| + 2|v_{xy}|) \\
&\qquad \leq 2R^4\delta^2(3|v_{xx}|^2 + 3|v_{yy}|^2 + 2|v_{xy}|^2);
\end{aligned}$$

thus, we have

$$\begin{aligned}
&[(R^2-y^2)v_{yy} - 2xyv_{xy} + (R^2-x^2)v_{xx}]^2 \\
&\qquad \geq R^4[(\Delta v)^2 - 2\delta^2(3|v_{xx}|^2 + 3|v_{yy}|^2 + 2|v_{xy}|^2)] \\
&\qquad \geq R^4[(\Delta v)^2 - 6\delta^2(|v_{xx}|^2 + |v_{yy}|^2 + 2|v_{xy}|^2)].
\end{aligned}$$

If $1 - 6\delta^2 = \eta > 0$, i.e., if $\delta < 1/\sqrt{6} \approx 0.4082$, then the above inequality implies

$$\begin{aligned}
&[(R^2-y^2)v_{yy} - 2xyv_{xy} + (R^2-x^2)v_{xx}]^2 \\
&\qquad \geq R^4[(\Delta v)^2 - (1-\eta)(|v_{xx}|^2 + |v_{yy}|^2 + 2|v_{xy}|^2)]
\end{aligned}$$

and

$$\begin{aligned}
&2\int_\Omega \beta H_{u_0}(v)H_{u_0}(v)dx\,dy \\
&\qquad \geq \beta_0\frac{1-2\delta^2}{2}\int_\Omega \{(\Delta v)^2 - (1-\eta)(|v_{xx}|^2 + |v_{yy}|^2 + 2|v_{xy}|^2)\}dx\,dy \\
&\qquad = \beta_0\frac{1-2\delta^2}{2}\eta|v|^2_{2,\Omega} \\
&\qquad = c_0|v|^2_{2,\Omega},
\end{aligned}$$

where $c_0 = \beta_0\frac{1-2\delta^2}{2}\eta > 0$. Here, we have used the fact that in $H_0^2(\Omega)$, $|\Delta v|_{0,\Omega} = |v|_{2,\Omega}$. Finally, by the Poincaré–Friedrichs inequality (4.6) and the Lax–Milgram lemma (Lemma 4.1), equation (5.5) has a unique solution in $H_0^2(\Omega)$.

That is, if the size of the square region is less than $0.8165\,R$, then the problem could be uniquely solved. The regularity of the solution follows from the result in Friedrichs [8]. ∎

Remarks.

1. If we use the Cauchy ϵ-inequality ($ab \le \epsilon a^2 + \frac{b^2}{4\epsilon}$) instead of the Schwarz inequality, we will have

$$\begin{aligned}
&|2R^2\Delta v(y^2 v_{yy} + x^2 v_{xx} + 2xyv_{xy})| \\
&\qquad \le 2R^4\delta^2(|v_{xx}| + |v_{yy}|)(|v_{xx}| + |v_{yy}| + 2|v_{xy}|) \\
&\qquad \le 2R^4\delta^2\left((2+2\epsilon)(|v_{xx}|^2 + |v_{yy}|^2) + 2\frac{1}{2\epsilon}|v_{xy}|^2\right).
\end{aligned}$$

   By choosing $\epsilon = (\sqrt{2}-1)/2$, $2 + 2\epsilon = 1/(2\epsilon)$, we have

$$\begin{aligned}
&[(R^2 - y^2)v_{yy} - 2xyv_{xy} + (R^2 - x^2)v_{xx}]^2 \\
&\qquad \ge R^4[(\Delta v)^2 - 2(\sqrt{2}+1)\delta^2(|v_{xx}|^2 + |v_{yy}|^2 + 2|v_{xy}|^2)].
\end{aligned}$$

   Thus if $1 - 2(\sqrt{2}+1)\delta^2 > 0$, i.e., if $\delta < \sqrt{1/(2(\sqrt{2}+1))} \approx 0.4551$, then the conclusion is still true.

2. If we solve the equation on a circular region $x^2 + y^2 \le r^2$ instead of a square region and assume $r = \delta R$, then with the Schwarz inequality, we will have

$$\begin{aligned}
&[(R^2 - y^2)v_{yy} - 2xyv_{xy} + (R^2 - x^2)v_{xx}]^2 \\
&\qquad \ge R^4[(\Delta v)^2 - 5\delta^2(|v_{xx}|^2 + |v_{yy}|^2 + 2|v_{xy}|^2)].
\end{aligned}$$

   This means that if $\delta < 1/\sqrt{5} \approx 0.4472$, then the equation has a unique solution in the circular region.
   If we use the Cauchy ϵ-inequality, we will have

$$\begin{aligned}
&[(R^2 - y^2)v_{yy} - 2xyv_{xy} + (R^2 - x^2)v_{xx}]^2 \\
&\qquad \ge R^4[(\Delta v)^2 - (\sqrt{5}+2)\delta^2(|v_{xx}|^2 + |v_{yy}|^2 + 2|v_{xy}|^2)].
\end{aligned}$$

   Thus if $\delta < \sqrt{1/(\sqrt{5}+2)} \approx 0.4859$, the above conclusion is still true.

5.2. BC (ii): Partially clamped boundary condition. Let $H^2_\Gamma(\Omega)$ be the subspace of $H^2(\Omega)$ such that $f(x,y)|_\Gamma = 0$, $\frac{\partial f}{\partial \nu}|_\Gamma = 0$. Here Γ is part of the boundary of Ω with positive measure.

THEOREM 5.2. *Let Ω be a subdomain of a square region of size d centered at $(0,0)$ in $\mathbb{R}^2$, and assume that $P(x,y)$, $\alpha(x,y)$, and $\beta(x,y)$ are bounded above in Ω. If $\beta(x,y) \ge \alpha(x,y) \ge \alpha_0$ for all $(x,y) \in \Omega$, where $\alpha_0 > 0$ is a constant, then if $d \le R$, the variational equation (5.5) has a unique solution in $H^2_\Gamma(\Omega)$. In addition, if $\alpha(x,y), \beta(x,y) \in C^k(\Omega)$ and $P(x,y) \in H^k(\Omega)$, then for every compact subdomain Ω_1 of Ω, the solution $u \in H^{k+2}(\Omega_1)$.*

Proof. If $\beta(x,y) \ge \alpha(x,y) \ge \alpha_0$, we have

$$B(u,u) \ge \int_\Omega \left[4\alpha H^2_{u_0}(v) - \alpha K_{u_0}(v,v)\right] dx\, dy,$$

where $H_{u_0}(v)$ and $K_{u_0}(v,v)$ are given by (5.6) and (5.7).

Since $R^2(R^2-x^2-y^2) = (R^2-x^2)(R^2-y^2) - x^2y^2$, $K_{u_0}(v,v)$ can be rewritten as

$$\begin{aligned} K_{u_0}(v,v) &= \frac{(R^2-x^2-y^2)^2(2v_{xx}v_{yy}-2v_{xy}^2)}{R^4} \\ &= \frac{2(R^2-x^2-y^2)}{R^6}\left[((R^2-x^2)(R^2-y^2)-x^2y^2)(v_{xx}v_{yy}-v_{xy}^2)\right]. \end{aligned}$$

Thus

$$B(v,v) \geq \int_\Omega \alpha \frac{(R^2-x^2-y^2)}{R^6}(\mathrm{I}+2x^2y^2v_{xy}^2-\mathrm{II})dx\,dy,$$

where

$$\begin{aligned} \mathrm{I} &= (R^2-x^2)^2v_{xx}^2 + (R^2-y^2)^2v_{yy}^2 + 2(R^2-x^2)(R^2-y^2)v_{xy}^2, \\ \mathrm{II} &= 4xy(R^2-y^2)v_{xy}v_{yy} + 4xy(R^2-x^2)v_{xy}v_{xx} - 2x^2y^2v_{xx}v_{yy}. \end{aligned}$$

Assume $|x|, |y| \leq r = \delta R$, where $\delta \in [0,1]$ is to be chosen later. Then

$$\begin{aligned} \mathrm{I} &= (R^2-x^2)^2v_{xx}^2 + (R^2-y^2)^2v_{yy}^2 + 2(R^2-x^2)(R^2-y^2)v_{xy}^2 \\ &\geq (1-\delta^2)^2R^4(|v_{xx}|^2 + |v_{yy}|^2 + 2|v_{xy}|^2), \end{aligned}$$

and

$$\begin{aligned} \mathrm{II} &= 4xy(R^2-y^2)v_{xy}v_{yy} + 4xy(R^2-x^2)v_{xy}v_{xx} - 2x^2y^2v_{xx}v_{yy} \\ &\leq 2\delta^2R^4(|v_{xx}|^2 + |v_{yy}|^2 + 2|v_{xy}|^2) + \delta^4R^4(|v_{xx}|^2 + |v_{yy}|^2) \\ &\leq (2\delta^2+\delta^4)R^4(|v_{xx}|^2 + |v_{yy}|^2 + 2|v_{xy}|^2). \end{aligned}$$

Thus, if $(1-\delta^2)^2 > 2\delta^2+\delta^4$, i.e., if $1-4\delta^2 > 0$, which implies $\delta < 1/2$, then $R^2-x^2-y^2 \geq R^2/2$ and

$$\begin{aligned} B(v,v) &\geq \int_\Omega \alpha \frac{(R^2-x^2-y^2)}{R^2}\left((1-\delta^2)^2 - (2\delta^2+\delta^4)\right)|D^2v|dx\,dy \\ &\geq \int_\Omega \alpha_0 \frac{(1-\delta^2)^2-(2\delta^2+\delta^4)}{2}|D^2v|dx\,dy \\ &= c_0|v|_{2,\Omega}^2, \end{aligned}$$

where $|D^2v|^2 = |v_{xx}|^2 + |v_{yy}|^2 + 2|v_{xy}|^2$ and $c_0 = \alpha_0\frac{(1-\delta^2)^2-(2\delta^2+\delta^4)}{2} > 0$.

By Lemma 4.4,

$$B(u,u) \geq c_1\|v\|_{2,\Omega}^2.$$

Then the Lax–Milgram lemma (Lemma 4.1) implies that the problem has a unique solution. The regularity of the solution follows from the result in Friedrichs [8]. ∎

This means that if the solution is fixed on a part of the boundary and free on the rest of the boundary, then the PDE has a unique solution in $H_\Gamma^2(\Omega)$.

From the above proof, it is clear that, to have a unique solution, δ cannot be arbitrarily large. It is desirable to find δ as large as possible for practical purposes.

Remarks.

1. If we use the Cauchy ϵ-inequality ($ab \le \epsilon a^2 + \frac{b^2}{4\epsilon}$) instead of the Schwarz inequality, we will have

$$\begin{aligned}\mathrm{II} &= 4xy(R^2-y^2)v_{xy}v_{yy} + 4xy(R^2-x^2)v_{xy}v_{xx} - 2x^2y^2v_{xx}v_{yy}\\ &\le \delta^2R^4\left(4\epsilon(|v_{xx}|^2+|v_{yy}|^2) + \frac{2}{\epsilon}|v_{xy}|^2\right) + \delta^4R^4(|v_{xx}|^2+|v_{yy}|^2)\\ &\le \delta^2R^4\left((4\epsilon+\delta^2)(|v_{xx}|^2+|v_{yy}|^2) + \frac{2}{\epsilon}|v_{xy}|^2\right).\end{aligned}$$

Solving $4\epsilon + \delta^2 = 1/\epsilon$, we get $\epsilon = \frac{\sqrt{\delta^4+16}-\delta^2}{8}$. With this choice of ϵ, we have

$$B(v,v) \ge \int_\Omega \alpha(1-2\delta^2)\left((1-\delta^2)^2 - \frac{8\delta^2}{\sqrt{\delta^4+16}-\delta^2}\right)|D^2v|dx\,dy.$$

By solving $(1-\delta^2)^2 = 8\delta^2/(\sqrt{\delta^4+16}-\delta^2)$, we derive $1-4\delta^2+\delta^4-2\delta^6 = 0$, and with the help of Mathematica we get

$$\delta = \sqrt{\frac{1}{6} - \frac{23}{6(19+12\sqrt{23})^{1/3}} + \frac{1}{6}(19+12\sqrt{23})^{1/3}} \approx 0.507992.$$

Thus if $r < 0.50799\,R$, then the conclusion of the above theorem is still true.

2. If we solve the equation on a circular region $x^2+y^2 \le r^2$ instead of a square region and assume $r = \delta\,R$, then with the Schwarz inequality, we will have

$$\begin{aligned}\mathrm{II} &= 4xy(R^2-y^2)v_{xy}v_{yy} + 4xy(R^2-x^2)v_{xy}v_{xx} - 2x^2y^2v_{xx}v_{yy}\\ &\le \delta^2R^4(|v_{xx}|^2+|v_{yy}|^2+2|v_{xy}|^2) + \frac{\delta^4}{4}R^4(|v_{xx}|^2+|v_{yy}|^2)\\ &\le \delta^2\left(1+\frac{\delta^4}{4}\right)R^4(|v_{xx}|^2+|v_{yy}|^2+2|v_{xy}|^2).\end{aligned}$$

Solving $(1-\delta^2)^2 > \delta^2(1+\frac{\delta^4}{4})$ gives $\delta < \sqrt{\frac{3-\sqrt{6}}{2}} \approx 0.5246$.
If we use the Cauchy ϵ-inequality, we will have

$$\begin{aligned}\mathrm{II} &= 4xy(R^2-y^2)v_{xy}v_{yy} + 4xy(R^2-x^2)v_{xy}v_{xx} - 2x^2y^2v_{xx}v_{yy}\\ &\le 2\delta^2R^4(|v_{xx}||v_{xy}| + |v_{yy}||v_{xy}|) + \frac{\delta^4R^4}{4}(|v_{xx}|^2+|v_{yy}|^2)\\ &\le 2\delta^2R^4\left(\epsilon(|v_{xx}|^2+|v_{yy}|^2) + 2\frac{|v_{xy}|^2}{4\epsilon}\right) + \frac{\delta^4R^4}{4}(|v_{xx}|^2+|v_{yy}|^2)\\ &= R^4\delta^2\left[\left(2\epsilon+\frac{\delta^2}{4}\right)(|v_{xx}|^2+|v_{yy}|^2) + 2\frac{|v_{xy}|^2}{2\epsilon}\right].\end{aligned}$$

Let $2\epsilon + \frac{\delta^2}{4} = \frac{1}{2\epsilon}$; we have $\epsilon = (-\delta^2 + \sqrt{\delta^4+64})/8$. Solving $(1-\delta^2)^2 > \delta^2(\delta^2+\sqrt{\delta^4+64})/8$ via Mathematica, we get $\delta^2 \approx 0.37399$, which gives $\delta \approx 0.611549$.

5.3. BC (iii): Natural boundary condition. For the same reason that the variational equations involve only the second derivatives of u, the analysis of the solution for the natural boundary condition is almost the same as for the linearization about a flat surface. Therefore, we state the following theorem but omit its proof.

THEOREM 5.3. *Let Ω be a subdomain of a square region of size d centered at $(0,0)$ in $\mathbb{R}^2$, and assume that $P(x,y)$, $\alpha(x,y)$, and $\beta(x,y)$ are bounded above in Ω. If $\beta(x,y) \geq \alpha(x,y) \geq \alpha_0$ for all $(x,y) \in \Omega$, where $\alpha_0 > 0$ is a constant, then if $d \leq R$, the solution to the variational equation (5.5) exists in $H^2(\Omega)$ and is unique up to an addition of a linear function. In addition, if $\alpha(x,y), \beta(x,y) \in C^k(\Omega)$ and $P(x,y) \in H^k(\Omega)$, then for every compact subdomain Ω_1 of Ω, the solution $u \in H^{k+2}(\Omega_1)$.*

6. Numerical results. In practice, lenses are designed on a circular region of diameter 80mm. For computational simplicity, we solve the design problem in the square domain $[-40,40] \otimes [-40,40]$.

We considered both of the linearizations analyzed above and solved the problems for all three types of boundary conditions by finite element methods. We found that linearization about a spherical surface with natural boundary conditions produces the best lens. We will describe the findings below. Detailed analysis of the numerical methods and description of the setting of the weight functions as well as the prescribed power functions which produced the result shown can be found in [11].

Figure 6.1 shows the results where the radius of the spherical surface $R = 107$mm, the index of the material $n = 1.53$, and the add-power (power addition from the base power) is 2 diopters (diopter $= (n-1)/($focal length$)$) for the near-vision area. The progressive lens surface is shown in Figure 6.1(a), while Figure 6.1(b) shows the difference between the progressive lens surface and the initial spherical surface.

To optically analyze the surface, we need to calculate the power and the cylinder of the surface. However, since the surface produced by the method is piecewise quadratic, the second derivatives at the grid points are not well defined. Thus it is not possible to calculate the cylinder and the power of the surface directly. An efficient and popular method in the lens industry for analyzing lenses is to fit a lens by tensor product B-splines of degree 3 to 5 and compute the cylinder and the power of the interpolated surface. This gives a very good approximation of the cylinder and the power of the true surface. We used the same method with tensor product B-splines of degree 5 for analyzing our lenses.

Figure 6.2 shows the power and cylinder contours of the surface given in Figure 6.1(a).

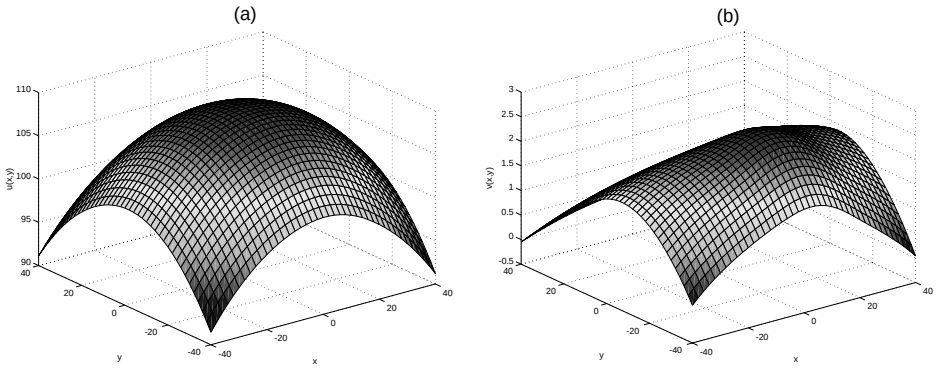


FIG. 6.1. *Numerical results:* (a) *progressive surface,* (b) *addition to the spherical surface.*

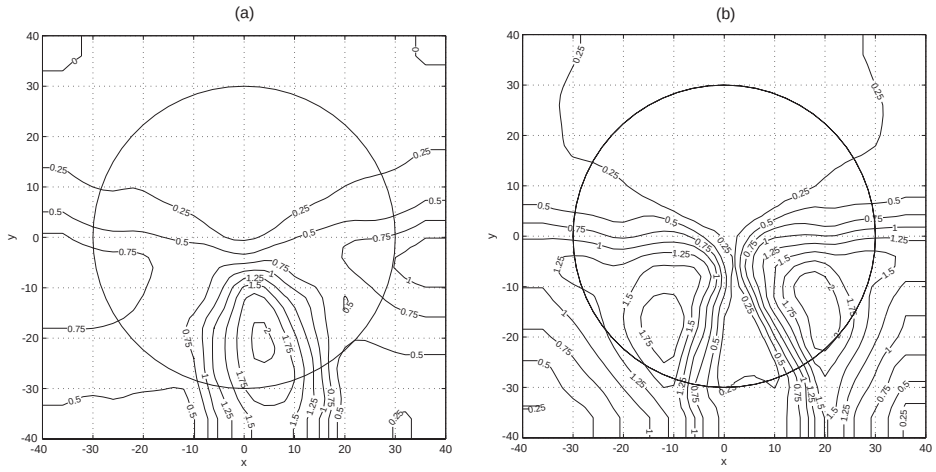


FIG. 6.2. *Numerical results with the natural boundary condition.* (a) *power contour,* (b) *cylinder contour.*

The circles in Figures 6.2(a)–(b) indicate the area of radius < 30mm that is actually used in making prescription lenses. We can see that the add-power of 2 diopters is achieved and that both the power and the cylinder change smoothly through out the lens. Moreover, there is a clear corridor connecting the far and near regions in the cylinder. Note also that the maximum cylinder is about the same as the add-power, which is typical in progressive lenses.

7. Discussions. In this first of a two-part paper, we have analyzed the variational problem of progressive lens design. The design problem consists of finding a surface whose power distribution over a lens is close to a desired distribution, while minimizing the astigmatism over a specific region. This is a nonlinear optimization problem. We considered linearizations about a plane and a sphere. The resulting equations in each case are fourth-order PDEs similar to those that describe deformations of plates. We considered the existence and uniqueness of solutions under various boundary conditions. For completeness, a numerical example is included, which demonstrates the effectiveness of the methods.

In the second part of our work [11], we will address in detail the computational issues arising in this design problem. The results from the present work will serve to guide the design of accurate and convergent numerical methods for solving the progressive lens design problem.

Acknowledgments. The authors express their gratitude to Dr. Jianping Wei for helpful discussions on this work. We are grateful to the anonymous referees for their careful reading of the manuscript and for pointing out references and patents relevant to this work.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1978.
[2] P. ALLIONE, F. AHSBAHS, AND G. L. SAUX, *Application of optimization in computer-aided ophthalmic lens design*, in Design and Engineering of Optical System II, F. Merkle, ed., Proceeding of SPIE, Vol. 3737, 1999, pp. 138–148.

[3] M. P. Do Carmo, *Differential Geometry of Curves and Surfaces*, Prentice-Hall, Englewood Cliffs, NJ, 1976.
[4] T. Baudart, F. Ahsbahs, and C. Miege, *Multifocal Ophthalmic Lens*, United States patent 6,102544, 2000.
[5] Ph. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[6] Ph. G. Ciarlet, *Mathematical Elasticity*, Vol. I, North-Holland, Amsterdam, 1997.
[7] Ph. G. Ciarlet, *Mathematical Elasticity*, Vol. II, North-Holland, Amsterdam, 1997.
[8] K. O. Friedrichs, *On the differentiability of the solutions of linear elliptic differential equations*, Comm. Pure Appl. Math., 6 (1953), pp. 299–325.
[9] J. Loos, G. Greiner, and H. P. Seidel, *A variational approach to progressive lens design*, Comput. Aided Design, 30 (1998), pp. 595–602.
[10] J. C. C. Nitsche, *Boundary value problems for variational integrals involving surface curvatures*, Quart. Appl. Math., 51 (1993), pp. 363–387.
[11] J. Wang and F. Santosa, *Finite element methods for progressive lens design*, Math. Models Methods Appl. Sci., to appear.
[12] J. T. Winthrop, *Progressive Addition Spectacle Lens*, United States patent 4,861,153, 1989.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 297–308

ACOUSTIC SCATTERING BY INHOMOGENEOUS OBSTACLES*

P. A. MARTIN†

Abstract. Acoustic scattering problems are considered when the material parameters (density and speed of sound) are functions of position within a bounded region. An integro-differential equation for the pressure in this region is obtained. It is proved that solving this equation is equivalent to solving the scattering problem. Problems of this kind are often solved by regarding the effects of the inhomogeneity as an unknown source term driving a Helmholtz equation, leading to an equation of Lippmann–Schwinger type. It is shown that this approach is incomplete when the density is discontinuous. Analogous scattering problems for elastic waves and for electromagnetic waves are also discussed briefly.

Key words. Bergmann's equation, Lippmann–Schwinger equation

AMS subject classifications. 76Q05, 35J05, 35J25, 45B05

DOI. 10.1137/S0036139902414379

1. Introduction.

Time-harmonic acoustic waves in an inhomogeneous compressible fluid can be modelled using Bergmann's equation (see (2.3) below). If the waves are generated by a point source located at $\boldsymbol{r}'$, the pressure at $\boldsymbol{r}$, $\mathcal{G}(\boldsymbol{r};\boldsymbol{r}')$, satisfies

$$\nabla^2\mathcal{G}(\boldsymbol{r};\boldsymbol{r}') - \rho^{-1}(\operatorname{grad}\rho)\cdot\operatorname{grad}\mathcal{G}(\boldsymbol{r};\boldsymbol{r}') + k^2(\boldsymbol{r})\,\mathcal{G}(\boldsymbol{r};\boldsymbol{r}') = \delta(\boldsymbol{r}-\boldsymbol{r}'), \tag{1.1}$$

where $k^2(\boldsymbol{r}) = [\omega/c(\boldsymbol{r})]^2$, ω is the frequency, $c(\boldsymbol{r})$ is the speed of sound, and $\rho(\boldsymbol{r})$ is the density. ($\mathcal{G}$ is an exact Green's function for the problem.) Equation (1.1) is supposed to hold everywhere in space, and it is to be solved subject to a radiation condition at infinity.

How does one solve (1.1)? According to a recent review article by Tourin, Fink, and Derode [39], "the solution of (1.1) can be written as"

$$\mathcal{G}(\boldsymbol{r};\boldsymbol{r}') = G_{\mathrm{e}}(\boldsymbol{r};\boldsymbol{r}') + \int G_{\mathrm{e}}(\boldsymbol{r};\boldsymbol{r}_1)\,V(\boldsymbol{r}_1)\mathcal{G}(\boldsymbol{r}_1;\boldsymbol{r}')\,d\boldsymbol{r}_1, \tag{1.2}$$

where the integration is over all of space, V is a "potential operator", defined by

$$V(\boldsymbol{r}) = k_{\mathrm{e}}^2 - k^2(\boldsymbol{r}) + \rho^{-1}(\operatorname{grad}\rho(\boldsymbol{r}))\cdot\operatorname{grad}, \tag{1.3}$$

k_{e} is the wavenumber for a related homogeneous medium, and G_{e} is the (known) solution of the problem for that medium: G_{e} satisfies

$$\nabla^2 G_{\mathrm{e}}(\boldsymbol{r};\boldsymbol{r}_1) + k_{\mathrm{e}}^2 G_{\mathrm{e}}(\boldsymbol{r};\boldsymbol{r}_1) = \delta(\boldsymbol{r}-\boldsymbol{r}_1) \tag{1.4}$$

and the radiation condition and is given explicitly by (3.1) below. Equation (1.2) is not derived in [39] and no indication of its range of validity is given. In fact, as we shall see, (1.2) is not valid when $\rho(\boldsymbol{r})$ is discontinuous. (This is unfortunate, because most of the applications in [39] are to arrays of discrete scatterers, such as steel rods in water.)

*Received by the editors September 10, 2002; accepted for publication (in revised form) May 13, 2003; published electronically November 19, 2003.
http://www.siam.org/journals/siap/64-1/41437.html
†Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401-1887 (pamartin@mines.edu).

1.1. A formal derivation. Equations such as (1.1) are often treated by moving all the complicated terms to the right-hand side where they are regarded as a forcing term. Thus, write (1.1) as

$$\nabla^2\mathcal{G}(\boldsymbol{r};\boldsymbol{r}') + k_{\mathrm{e}}^2\,\mathcal{G}(\boldsymbol{r};\boldsymbol{r}') = V(\boldsymbol{r})\mathcal{G}(\boldsymbol{r};\boldsymbol{r}') + \delta(\boldsymbol{r}-\boldsymbol{r}').$$

Equation (1.2) then follows by noting that

$$u(\boldsymbol{r}) = \int G_{\mathrm{e}}(\boldsymbol{r};\boldsymbol{r}')\,f(\boldsymbol{r}')\,d\boldsymbol{r}' \quad \text{solves} \quad (\nabla^2 + k_{\mathrm{e}}^2)u = f. \tag{1.5}$$

Formal derivations of this kind are often found in textbooks; see, for example, [9, sect. 8.9.1] or [18, eqn. (21.37)]. The result (1.5) can be justified readily if one assumes that f is (Hölder) continuous. However, for discrete scatterers, there will be interfaces across which $k(\boldsymbol{r})$ and the normal derivative of $\mathcal{G}$, $\partial\mathcal{G}/\partial n$, will be discontinuous (although $\mathcal{G}$ and $\rho^{-1}\partial\mathcal{G}/\partial n$ are both continuous across such interfaces).

1.2. The present paper. The formal derivation in section 1.1 is incomplete. It can be repaired so as to give the correct result. Thus, when considered as a distribution, we have

$$V\mathcal{G} = \{V\mathcal{G}\} + [\rho]\,\delta(S)\,(\rho^{-1}\partial\mathcal{G}/\partial n),$$

where $\{V\mathcal{G}\}$ denotes the value of $V\mathcal{G}$ anywhere but on the interfaces S and $[\rho]$ denotes the discontinuity in ρ across S [41, sect. 1.13]. Then, (1.5) suggests an additional term on the right-hand side of (1.2), namely

$$\int_S G_{\mathrm{e}}(\boldsymbol{r};\boldsymbol{r}_s)\,[\rho](\boldsymbol{r}_s)\left(\frac{1}{\rho}\frac{\partial\mathcal{G}}{\partial n}\right)(\boldsymbol{r}_s;\boldsymbol{r}')\,ds(\boldsymbol{r}_s). \tag{1.6}$$

In this paper, we shall derive an equation, similar to (1.2), that respects the proper transmission conditions across interfaces: we do not use distribution theory. We prove that solving this equation is equivalent to solving the transmission problem for the acoustic pressure (Theorem 3.1).

We are mainly concerned with the following problem: acoustic scattering by a bounded inhomogeneity embedded in an unbounded homogeneous medium. The density and sound-speed are assumed to be functions of position within the inhomogeneity, and they can be discontinuous across the interface between the inhomogeneity and the surrounding fluid. Problems in which the inhomogeneity is *spherically symmetric*, so that ρ and c are assumed to be given functions of the spherical polar coordinate r (only), have been studied by several authors; see [28] for references.

The new equation is derived in section 3. It reduces to the well-known Lippmann–Schwinger equation when the density in the inhomogeneity is constant and equal to the density of the surrounding homogeneous fluid. It also reduces to the equation derived formally above, namely (1.2), *but only when there is no discontinuity in the density across the boundary of the inhomogeneity.* If there is such a discontinuity (as is typical in applications), an extra term is needed; see (1.6) and (4.4) below. Analogous scattering problems for electromagnetic waves and for elastic waves are discussed briefly in sections 4.4 and 4.5, respectively.

2. Formulation. Consider the scattering of time-harmonic sound waves in a homogeneous compressible fluid by an inhomogeneous obstacle. In the exterior fluid, B_{e}, we can write

$$p_{\mathrm{e}} = p_{\mathrm{inc}} + p_{\mathrm{sc}}, \tag{2.1}$$

where p_e is the (total) acoustic pressure, p_{inc} is the given incident field, and p_{sc} is the scattered field. The governing equation for p_{sc} is

$$(\nabla^2 + k_e^2)p_{sc} = 0 \quad \text{in } B_e, \tag{2.2}$$

where $k_e = \omega/c_e$ is the wave number (assumed to be real and positive), ω is the frequency, and c_e is the constant speed of sound. We assume that the incident field p_{inc} satisfies (2.2) everywhere, except possibly at some places in B_e (so that p_{inc} could correspond to a point source in B_e, for example). We require that p_{sc} satisfies the Sommerfeld radiation condition at infinity.

Within the obstacle, B, the governing equation is Bergmann's equation ([3, 28] and [29, p. 408])

$$\rho_i \operatorname{div}\left(\rho_i^{-1} \operatorname{grad} p_i\right) + k_i^2 p_i = 0 \quad \text{in } B, \tag{2.3}$$

where p_i is the pressure and $k_i = \omega/c_i$. The interior density ρ_i and speed of sound c_i can vary with position in B. At the interface S between B and B_e, we have a pair of transmission conditions, expressing continuity of pressure and normal velocity. These are

$$p_e = p_i \quad \text{and} \quad \frac{1}{\rho_e}\frac{\partial p_e}{\partial n} = \frac{1}{\rho_i}\frac{\partial p_i}{\partial n} \quad \text{on } S, \tag{2.4}$$

where ρ_e is the (constant) density of the fluid in B_e.

Summarizing, we have the following problem to solve.

Scattering Problem. Let p_{inc} be a given incident field. Find a pair of functions, $\{p_e, p_i\}$, where $p_{sc} = p_e - p_{inc}$ satisfies (2.2) and the Sommerfeld radiation condition, p_i satisfies (2.3), and p_e and p_i satisfy the transmission conditions (2.4) across the interface S.

Werner wrote an important paper on the Scattering Problem in 1963 [43]. He reduced the problem to a system of coupled integral equations, using single-layer, double-layer, and volume potentials; see Appendix A. Werner's approach is an example of an *indirect method*, meaning that the unknown quantities do not have any physical relevance. He proved that the Scattering Problem has exactly one solution; his uniqueness result is in [42]. However, as far as we know, his system of integral equations has not been used in computations.

We shall use a *direct method*, meaning that the unknown quantity is recognized as a physical variable, namely p_i. Moreover, we shall use only volume potentials; this is convenient from a computational point of view, because one does not have to approximate a mixture of surface and volume contributions.

In solving the Scattering Problem, we seek classical solutions. We shall appeal to Werner's existence result, so we suppose that $p_{sc} \in C^2(B_e) \cap C^1(\overline{B_e})$ and $p_i \in C^2(B) \cap C^1(\overline{B})$. We assume that $\rho_i \in C^2(\overline{B})$ and $c_i \in C^1(\overline{B})$ are both positive. Finally, we assume that S is smooth (C^2). It is likely that these conditions can be weakened; for example, Werner's uniqueness theorem [42] requires that ρ_i and c_i be in Hölder spaces, with $\rho_i \in C^{1,\alpha}(\overline{B})$ and $c_i \in C^{0,\alpha}(\overline{B})$.

Bergmann's equation (2.3) can be written in other ways. One alternative is

$$\nabla^2 p_i + \rho_i \left(\operatorname{grad} \rho_i^{-1}\right) \cdot \operatorname{grad} p_i + k_e^2 N p_i = 0 \quad \text{in } B, \tag{2.5}$$

where $N = (k_i/k_e)^2 = (c_e/c_i)^2$ is the (square of the) *refractive index*. Another is

$$\nabla^2 p_i + k_e^2 p_i = V p_i \quad \text{in } B,$$

where (see (1.3))

$$Vu = k_{\mathrm{e}}^2(1-N)u + \rho_{\mathrm{i}}^{-1}(\operatorname{grad}\rho_{\mathrm{i}})\cdot\operatorname{grad} u. \tag{2.6}$$

Bergmann's equation can also be reduced to an equation without first derivatives by introducing a new dependent variable [3], $u = p_{\mathrm{i}}\rho_{\mathrm{i}}^{-1/2}$: u is found to satisfy

$$\nabla^2 u + (k_{\mathrm{i}}^2 + K)u = 0, \tag{2.7}$$

where

$$K = \tfrac{1}{2}\rho_{\mathrm{i}}^{-1}\nabla^2\rho_{\mathrm{i}} - \tfrac{3}{4}\rho_{\mathrm{i}}^{-2}\left|\operatorname{grad}\rho_{\mathrm{i}}\right|^2 \tag{2.8}$$

$$= -\rho_{\mathrm{i}}^{1/2}\nabla^2\left(\rho_{\mathrm{i}}^{-1/2}\right). \tag{2.9}$$

Equations (2.7) and (2.8) (but not (2.9)) can be found in [4, p. 171]. Equation (2.7) can also be written as *Schrödinger's equation* [32, eqn. (10.59)].

Much has been written on the case where ρ_{i} is constant, so that the second term on the left-hand side of (2.5) can be deleted [12, Chap. 8]; see [23] for a review of available point-source solutions (Green's functions) for various functional forms of N. Here, we do not make this assumption: we allow both ρ_{i} and N to vary with position. Note that if the material in B is actually homogeneous, so that ρ_{i} and k_{i} are both constants, boundary integral equations over S can be used; see [20] for a review.

Kriegsmann and Reiss [22] have given long-wave approximations to the solution of the Scattering Problem, assuming that $\rho_{\mathrm{i}}/\rho_{\mathrm{e}} \simeq 1$. Specifically, they take $\rho_{\mathrm{i}}/\rho_{\mathrm{e}} = 1 + O(\varepsilon^2)$ and $k_{\mathrm{e}}a = O(\varepsilon)$, where a is the diameter of B and $0 < \varepsilon \ll 1$.

Colton and Monk [13] have reviewed progress with inverse problems for (2.3), which they write as $\operatorname{div}\left(\rho_{\mathrm{i}}^{-1}\operatorname{grad} p_{\mathrm{i}}\right) + k_{\mathrm{e}}^2\tilde{N}p_{\mathrm{i}} = 0$, where $\tilde{N} = N/\rho_{\mathrm{i}}$. They also assume that ρ_{i} is constant near S so that the ratio $\rho_{\mathrm{e}}/\rho_{\mathrm{i}}$ occurring in $(2.4)_2$ is constant; in general, this need not be true.

2.1. Uniqueness. The Scattering Problem has at most one solution [42]. This uniqueness theorem can be proved as follows. Set $p_{\mathrm{inc}} \equiv 0$ and then apply Green's theorem to p_{e} and its complex conjugate, $\overline{p_{\mathrm{e}}}$, in the exterior, giving

$$2ik_{\mathrm{e}}\lim_{R\to\infty}\int_{S_R}|p_{\mathrm{e}}|^2\,ds + \int_S\left(p_{\mathrm{e}}\,\frac{\partial\overline{p_{\mathrm{e}}}}{\partial n} - \overline{p_{\mathrm{e}}}\,\frac{\partial p_{\mathrm{e}}}{\partial n}\right)ds = 0. \tag{2.10}$$

Here, the unit normal to S, $\boldsymbol{n}$, points out of B, S_R is a large sphere of radius R that encloses S, we have used the radiation condition, and we have assumed that k_{e} is real. Next, apply the divergence theorem in B to the vector field $(\overline{p_{\mathrm{i}}}/\rho_{\mathrm{i}})\operatorname{grad} p_{\mathrm{i}}$, giving

$$\int_B\left(|\operatorname{grad} p_{\mathrm{i}}|^2 - k_{\mathrm{i}}^2|p_{\mathrm{i}}|^2\right)\frac{dV}{\rho_{\mathrm{i}}} = \int_S\frac{\overline{p_{\mathrm{i}}}}{\rho_{\mathrm{i}}}\,\frac{\partial p_{\mathrm{i}}}{\partial n}\,ds,$$

where we have used (2.3). The imaginary part of this equation gives

$$\int_S\left(p_{\mathrm{i}}\,\frac{\partial\overline{p_{\mathrm{i}}}}{\partial n} - \overline{p_{\mathrm{i}}}\,\frac{\partial p_{\mathrm{i}}}{\partial n}\right)\frac{ds}{\rho_{\mathrm{i}}} = \int_B\left(k_{\mathrm{i}}^2 - \overline{k_{\mathrm{i}}^2}\right)|p_{\mathrm{i}}|^2\,\frac{dV}{\rho_{\mathrm{i}}}.$$

Then, making use of the transmission conditions (2.4), (2.10) gives

$$k_{\mathrm{e}}\lim_{R\to\infty}\int_{S_R}|p_{\mathrm{e}}|^2\,ds + 2\int_B\operatorname{Re}(k_{\mathrm{i}})\operatorname{Im}(k_{\mathrm{i}})\,|p_{\mathrm{i}}|^2\,\frac{\rho_{\mathrm{e}}}{\rho_{\mathrm{i}}}\,dV = 0.$$

Rellich's lemma [11, Lem. 3.11] then implies that $p_\mathrm{e} \equiv 0$ in B_e, provided that

$$\mathrm{Re}\,(k_\mathrm{i})\,\mathrm{Im}\,(k_\mathrm{i}) \geq 0.$$

The transmission conditions then imply that $p_\mathrm{i} = 0$ and $\partial p_\mathrm{i}/\partial n = 0$ on S. Thus, p_i solves the Cauchy problem for the elliptic partial differential equation (2.3), in which the (positive) coefficients ρ_i^{-1} and $k_\mathrm{i}^2/\rho_\mathrm{i}$ are C^2 and C^1, respectively. It follows that $p_\mathrm{i} \equiv 0$ in B, as required. Here, we have used a unique continuation result due to Müller [30] and Aronszajn [2]; see [21] for a brief review.

3. An integral representation and an integro-differential equation. We shall consider integral representations obtained using the free-space Green's function for the exterior fluid,

$$G_\mathrm{e}(P,Q) = -\exp(ik_\mathrm{e}R)/(4\pi R), \tag{3.1}$$

where P and Q are typical points in three-dimensional space and $R = |\boldsymbol{r}_P - \boldsymbol{r}_Q|$ is the distance between P and Q.

An application of Green's second theorem in B_e to p_sc and G_e gives

$$\int_S \left\{ G_\mathrm{e}(P,q)\,\frac{\partial p_\mathrm{sc}}{\partial n_q} - p_\mathrm{sc}(q)\,\frac{\partial}{\partial n_q} G_\mathrm{e}(P,q) \right\} ds_q = \begin{cases} p_\mathrm{sc}(P), & P \in B_\mathrm{e}, \\ 0, & P \in B. \end{cases}$$

A similar application in B to p_inc and G_e gives

$$\int_S \left\{ G_\mathrm{e}(P,q)\,\frac{\partial p_\mathrm{inc}}{\partial n_q} - p_\mathrm{inc}(q)\,\frac{\partial}{\partial n_q} G_\mathrm{e}(P,q) \right\} ds_q = \begin{cases} 0, & P \in B_\mathrm{e}, \\ -p_\mathrm{inc}(P), & P \in B. \end{cases}$$

Adding these gives

$$\int_S \left\{ G_\mathrm{e}(P,q)\,\frac{\rho_\mathrm{e}}{\rho_\mathrm{i}}\frac{\partial p_\mathrm{i}}{\partial n_q} - p_\mathrm{i}(q)\,\frac{\partial}{\partial n_q} G_\mathrm{e}(P,q) \right\} ds_q = \begin{cases} p_\mathrm{sc}(P), & P \in B_\mathrm{e}, \\ -p_\mathrm{inc}(P), & P \in B, \end{cases} \tag{3.2}$$

where we have used (2.1) and the transmission conditions (2.4). The first of these gives an integral representation for $p_\mathrm{sc}(P)$ in terms of a distribution of sources and dipoles over S. Such representations are common in scattering theory. However, it is not very convenient here because we do not know p_i or $\partial p_\mathrm{i}/\partial n$ on S.

To make progress, recall Green's first theorem,

$$\int_B \left\{ \phi\,\nabla^2\psi + (\mathrm{grad}\,\phi)\cdot(\mathrm{grad}\,\psi) \right\} dV = \int_S \phi\,\frac{\partial\psi}{\partial n}\,ds,$$

where ϕ and ψ are sufficiently smooth in B. Choose $\phi(Q) = p_\mathrm{i}(Q)$ and $\psi(Q) = G_\mathrm{e}(P,Q)$ with $P \in B_\mathrm{e}$, whence

$$\begin{aligned} &\int_S p_\mathrm{i}(q)\,\frac{\partial}{\partial n_q} G_\mathrm{e}(P,q)\,ds_q \\ &\qquad = \int_B \left\{ (\mathrm{grad}\,p_\mathrm{i})\cdot(\mathrm{grad}_Q\,G_\mathrm{e}) - k_\mathrm{e}^2 p_\mathrm{i}(Q)\,G_\mathrm{e}(P,Q) \right\} dV_Q, \end{aligned} \tag{3.3}$$

where we have used $(\nabla^2 + k_\mathrm{e}^2)G_\mathrm{e}(P,Q) = 0$ for $P \neq Q$. Similarly, if we choose $\psi(Q) = p_\mathrm{i}(Q)$ and $\phi(Q) = (\rho_\mathrm{e}/\rho_\mathrm{i})G_\mathrm{e}(P,Q)$ with $P \in B_\mathrm{e}$, we obtain

$$\begin{aligned} &\int_S \frac{\rho_\mathrm{e}}{\rho_\mathrm{i}}\,\frac{\partial p_\mathrm{i}}{\partial n_q}\,G_\mathrm{e}(P,q)\,ds_q \\ &\qquad = \int_B \frac{\rho_\mathrm{e}}{\rho_\mathrm{i}} \left\{ (\mathrm{grad}\,p_\mathrm{i})\cdot(\mathrm{grad}_Q\,G_\mathrm{e}) - k_\mathrm{e}^2\,N(Q)\,p_\mathrm{i}(Q)\,G_\mathrm{e}(P,Q) \right\} dV_Q, \end{aligned} \tag{3.4}$$

where we have used (2.5). Subtracting (3.3) from (3.4) gives the left-hand side of (3.2) for $P \in B_{\rm e}$, whence $p_{\rm sc}(P) = (\mathcal{L}p_{\rm i})(P)$ for $P \in B_{\rm e}$, where

$$(\mathcal{L}v)(P) = \int_B \big\{(\alpha(Q)-1)\,(\operatorname{grad} v)\cdot(\operatorname{grad}_Q G_{\rm e}(P,Q)) + (1-N\alpha)\,k_{\rm e}^2\,v(Q)\,G_{\rm e}(P,Q)\big\}\,dV_Q \tag{3.5}$$

and

$$\alpha(P) = \rho_{\rm e}/\rho_{\rm i}(P).$$

We repeat the calculations for $P \in B$, having excised a small sphere centered at P. The singularity at $P = Q$ has no effect on (3.4) but it causes $-p_{\rm i}(P)$ to be added to the left-hand side of (3.3). Then, (3.2) for $P \in B$ becomes

$$-p_{\rm inc}(P) = -p_{\rm i}(P) + (\mathcal{L}p_{\rm i})(P), \quad P \in B.$$

At this stage, we have proved one half of the following theorem.

THEOREM 3.1. *Let the pair $\{p_{\rm e}, p_{\rm i}\}$ solve the Scattering Problem. Then $v(P) \equiv p_{\rm i}(P) \in C^2(B)$ solves*

$$v(P) - (\mathcal{L}v)(P) = p_{\rm inc}(P), \quad P \in B, \tag{3.6}$$

where $\mathcal{L}v$ is defined by (3.5). *Conversely, let v solve* (3.6). *Then the pair $\{p_{\rm e}, p_{\rm i}\}$, defined by*

$$p_{\rm e}(P) = p_{\rm inc}(P) + (\mathcal{L}v)(P) \quad \textit{for } P \in B_{\rm e} \tag{3.7}$$

and $p_{\rm i}(P) = v(P)$ for $P \in B$ solves the Scattering Problem.

Proof. We have to prove the second half of the theorem. From (3.7), we define $p_{\rm sc}$ using

$$p_{\rm sc}(P) = (\mathcal{L}v)(P), \quad P \in B_{\rm e}; \tag{3.8}$$

evidently, $p_{\rm sc}$ satisfies (2.2) and the Sommerfeld radiation condition, as it inherits these properties from $G_{\rm e}$.

Next, let us show that $p_{\rm i} \equiv v$ satisfies (2.3). As $p_{\rm inc}$ satisfies (2.2) in B, (3.6) gives

$$(\nabla^2 + k_{\rm e}^2)(v - \mathcal{L}v) = 0 \quad \text{in } B. \tag{3.9}$$

Now, from the definition (3.5), we have

$$(\mathcal{L}v)(P) = -\frac{\partial}{\partial x_j^P}\int_B (\alpha-1)\,\frac{\partial v}{\partial x_j^Q}\,G_{\rm e}(P,Q)\,dV_Q + k_{\rm e}^2\int_B (1-N\alpha)\,v(Q)\,G_{\rm e}(P,Q)\,dV_Q, \tag{3.10}$$

where $P \equiv (x_1^P, x_2^P, x_3^P)$, $Q \equiv (x_1^Q, x_2^Q, x_3^Q)$, and summation over j is implied. The second integral in (3.10) is an acoustic volume potential, and the first term is the sum of three first derivatives of volume potentials. The properties of volume potentials are

summarized in Appendix B. In particular, the result of applying $(\nabla^2 + k_e^2)$ is given by (B.1), so that we obtain

$$(\nabla^2 + k_e^2)(\mathcal{L}v) = -\frac{\partial}{\partial x_j^P}\left\{(\alpha - 1)\,\frac{\partial v}{\partial x_j^P}\right\} + k_e^2\,(1 - N\alpha)\,v(P)$$
$$= (\nabla^2 + k_e^2)v - \rho_e \operatorname{div}\left(\rho_i^{-1}\operatorname{grad} v\right) - k_i^2(\rho_e/\rho_i)v, \quad P \in B,$$

whence (3.9) gives the desired result.

To verify the transmission conditions, observe that (3.6) gives

$$p_i(P) - p_{inc}(P) = (\mathcal{L}v)(P), \quad P \in B. \tag{3.11}$$

However, as $\mathcal{L}v$ comprises a volume potential and first derivatives of volume potentials, it follows that $(\mathcal{L}v)(P)$ is continuous as P crosses S (see Appendix B). Thus, (3.8) and (3.11) show that the first transmission condition, $(2.4)_1$, is satisfied.

For the second transmission condition, we take the normal derivative of (3.8) and (3.11) to give

$$\frac{\partial}{\partial n}\left\{p_{sc} - (p_i - p_{inc})\right\} = \left[\frac{\partial}{\partial n}\,\mathcal{L}v\right] \quad \text{on } S, \tag{3.12}$$

where $[f]$ is the discontinuity in f across S, defined by

$$[f(p)] = \lim_{P_e \to p} f(P_e) - \lim_{P \to p} f(P), \quad P_e \in B_e, \quad P \in B, \quad p \in S. \tag{3.13}$$

It is shown in Appendix B that

$$\left[\frac{\partial}{\partial n}\,\mathcal{L}v\right] = \left(\frac{\rho_e}{\rho_i} - 1\right)\frac{\partial v}{\partial n}, \tag{3.14}$$

and then (3.12) and $v \equiv p_i$ imply that $(2.4)_2$ is satisfied. This completes the proof of Theorem 3.1. □

4. Discussion of the integro-differential equation (3.6).

4.1. Solvability. We have seen that solving the Scattering Problem is equivalent to solving (3.6), which is an integro-differential equation for $v(P)$, $P \in B$. This equation is uniquely solvable. To see this, we appeal to Werner's existence result [43]: the solution $\{p_e, p_i\}$ of the Scattering Problem exists and, by the first half of Theorem 3.1, p_i solves (3.6). For uniqueness, suppose that $v_0(P)$ solves (3.6) with $p_{inc} \equiv 0$. Construct $p_e = (\mathcal{L}v_0)(P)$ for $P \in B_e$ and $p_i = v_0(P)$ for $P \in B$. By the second half of Theorem 3.1, these fields solve the homogeneous Scattering Problem; they must vanish identically by the uniqueness theorem for the Scattering Problem (section 2.1). In particular, $v_0(P) \equiv 0$ for $P \in B$, as required.

We note that an integro-differential equation equivalent to (3.6) was derived by Gerjuoy and Saxon [15] in 1954. In fact, they derived a coupled system, involving the pressure and the velocity, which they regarded as preferable to a single equation for the pressure as they were motivated by a desire to obtain variational principles.

4.2. The Lippmann–Schwinger equation. As a special case of the Scattering Problem, suppose that $\rho_i(Q) = \rho_e$ for all $Q \in B$, so that the density of the scatterer is

the same as that of the surrounding homogeneous fluid. Then, the integro-differential equation (3.6) reduces to the integral equation

$$v(P) - k_{\mathrm{e}}^2 \int_B \{1 - N(Q)\} v(Q)\, G_{\mathrm{e}}(P,Q)\, dV_Q = p_{\mathrm{inc}}(P), \quad P \in B, \tag{4.1}$$

where $N(Q) = (k_{\mathrm{i}}/k_{\mathrm{e}})^2 = \{c_{\mathrm{e}}/c_{\mathrm{i}}(Q)\}^2$. This integral equation and its numerical treatment have been discussed in [10, 5, 46, 8] and [9, sect. 8.9.1]

Let us define $N(P) = 1$ for $P \in B_{\mathrm{e}}$ and

$$w(P) = \begin{cases} p_{\mathrm{e}}(P), & P \in B_{\mathrm{e}}, \\ p_{\mathrm{i}}(P), & P \in B. \end{cases}$$

Then, we can combine (4.1) with the representation (3.7) to obtain

$$w(P) - k_{\mathrm{e}}^2 \int \{1 - N(Q)\} w(Q)\, G_{\mathrm{e}}(P,Q)\, dV_Q = p_{\mathrm{inc}}(P) \tag{4.2}$$

for all $P \in B \cup B_{\mathrm{e}}$, where the integration is over all Q. We recognize this equation as the *Lippmann–Schwinger equation* [24]; see, for example, [1], [12, sect. 8.2] [32, sect. 10.3], and [35, Thm. 9.4]. Notice that our derivation shows that the Lippmann–Schwinger equation is valid even when $N(Q)$ is discontinuous as Q crosses S. This fact is implicit in [34] and explicit in [44].

4.3. An alternative equation. As we know that $v \equiv p_{\mathrm{i}}$ solves (2.5) in B, we can use this fact to rewrite the expression for $\mathcal{L}v$. Thus

$$(\alpha - 1)\, (\operatorname{grad} v) \cdot (\operatorname{grad}_Q G_{\mathrm{e}}) = \operatorname{div}\{(\alpha - 1)\, G_{\mathrm{e}} \operatorname{grad} v\} - G_{\mathrm{e}} \operatorname{div}\{(\alpha - 1) \operatorname{grad} v\}$$

and

$$\begin{aligned} \operatorname{div}\{(\alpha - 1) \operatorname{grad} v\} &= (\alpha - 1)\, \nabla^2 v + \rho_{\mathrm{e}} \left(\operatorname{grad} \rho_{\mathrm{i}}^{-1}\right) \cdot \operatorname{grad} v \\ &= (\alpha - 1) \left\{\nabla^2 v + \rho_{\mathrm{i}} \left(\operatorname{grad} \rho_{\mathrm{i}}^{-1}\right) \cdot \operatorname{grad} v\right\} + \rho_{\mathrm{i}} \left(\operatorname{grad} \rho_{\mathrm{i}}^{-1}\right) \cdot \operatorname{grad} v \\ &= (1 - \alpha)\, k_{\mathrm{e}}^2 N v - \rho_{\mathrm{i}}^{-1} (\operatorname{grad} \rho_{\mathrm{i}}) \cdot \operatorname{grad} v. \end{aligned}$$

Hence, substituting in (3.5), we obtain

$$(\mathcal{L}v)(P) = \int_B G_{\mathrm{e}}(P,Q)\, (Vv)(Q)\, dV_Q + (\mathcal{L}_E v)(P),$$

where Vv is defined by (2.6) and

$$\begin{aligned} (\mathcal{L}_E v)(P) &= \int_B \operatorname{div}\{(\alpha(Q) - 1)\, G_{\mathrm{e}} \operatorname{grad} v\}\, dV_Q \\ &= \int_S \{\alpha(q) - 1\}\; G_{\mathrm{e}}(P,q)\, \frac{\partial v}{\partial n}\, ds_q, \end{aligned} \tag{4.3}$$

by the divergence theorem. Thus, the Scattering Problem can be reduced to solving

$$p_{\mathrm{i}}(P) = p_{\mathrm{inc}}(P) + \int_B G_{\mathrm{e}}(P,Q)\, (Vp_{\mathrm{i}})(Q)\, dV_Q + p_E(P), \quad P \in B, \tag{4.4}$$

where

$$p_E(P) = (\mathcal{L}_E p_{\mathrm{i}})\,(P) = \int_S \left(\frac{\partial p_{\mathrm{e}}}{\partial n} - \frac{\partial p_{\mathrm{i}}}{\partial n}\right) G_{\mathrm{e}}(P,q)\, ds_q$$

and we have used $(2.4)_2$ in (4.3).

When both $\rho_{\rm i}$ and N are constants, (4.4) reduces to an equation obtained previously by Ramm [37]. However, in this situation, the scatterer is homogeneous and the problem can be reduced to boundary integral equations over S; see [20].

If we had attempted to solve the Scattering Problem using the formal method described in section 1.1, we would have obtained precisely (4.4) but with $p_E(P) \equiv 0$. In general, this extra term is not zero, and its magnitude is difficult to estimate. Observe that, from (4.3), p_E does vanish if $\rho_{\rm i}(q) = \rho_{\rm e}$ for all $q \in S$, which means that the density is continuous across S. Otherwise, the single-layer potential $p_E(P)$ should be retained.

4.4. Electromagnetic waves. For Maxwell's equations, we can encounter exactly the same difficulty as in acoustics. Thus, in an inhomogeneous medium, the electric field $\boldsymbol{E}$ satisfies

$$\mu \operatorname{curl} \left\{\mu^{-1} \operatorname{curl} \boldsymbol{E}\right\} - k^2 \boldsymbol{E} = \boldsymbol{0},$$

where $\mu(\boldsymbol{r})$ is the magnetic permeability, $k(\boldsymbol{r}) = \omega\sqrt{\mu\varepsilon}$, and $\varepsilon(\boldsymbol{r})$ is the electric permittivity. Moreover, the transmission conditions across an interface S are that $\boldsymbol{n} \times \boldsymbol{E}$ and $\boldsymbol{n} \times (\mu^{-1} \operatorname{curl} \boldsymbol{E})$ should both be continuous. (See, for example, [9, sect. 8.9.2]; for existence and uniqueness theorems, see [31].) We can then mimic the derivations in section 3 to show that discontinuities in μ across S will lead to an extra term similar to p_E in (4.4).

Specifically, we find the following electromagnetic analogue of (3.6):

$$\left\{1 - \frac{1}{3}\left(1 - \frac{\varepsilon_{\rm i}(P)}{\varepsilon_{\rm e}}\right)\right\} \boldsymbol{E}_{\rm i}(P) - \int_B \boldsymbol{W}(P,Q)\, dV_Q = \boldsymbol{E}_{\rm inc}(P), \quad P \in B. \tag{4.5}$$

Here, $\boldsymbol{E}_{\rm i}$ is the field in B, $\boldsymbol{E}_{\rm inc}$ is the incident field, $\varepsilon_{\rm e}$ is the (constant) electric permittivity in $B_{\rm e}$, and $\varepsilon_{\rm i}$ is the electric permittivity in B. The field $\boldsymbol{W}$ is defined by

$$\begin{aligned}\boldsymbol{W}(P,Q) = {} & \left(1 - \frac{\mu_{\rm e}}{\mu_{\rm i}}\right) (\operatorname{grad}_Q G_{\rm e}) \times \operatorname{curl} \boldsymbol{E}_{\rm i}(Q) \\ & + \left(1 - \frac{\varepsilon_{\rm i}}{\varepsilon_{\rm e}}\right) \left\{k_{\rm e}^2 G_{\rm e} \boldsymbol{E}_{\rm i} - \operatorname{grad}_P \left(\boldsymbol{E}_{\rm i} \cdot \operatorname{grad}_Q G_{\rm e}\right)\right\},\end{aligned}$$

where $\mu_{\rm e}$ is the (constant) magnetic permeability in $B_{\rm e}$, $\mu_{\rm i}$ is the magnetic permeability in B, and $k_{\rm e} = \omega\sqrt{\mu_{\rm e}\varepsilon_{\rm e}}$. The integral in (4.5) is to be interpreted in the Cauchy principal-value sense with a spherical exclusion volume. In the special case that $\mu_{\rm i}(Q) \equiv \mu_{\rm e}$, (4.5) reduces to eqn. (2.1.41) in [40].

The electromagnetic analogue of (4.4) is

$$\boldsymbol{E}_{\rm i}(P) = \boldsymbol{E}_{\rm inc}(P) + \int_B G_{\rm e}(P,Q)\, \boldsymbol{V}(Q)\, dV_Q + \int_S \boldsymbol{F}(P,q)\, ds_q, \tag{4.6}$$

where

$$\begin{aligned}\boldsymbol{V} &= (k_{\rm e}^2 - k_{\rm i}^2)\boldsymbol{E}_{\rm i} - (\mu_{\rm i}^{-1} \operatorname{grad} \mu_{\rm i}) \times \operatorname{curl} \boldsymbol{E}_{\rm i} + \operatorname{grad} \operatorname{div} \boldsymbol{E}_{\rm i}, \\ \boldsymbol{F} &= \left(1 - \frac{\varepsilon_{\rm i}}{\varepsilon_{\rm e}}\right) (\boldsymbol{n} \cdot \boldsymbol{E}_{\rm i}) \operatorname{grad}_q G_{\rm e} \\ &\quad + \left\{\left(1 - \frac{\mu_{\rm e}}{\mu_{\rm i}}\right) (\boldsymbol{n} \times \operatorname{curl} \boldsymbol{E}_{\rm i}) + \boldsymbol{n}(\varepsilon_{\rm i}^{-1} \operatorname{grad} \varepsilon_{\rm i}) \cdot \boldsymbol{E}_{\rm i}\right\} G_{\rm e},\end{aligned}$$

and $k_{\rm i} = \omega\sqrt{\mu_{\rm i}\varepsilon_{\rm i}}$. Notice that $\boldsymbol{F}$ vanishes if $\varepsilon_{\rm i}(q) = \varepsilon_{\rm e}$, $\mu_{\rm i}(q) = \mu_{\rm e}$ and $\varepsilon_{\rm i}(P)$ is constant near S. Also, in the special case that $\mu_{\rm i}(Q) \equiv \mu_{\rm e}$, (4.6) reduces to eqn. (4.18) in [36].

4.5. Elastic waves. We can also consider analogous problems for elastic waves: scattering of elastic waves in a homogeneous solid by an inhomogeneous (and anisotropic) inclusion. It turns out that the formal method of section 1.1 (described in detail in [33, 17]) and the method of section 3 yield exactly the same equation for the displacement within the inclusion, $\boldsymbol{u}$. This is because the "extra term" analogous to p_E involves the discontinuity in the traction vector across the interface: this is zero for a perfect (welded) interface. (For imperfect interfaces [27], a nonzero contribution is obtained.) Some applications of the volume equation for $\boldsymbol{u}$ can be found in [38, 6, 7]. A polarization approach (leading to a coupled system) has been developed by Willis [45]. Long-wave approximations can be found in [45, 38].

Appendix A. Werner's solution. Werner [43] proved an existence theorem for a problem that is very similar to our Scattering Problem: he considered inhomogeneous forms of (2.2) and (2.3) but supposed that $p_{\rm inc} \equiv 0$. His method leads to the following integral representations:

$$p_{\rm sc}(P) = -\int_S \left\{ \alpha(q)\,\mu(q)\,G_{\rm e}(P,q) - \nu(q)\,\frac{\partial}{\partial n_q} G_{\rm e}(P,q) \right\} ds_q, \quad P \in B_{\rm e},$$

$$p_{\rm i}(P) = \int_S \left\{ \mu\,G_{\rm e}(P,q) - \nu\,\frac{\partial}{\partial n_q} G_{\rm e}(P,q) \right\} ds_q + \int_B \varphi(Q)\,G_{\rm e}(P,Q)\,dV_Q, \quad P \in B.$$

The two surface densities, $\mu(q)$ and $\nu(q)$, and the volume density, $\varphi(Q)$, satisfy the following system of integral equations:

$$\nu(p) + \int_S \mu(q)\,\{1-\alpha(q)\}\,G_{\rm e}(p,q)\,ds_q + \int_B \varphi(Q)\,G_{\rm e}(p,Q)\,dV_Q = -p_{\rm inc}, \tag{A.1}$$

$$\frac{1+\alpha}{2}\mu + \int_S \mu(1-\alpha)\frac{\partial}{\partial n_p} G_{\rm e}(p,q)\,ds_q + \int_B \varphi \frac{\partial}{\partial n_p} G_{\rm e}(p,Q)\,dV_Q = -\frac{\partial p_{\rm inc}}{\partial n_p}, \tag{A.2}$$

$$\varphi + \int_S V_P \left\{ \mu\,G_{\rm e}(P,q) - \nu\,\frac{\partial G_{\rm e}}{\partial n_q} \right\} ds_q - \int_B \varphi\,(V_P G_{\rm e}(P,Q))\,dV_Q = 0. \tag{A.3}$$

Equations (A.1) and (A.2) hold for $p \in S$, whereas (A.3) holds for $P \in B$. V_P denotes the operator V (defined by (2.6)) applied with respect to P.

Werner's proof [43] can be adapted to show that the system (A.1)–(A.3) is uniquely solvable.

Appendix B. Volume potentials. Define a *volume potential* $W(P)$ by

$$W(P) = \int_B \varphi(Q)\,G_{\rm e}(P,Q)\,dV_Q,$$

where $G_{\rm e}(P,Q) = -\exp(ik_{\rm e}R)/(4\pi R)$. The properties of such potentials are similar to those of *Newtonian potentials* for which $k_{\rm e} = 0$: thus, define

$$W_0(P) = -\frac{1}{4\pi}\int_B \frac{\varphi(Q)}{R}\,dV_Q.$$

From [12, sect. 8.2], we have

$$(\nabla^2 + k_{\rm e}^2)W = \begin{cases} 0, & P \in B_{\rm e}, \\ \varphi(P), & P \in B, \end{cases} \tag{B.1}$$

where $B_{\rm e}$ is the region exterior to B. Also, if φ is piecewise continuous, then $W(P)$ and its first partial derivatives are continuous everywhere in three-dimensional space;

see, for example, [19, Chapter VI, sect. 3], [14, Chapter IV, sect. 1.2], [41, sect. 3.9], [12, Thm. 8.1], and [16, sect. 4.2].

We also require the behavior of the second derivatives of W near the boundary of B, S. As Kellogg remarks [19, p. 156], "in general, the derivatives of second order will not exist. It is clear that they cannot all be continuous, for as we pass from an exterior to an interior point through the boundary where φ is not 0, $\nabla^2 W_0$ experiences a break of φ." This discontinuous behavior is described in [25, p. 175] and [26, p. 125]:

$$\left[\frac{\partial^2}{\partial n\, \partial x_i^P} W_0(P)\right] = -\varphi(p)\, n_i(p), \quad p \in S.$$

The same formula holds for W because the difference, $W - W_0$, is less singular. Finally, the result (3.14) is obtained from (3.10), using $\varphi(Q) = (1-\alpha)\partial v/\partial x_i^Q$. Note that the second term in (3.10) is a volume potential: it does not contribute as its first derivatives are continuous across S.

Acknowledgments. I am grateful for the constructive comments provided by two anonymous referees. I also thank Prof. D. Tataru for his help with the unique continuation aspect of section 2.1.

REFERENCES

[1] J. F. Ahner, *Scattering by an inhomogeneous medium*, J. Inst. Math. Appl., 19 (1977), pp. 425–439.
[2] N. Aronszajn, *A unique continuation theorem for solutions of elliptic partial differential equations or inequalities of second order*, J. Math. Pures Appl., 36 (1957), pp. 235–249.
[3] P. G. Bergmann, *The wave equation in a medium with a variable index of refraction*, J. Acoust. Soc. Amer., 17 (1946), pp. 329–333.
[4] L. M. Brekhovskikh, *Waves in Layered Media*, Academic Publishers, New York, 1960.
[5] O. P. Bruno and A. Sei, *A fast high-order solver for EM scattering from complex penetrable bodies: TE case*, IEEE Trans. Antennas and Propagation, 48 (2000), pp. 1862–1864.
[6] D. E. Budreck and J. H. Rose, *Three-dimensional inverse scattering in anisotropic elastic media*, Inverse Problems, 6 (1990), pp. 331–348.
[7] D. E. Budreck and J. H. Rose, *Elastodynamic completeness relations for scattered wavefields*, SIAM J. Appl. Math., 51 (1991), pp. 1568–1584.
[8] Y. Chen, *A fast, direct algorithm for the Lippmann–Schwinger integral equation in two dimensions*, Adv. Comput. Math., 16 (2002), pp. 175–190.
[9] W. C. Chew, *Waves and Fields in Inhomogeneous Media*, Van Nostrand, New York, 1990.
[10] W. C. Chew and C.-C. Lu, *The use of Huygens' equivalence principle for solving the volume integral equation of scattering*, IEEE Trans. Antennas and Propagation, 41 (1993), pp. 897–904.
[11] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, Wiley, New York, 1983.
[12] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer, Berlin, 1992.
[13] D. Colton and P. Monk, *Mathematical and numerical methods in inverse acoustic scattering theory*, ZAMM Z. Angew. Math. Mech., 81 (2001), pp. 723–731.
[14] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, Wiley-Interscience, New York, 1962.
[15] E. Gerjuoy and D. S. Saxon, *Variational principles for the acoustic field*, Physical Rev., 94 (1954), pp. 1445–1458.
[16] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, Berlin, 1998.
[17] J. E. Gubernatis, E. Domany, and J. A. Krumhansl, *Formal aspects of the theory of the scattering of ultrasound by flaws in elastic materials*, J. Appl. Phys., 48 (1977), pp. 2804–2811.
[18] S. Hassani, *Mathematical Physics*, Springer, New York, 1999.
[19] O. D. Kellogg, *Foundations of Potential Theory*, Springer, Berlin, 1929.

[20] R. E. Kleinman and P. A. Martin, *On single integral equations for the transmission problem of acoustics*, SIAM J. Appl. Math., 48 (1988), pp. 307–325.
[21] H. Koch and D. Tataru, *Carleman estimates and unique continuation for second-order elliptic equations with nonsmooth coefficients*, Comm. Pure Appl. Math., 54 (2001), pp. 339–360.
[22] G. A. Kriegsmann and E. L. Reiss, *Low frequency scattering by local inhomogeneities*, SIAM J. Appl. Math., 43 (1983), pp. 923–934.
[23] Y. L. Li, C. H. Liu, and S. J. Franke, *Three-dimensional Green's function for wave propagation in a linearly inhomogeneous medium—the exact analytic solution*, J. Acoust. Soc. Amer., 87 (1990), pp. 2285–2291.
[24] B. A. Lippmann and J. Schwinger, *Variational principles for scattering processes.* I, Physical Rev., 79 (1950), pp. 469–480.
[25] W. D. MacMillan, *The Theory of the Potential*, Dover, New York, 1958.
[26] E. Martensen, *Potentialtheorie*, Teubner, Stuttgart, 1968.
[27] P. A. Martin, *Boundary integral equations for the scattering of elastic waves by elastic inclusions with thin interface layers*, J. Nondestructive Evaluation, 11 (1992), pp. 167–174.
[28] P. A. Martin, *Acoustic scattering by inhomogeneous spheres*, J. Acoust. Soc. Amer., 111 (2002), pp. 2013–2018.
[29] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*, Princeton University Press, Princeton, NJ, 1986.
[30] C. Müller, *On the behavior of the solutions of the differential equation $\Delta U = F(x, U)$ in the neighborhood of a point*, Comm. Pure Appl. Math., 7 (1954), pp. 505–515.
[31] C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer, Berlin, 1969.
[32] R. G. Newton, *Scattering Theory of Waves and Particles*, 2nd ed., Springer, New York, 1982.
[33] Y.-H. Pao and V. Varatharajulu, *Huygens' principle, radiation conditions, and integral formulas for the scattering of elastic waves*, J. Acoust. Soc. Amer., 59 (1976), pp. 1361–1371.
[34] D. N. Pattanayak and E. Wolf, *Scattering states and bound states as solutions of the Schrödinger equation with nonlocal boundary conditions*, Physical Rev. D, 13 (1976), pp. 913–923.
[35] D. B. Pearson, *Quantum Scattering and Spectral Theory*, Academic Press, London, 1988.
[36] A. J. Poggio and E. K. Miller, *Integral equation solutions of three-dimensional scattering problems*, in Computer Techniques for Electromagnetics, R. Mittra, ed., Pergamon, Oxford, 1973, pp. 159–264.
[37] A. G. Ramm, *Scattering by a penetrable body*, J. Math. Phys., 25 (1984), pp. 469–471.
[38] J. M. Richardson, *Scattering of elastic waves from symmetric inhomogeneities at low frequencies*, Wave Motion, 6 (1984), pp. 325–336.
[39] A. Tourin, M. Fink, and A. Derode, *Multiple scattering of sound*, Waves Random Media, 10 (2000), pp. R31–R60.
[40] L. Tsang, J. A. Kong, K.-H. Ding, and C. O. Ao, *Scattering of Electromagnetic Waves: Numerical Simulations*, Wiley, New York, 2001.
[41] J. Van Bladel, *Singular Electromagnetic Fields and Sources*, Clarendon Press, Oxford, 1991.
[42] P. Werner, *Zur mathematischen Theorie akustischer Wellenfelder*, Arch. Ration. Mech. Anal., 6 (1960), pp. 231–260.
[43] P. Werner, *Beugungsprobleme der mathematischen Akustik*, Arch. Ration. Mech. Anal., 12 (1963), pp. 155–184.
[44] V. H. Weston, *Multifrequency inverse problem for the reduced wave equation with sparse data*, J. Math. Phys., 25 (1984), pp. 1382–1390.
[45] J. R. Willis, *A polarization approach to the scattering of elastic waves*—I. *Scattering by a single inclusion*, J. Mech. Phys. Solids, 28 (1980), pp. 287–305.
[46] X. M. Xu and Q. H. Liu, *Fast spectral domain method for acoustic scattering problems*, IEEE Trans. Ultrasonics, Ferroelectrics, & Frequency Control, 48 (2001), pp. 522–529.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 309–321

ON LOCAL ISOTROPY IN STRATIFIED HOMOGENEOUS TURBULENCE*

B. A. PETTERSSON REIF† AND Ø. ANDREASSEN†

Abstract. This paper examines the postulate of local isotropy in stratified homogeneous turbulence from a theoretical point of view. The study is based on a priori analysis of the evolution equations governing single-point turbulence statistics that are formally consistent with the Navier–Stokes equations. The Boussinesq approximation has been utilized to account for the effect of buoyancy—a simplifying assumption which constitutes an excellent approximation for the case considered here. The study concludes that the hypothesis of local isotropy is formally inconsistent with the Navier–Stokes equations in homogeneous stratified turbulence. An estimate is provided that suggests that local isotropy may constitute only a physically justifiable approximation in the limit of a clear-cut separation between the time scales associated with the imposed buoyancy and the turbulent eddy turnover time scale. This is unlikely to happen in most flows, at least those not too far from equilibrium. The results also suggest that the dynamical dependence of the small-scale turbulence on large-scale anisotropies associated with imposed density stratification is significantly stronger than that caused by an imposed mean straining.

Key words. local isotropy, homogeneous turbulence, single-point correlation, density stratification, shear turbulence

AMS subject classification. 76F05

DOI. 10.1137/S0036139903421559

1. Introduction.

It is well established that the imposition of density stratification and mean straining significantly promotes anisotropy on the energetic large-scale turbulence motion. It is frequently also argued that the small-scale motion would remain virtually unaffected by the large-scale anisotropy at sufficiently high Reynolds number (Re). This view inherently assumes that any direct effects of the large-scale motion on the smallest scales would be negligible at high enough Re and that large-scale anisotropies would not mediate across the spectral gap fast enough to overcome the nonlinear scrambling of the cascade process. Small-scale turbulence is therefore expected to be statistically independent of the large-scale motion at sufficiently high Re. This is essentially the postulate of local isotropy put forward by Kolmogorov [6] more than 70 years ago, a postulate that has been enormously influential in turbulence research.

The conjecture of locally isotropic turbulence is sometimes also based on the notion of a clear-cut separation of characteristic time scales; since the limiting behavior of the small-to-large-scale time scale ratio asymptotes to $\tau/T \sim Re^{-1/2} \to 0$ as $Re \to \infty$, it is believed that small-scale turbulence would have sufficiently long time to interact with itself and to establish a state of directional independence, or local isotropy.

The terminology "local isotropy" alludes to statistical isotropy of the smallest, dissipative scales of motion, i.e., scales much smaller than the energetic large-scale motion. Mathematically, "isotropy" implies that any statistical measure must display invariance to arbitrary reflections and rotations. Local isotropy is, however, not only

*Received by the editors January 21, 2003; accepted for publication (in revised form) April 29, 2003; published electronically November 19, 2003.
http://www.siam.org/journals/siap/64-1/42155.html
†Norwegian Defence Research Establishment (FFI), NO-2027 Kjeller, Norway (bre@ffi.no, oya@ffi.no).

a concept of theoretical interest. It is in fact widely used, for instance, by experimentalists to infer the rate of viscous dissipation of turbulent kinetic energy (formally defined as $\varepsilon \equiv 2\nu s'_{ij}s'_{ij}$) by only conducting measurements of one of the six independent components of the fluctuating rate-of-strain tensor s'_{ij} (defined in (2.8)). In particular, by imposing the assumption of local isotropy, the number of derivative correlations that must be determined can be reduced from twelve to just a single one, e.g., $\varepsilon \sim \overline{(\partial_1 u_1)^2}$; see, e.g., [5].

There exist several hundred articles and papers on the concept of locally isotropic turbulence. Among the pioneering ones are those due to Kolmogorov [6] and Obukhov [10], to mention only a few. Monin and Yaglom [8] provide an extensive review on the early developments of the topic, whereas more recent reviews are provided by Nelkin [9], Frisch [3], Sreenivasan and Antonia [14], and Warhaft [20]. Among the many studies there are a growing number of theoretical, experimental, and numerical investigations that suggest that the concept of local isotropy is somewhat dubious. Townsend [17] and Uberoi [18] were probably among the first to suggest that there exists a direct effect of large-scale anisotropy on the dissipative scales of motion, in addition to the indirect influence through the cascading process. This view was supported by, e.g., Durbin and Speziale [2] who demonstrated that, as a formally consistent consequence of the Navier–Stokes equations, there must indeed exist a direct effect of mean straining on the dissipative scales. They concluded that local isotropy is a physically implausible argument in turbulence affected by mean straining.

Brasseur and Wei [1] and Yeung, Brasseur, and Wang [22] conducted numerical studies of the triadic interactions in forced turbulence. These studies demonstrated that triadic interactions between widely disparate scales directly modified the structure of the smallest scales in accordance with the structure of the large energetic ones. Experimental results in uniform turbulent shear flow [12] also imply a direct coupling between the large- and small-scales in strained turbulence. They further concluded, fully in line with Durbin and Speziale [2], that the hypothesis of local isotropy in isothermal turbulent shear flows seems untenable even in the limit of infinite Re.

Sreenivasan [13] reviewed experimental work on local isotropy of passive scalar fields and suggested that local isotropy is not a natural concept for scalar fields in shear flows, except perhaps for such extreme Re that are of no practical use on earth. Van Atta [19] analyzed experimental data in stably stratified turbulence and noted that the effects are surprisingly rapid, destroying the directional independence of the smallest scales as soon as buoyancy forces become dynamically important. This was essentially confirmed by the enormous numerical simulations of Werne and Fritts [21] who studied a stratified shear layer. They found that turbulence affected by mean straining tends to develop a state of local streamwise axisymmetry, as opposed to local isotropy. The concept of locally axisymmetric turbulence in strained homogeneous flows has been theoretically and experimentally considered by George and Hussain [5] who concluded that a theory of local axisymmetry provides more credibility to the numerous measurements that have failed to confirm local isotropy. These findings, along with many more not mentioned here, add to the body of literature that sheds new light on the concept of locally isotropic turbulence.

The present study examines local isotropy from a theoretical point of view. It extends the approach suggested in [2] to homogeneous flows affected by both density stratification and mean straining. The methodology is based on an examination of the dynamical equations governing single-point turbulence correlations that are characteristic of small-scale turbulence; these equations are formally consistent with the

Navier–Stokes equations. The objective of the study is to provide insight into whether or not the hypothesis of local isotropy is a formally consistent concept in stratified flows, and if not, to provide also an estimate of under what circumstances it would constitute a physically plausible approximation. The practical implications are related to the development of semiempirical models intended to describe the statistical coupling between large- and small-scale turbulence; a development which is crucial for improved turbulence model formulations.

2. The evolution of single-point turbulence statistics. The present analysis is based on the incompressible Navier–Stokes equations in the limit of homogeneous turbulence and to cases where the Boussinesq approximation constitutes a reasonable assumption. The latter assumption is not believed to be a severe limitation in the present context; the Boussinesq approximation represents a first-order perturbation of the fluid density. In cases where this approximation fails, an even stronger effect of buoyancy is expected.

Single-point turbulence statistics allude to correlations of fluctuating quantities evaluated at the same position in space and time. Dynamical equations governing these statistics can be rigorously derived from the conservation equations for mass, momentum (Navier–Stokes), and energy:

$$\partial_i \tilde{u}_i = 0, \tag{2.1}$$

$$\partial_t \tilde{u}_i + \tilde{u}_k \partial_k \tilde{u}_i = -\partial_i \tilde{p} + \nu \nabla^2 \tilde{u}_i + \frac{\rho}{\rho_0} g_i, \tag{2.2}$$

$$\partial_t \tilde{\theta} + \tilde{u}_k \partial_k \tilde{\theta} = \kappa \nabla^2 \tilde{\theta} + 2 \frac{\nu}{c_v} \tilde{s}_{ij} \tilde{s}_{ij}. \tag{2.3}$$

Repeated indices imply summation, e.g., $\tilde{u}_k \partial_k \tilde{\theta} = \tilde{u}_1 \partial_1 \tilde{\theta} + \tilde{u}_2 \partial_2 \tilde{\theta} + \tilde{u}_3 \partial_3 \tilde{\theta}$. The superscript $\tilde{}$ denotes instantaneous quantities, the subscript $_0$ denotes a constant reference state, and $\tilde{s}_{ij} \equiv \frac{1}{2}(\partial_i \tilde{u}_j + \partial_j \tilde{u}_i)$ is the instantaneous rate-of-strain tensor. Spatial and temporal differentiation are denoted $\partial_m \equiv \partial/\partial x_m$ and $\partial_t \equiv \partial/\partial t$, respectively, and $\nabla^2 = \partial^2_{mm} \equiv \partial^2/(\partial x_m \partial x_m)$. Here, $\nu = \mu/\rho_0$ is the kinematic viscosity and $\kappa = \alpha/(\rho_0 c_v)$ the thermal diffusivity, where μ, α, and c_v denote the dynamic viscosity, thermal conductivity, and specific heat, respectively. $\mathbf{g}$ is the gravitational acceleration. According to the Boussinesq approximation, the density ratio ρ/ρ_0 in (2.2) varies according to

$$\frac{\rho}{\rho_0} = 1 - \beta(\tilde{\theta} - \Theta_0), \tag{2.4}$$

where $\beta \equiv [-\partial \log(\rho)/\partial\theta]_\Theta$ defines the thermal expansion coefficient at fixed mean temperature $\Theta(\mathbf{x}, t)$.

Equations governing fluctuating quantities can systematically be derived using the following procedure:

1. Decompose the instantaneous velocity, pressure, and temperature fields into mean and fluctuating parts, i.e., $\tilde{a}(\mathbf{x}, t) = A(\mathbf{x}, t) + a(\mathbf{x}, t)$.

2. Average to obtain the dynamical equation for the mean field; $A(\mathbf{x}, t) \equiv \overline{\tilde{a}(\mathbf{x}, t)}$, since $\overline{a(\mathbf{x}, t)} \equiv 0$ by definition.

3. Obtain the evolution equations for the fluctuating fields $a(\mathbf{x}, t)$ by subtracting 2 from 1.

Using this procedure, the evolution of the ith component of the fluctuating velocity $u_i(\mathbf{x},t)$ for an incompressible fluid can then be written as

$$\partial_t u_i + U_k\partial_k u_i + u_k\partial_k U_i + u_k\partial_k u_i + \overline{u_k\partial_k u_i} = -\frac{1}{\rho_0}\partial_i p + \nu\nabla^2 u_i - \beta g_i\theta, \tag{2.5}$$

$$\partial_i u_i = 0. \tag{2.6}$$

Here, $\mathbf{U}(\mathbf{x},t)$ denotes the mean velocity field and $\theta(\mathbf{x},t)$ is the fluctuating temperature field. The corresponding dynamical equation governing the evolution of the fluctuating temperature field $\theta(\mathbf{x},t)$ reads as follows:

$$\begin{aligned}\partial_t\theta + U_m\partial_m\theta = &-u_m\partial_m\Theta - u_m\partial_m\theta + \kappa\nabla^2\theta + 4\frac{\nu}{c_v}S_{ij}s'_{ij}\\ &+ 2\frac{\nu}{c_v}\left(s'_{ij}s'_{ij} - \overline{s'_{ij}s'_{ij}}\right).\end{aligned} \tag{2.7}$$

Here,

$$s'_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i) \tag{2.8}$$

and

$$S_{ij} = \frac{1}{2}(\partial_i U_j + \partial_j U_i) \tag{2.9}$$

denote the fluctuating and mean rate-of-strain tensor, respectively.

Transport equations governing suitable turbulence correlation can now be constructed from (2.5)–(2.7), and the results are formally consistent with the incompressible Navier–Stokes equations in the limit of the Boussinesq approximation. The assumption of homogeneity constitutes the only additional simplification and it implies that statistical measures of the flow must be translational invariant, i.e., single-point correlations are spatially constant.

The fluctuating pressure field $p(\mathbf{x},t)$ in (2.5) is the solution to a Poisson equation which can be obtained by taking the divergence of (2.5). Invoking the incompressibility and homogeneity constraints then gives

$$\nabla^2 p = -\rho_0\partial_i u_k\partial_k u_i - 2\rho_0\partial_k u_i\partial_i U_k - \rho_0\beta g_i\partial_i\theta, \tag{2.10}$$

which represents nonlocal effects on single-point statistics.[1] The fluctuating momentum and temperature equations, (2.5) and (2.7), can symbolically be written in operator form as $\mathcal{R}u_i = 0$ and $\mathcal{R}\theta = 0$, respectively. The transport equation governing the second-order moments, $\tau_{ij} = \overline{\tau'_{ij}} \equiv \overline{u_i u_j}$, is readily obtained by multiplying (2.5) by u_j, adding the result to itself with i and j interchanged, and finally averaging. This can symbolically be written as $\overline{u_j\mathcal{R}u_i + u_i\mathcal{R}u_j} = 0$. The result for homogeneous turbulence reads

$$\begin{aligned}d_t\tau_{ij} = &-\frac{1}{\rho_0}\left(\overline{u_j\partial_i p} + \overline{u_i\partial_j p}\right) - (\tau_{ik}\partial_k U_j + \tau_{jk}\partial_k U_i)\\ &- \varepsilon_{ij} - \beta\left(g_i\overline{u_j\theta} + g_j\overline{u_i\theta}\right),\end{aligned} \tag{2.11}$$

[1] It is interesting to note that the solution of (2.10) shows that the evolution of single-point moments implicitly depends on two-point correlations, i.e., correlations of velocity components evaluated at different position in space; see, e.g., Rotta [11] for more details.

where $d_t \equiv d/dt$ is the local time derivative. Recall that all spatial derivatives of turbulence correlations are zero in homogeneous turbulence. The second-order viscous dissipation rate tensor is given by

$$\varepsilon_{ij} = 2\nu\overline{\varepsilon'_{ij}} \equiv 2\nu\overline{\partial_m u_i \partial_m u_j}. \tag{2.12}$$

The evolution equation governing the turbulent kinetic energy, $k \equiv \frac{1}{2}\tau_{ii}$, is obtained by taking the trace of (2.11) and multiplying by $\frac{1}{2}$:

$$d_t k = -\tau_{ik}\partial_k U_i - \varepsilon - \frac{1}{2}\beta g_i \overline{u_i\theta}, \tag{2.13}$$

where $\varepsilon \equiv \frac{1}{2}\varepsilon_{ii}$ is the rate of turbulent energy dissipation.

By first writing (2.11) as $\overline{\mathcal{R}\tau'_{ij}} = 0$, the corresponding transport equation for the third-order moments, $\tau_{ijk} \equiv \overline{\tau'_{ijk}} \equiv \overline{u_i u_j u_k}$, can then be derived as $\overline{u_k \mathcal{R}\tau'_{ij} + \tau'_{ij}\mathcal{R}u_k} = 0$. The result can be written as

$$\begin{aligned} d_t\tau_{ijk} = &-\frac{1}{\rho_0}\left(\overline{\tau'_{ij}\partial_k p} + \overline{\tau'_{ik}\partial_j p} + \overline{\tau'_{jk}\partial_i p}\right) \\ &- \left(\tau_{mij}\partial_m U_k + \tau_{mik}\partial_m U_j + \tau_{mjk}\partial_m U_i\right) \\ &- \varepsilon_{ijk} - \beta\left(g_k\overline{\tau'_{ij}\theta} + g_j\overline{\tau'_{ik}\theta} + g_i\overline{\tau'_{jk}\theta}\right), \end{aligned} \tag{2.14}$$

where

$$\varepsilon_{ijk} \equiv 2\nu\left(\overline{u_i\varepsilon'_{jk}} + \overline{u_j\varepsilon'_{ik}} + \overline{u_k\varepsilon'_{ij}}\right) \tag{2.15}$$

denotes the third-order viscous dissipation rate tensor.

The equation governing the transport of turbulent heat flux ($\overline{u_i\theta}$) can readily be derived as $\overline{\theta\mathcal{R}u_i + u_i\mathcal{R}\theta} = 0$:

$$d_t\overline{u_i\theta} = -\frac{1}{\rho_0}\overline{\theta\partial_i p} + \mathcal{P}_{i\theta} - \varepsilon_{i\theta} - \beta g_i\overline{\theta^2} + 4\frac{\nu}{c_v}S_{kj}\overline{u_i s'_{kj}} + 2\frac{\nu}{c_v}\overline{u_i s'_{kj}s'_{kj}}, \tag{2.16}$$

where

$$\varepsilon_{i\theta} \equiv (\kappa + \nu)\,\overline{\partial_m\theta\partial_m u_i} \tag{2.17}$$

and

$$\mathcal{P}_{i\theta} = -\left(\overline{u_m\theta}\partial_m U_i + \tau_{mi}\partial_m\Theta\right) \tag{2.18}$$

represent the rate of dissipation and production of turbulent heat flux, respectively.

To this end the dynamical equations governing the turbulent heat flux ($\overline{u_i\theta}$) and second- and third-order velocity-moments (τ_{ij} and τ_{ijk}) have been derived. This rather limited choice of basic single-point correlations suffices to assess the validity of the local isotropy postulate in stratified turbulence and to provide an estimate of when this hypothesis may constitute a physically plausible approximation. It should be noted, however, that the above-mentioned correlations are characteristic for the large-scale energetic part of the turbulence spectrum. In order to study the dynamics

of the dissipative scales, on the other hand, correlations characteristic for these scales must be considered. In particular, the dynamical equations governing the dissipation rate tensors ε_{ij}, ε_{ijk}, and $\varepsilon_{i\theta}$ appear in (2.11), (2.14), and (2.16), respectively. These tensors comprise correlations between fluctuating gradients and characterize therefore the high wave-number part in spectral space or the small scales in physical space.

The transport equation for dissipation rate $\varepsilon_{i\theta}$ of turbulent heat flux can be derived as $(\kappa+\nu)\overline{\partial_m\theta\partial_m(\mathcal{R}u_i)}+(\kappa+\nu)\overline{\partial_m u_i\partial_m(\mathcal{R}\theta)}=0$, and the result can be written as

$$d_t\varepsilon_{i\theta} = -\varepsilon_{k\theta}\partial_k U_i - 2\mathcal{E}_{mk\theta i}\partial_m U_k + 4\frac{\nu}{c_v}S_{kj}\mathcal{J}_{ikj}$$

$$-\frac{1}{2}\left(1+Pr^{-1}\right)\varepsilon_{ik}\partial_k\Theta - \frac{1}{2}\beta\left(1+Pr\right)g_i\varepsilon_\theta + \mathcal{F}_{i\theta} \tag{2.19}$$

for homogeneous turbulence, where $Pr\equiv\nu/\kappa$ is the Prandtl number, $\mathcal{E}_{mk\theta i}=\overline{\partial_m\theta\partial_k u_i}$, and $\mathcal{J}_{ikj}=\overline{\partial_m u_i\partial_m s'_{kj}}$. The dissipation rate of temperature variance $\overline{\theta^2}$ is defined as

$$\varepsilon_\theta \equiv 2\kappa\overline{\partial_m\theta\partial_m\theta}, \tag{2.20}$$

whereas the last term in (2.19) is

$$\mathcal{F}_{i\theta} = -\frac{\kappa+\nu}{\rho_0}\overline{\partial_i\theta\nabla^2 p} + (\kappa+\nu)\left(\overline{u_n\partial_n u_i\nabla^2\theta}+\overline{u_n\partial_n\theta\nabla^2 u_i}\right)$$

$$+(\kappa+\nu)^2\,\overline{\nabla^2\theta\nabla^2 u_i} + 2\frac{\nu}{c_v}\overline{\partial_m u_i\partial_m(s'_{kj}s'_{kj})}. \tag{2.21}$$

The evolution equation for ε_{ij} is derived as $\mathcal{R}\varepsilon_{ij}=2\nu[\overline{u_i\partial_m(\mathcal{R}u_j)}+\overline{u_j\partial_m(\mathcal{R}u_i)}]=0$ and the result reads

$$d_t\varepsilon_{ij} = \mathcal{H}_{ij} - 2\mathcal{E}_{mkij}\partial_m U_k - (\varepsilon_{jk}\partial_k U_i + \varepsilon_{ik}\partial_k U_j) - \beta\left(g_i\overline{\varepsilon'_{j\theta}}+g_j\overline{\varepsilon'_{i\theta}}\right), \tag{2.22}$$

where $\mathcal{E}_{ijkm}\equiv 2\nu\overline{\partial_i u_k\partial_j u_m}$ and

$$\mathcal{H}_{ij} = -4\nu^2\overline{\partial^2_{mk}u_i\partial^2_{mk}u_j} - 2\nu\left(\overline{\varepsilon'_{jk}\partial_k u_i}+\overline{\varepsilon'_{ik}\partial_k u_j}\right)$$

$$-\frac{2\nu}{\rho_0}\overline{(\partial_j u_i+\partial_i u_j)\,\nabla^2 p}. \tag{2.23}$$

The corresponding evolution equation for third-order dissipation rate tensor ε_{ijk}, (2.15), is obtained as $\mathcal{R}\varepsilon_{ijk}=2\nu(\mathcal{L}_{kij}+\mathcal{L}_{jik}+\mathcal{L}_{ijk})=0$, where $\mathcal{L}_{kij}=\overline{(u_k\mathcal{R}\varepsilon'_{ij}+\varepsilon'_{ij}\mathcal{R}u_k)}$. After some algebra, the final result can be symbolically written as

$$d_t\varepsilon_{ijk} = \mathcal{P}_{ijk}+\mathcal{U}_{ijk}+\mathcal{G}_{ijk}+\mathcal{N}_{ijk}, \tag{2.24}$$

where

$$\mathcal{P}_{ijk} = P_{ijk}+P_{jik}+P_{kij}, \tag{2.25}$$

$$\mathcal{U}_{ijk} = U_{ijk}+U_{jik}+U_{kij}, \tag{2.26}$$

$$\mathcal{G}_{ijk} = G_{ijk}+G_{jik}+G_{kij}, \tag{2.27}$$

$$\mathcal{N}_{ijk} = N_{ijk}+N_{jik}+N_{kij} \tag{2.28}$$

and

$$\begin{aligned} P_{kij} = &-\frac{2\nu}{\rho_0}\left(\overline{(u_i\nabla^2 u_j + u_j\nabla^2 u_i)\partial_k p}\right) \\ &-\frac{2\nu}{\rho_0}\left(\overline{(u_i\nabla^2 u_k + u_k\nabla^2 u_i)\partial_j p}\right) \\ &-\frac{2\nu}{\rho_0}\left(\overline{(u_k\nabla^2 u_j + u_j\nabla^2 u_k)\partial_i p}\right) \\ &-\frac{4\nu}{\rho_0}\overline{(\partial_i(u_j u_k) + \partial_j(u_i u_k) + \partial_k(u_i u_j))\nabla^2 p}, \end{aligned} \tag{2.29}$$

$$\begin{aligned} U_{kij} = &-2\nu\left(\overline{u_k\left(\partial_m u_i \partial_n u_j + \partial_m u_i \partial_n u_j\right)}\right)\partial_m U_n \\ &-\varepsilon_{nij}\partial_n U_k - \varepsilon_{knj}\partial_n U_i - \varepsilon_{kin}\partial_n U_j, \end{aligned} \tag{2.30}$$

$$G_{kij} = -2\nu\beta\left(g_i\overline{u_k\varepsilon'_{j\theta}} + \overline{u_k\varepsilon'_{i\theta}} + \overline{u_i\varepsilon'_{j\theta}}\right), \tag{2.31}$$

$$\begin{aligned} N_{kij} = &\ 2\nu\left(\overline{u_k\partial_m u_n\left(\partial_n u_i\partial_m u_j + \partial_n u_j\partial_m u_i\right)} - \overline{u_n\partial_n u_k\partial_m u_i\partial_m u_j}\right) \\ &-2\nu^2\left(\overline{\partial_m u_i\partial_m u_j\nabla^2 u_k} + \overline{\partial_m u_k\left(\partial_m u_j\nabla^2 u_i + \partial_m u_i\nabla^2 u_j\right)}\right) \\ &-4\nu^2\overline{u_k\nabla^2 u_i\nabla^2 u_j}. \end{aligned} \tag{2.32}$$

It follows directly from (2.29)–(2.32) that $P_{kij} = P_{kji}$, $U_{kij} = U_{kji}$, $G_{kij} = G_{kji}$, and $N_{kij} = N_{kji}$. Consequently, $\mathcal{P}_{ijk}$, $\mathcal{U}_{ijk}$, $\mathcal{G}_{ijk}$, and $\mathcal{N}_{ijk}$ are symmetric for any permutation of indices; see (2.28). This property is obviously required by the definition of ε_{ijk} (2.15).

3. Imposing local isotropy a priori. The theory of isotropic turbulence is essentially based on the fact that all statistical measures of the flow must display invariance to arbitrary reflections and rotations. The properties of isotropic tensors can here be put to good use in order to establish whether the postulate is formally consistent with the Navier–Stokes equations. This methodology was first used by Durbin and Speziale [2], where it was applied to the second-order dissipation rate equations (2.12) to investigate the impact of mean straining on the small scales. The objective here is not only to elucidate the impact of density stratification on small-scale turbulence, but also to relate it to the impact of mean straining.

It is well known that, at any given order, a general isotropic tensor can be written as a linear combination of a set of linearly independent isotropic tensors. The number of independent isotropic tensors depends on the order of the tensor itself. Here, we will consider tensors up to fourth rank. The most general isotropic forms of *any* first-, second-, third-, and fourth-order isotropic tensor[2] can be written as

$$\mathcal{X}_i = 0, \tag{3.1}$$

$$\mathcal{X}_{ij} = \alpha_0\delta_{ij}, \tag{3.2}$$

$$\mathcal{X}_{ijk} = \alpha_1\epsilon_{ijk} = 0, \tag{3.3}$$

$$\mathcal{X}_{ijkl} = \alpha_2\delta_{ij}\delta_{kl} + \alpha_3\delta_{ik}\delta_{jl} + \alpha_4\delta_{il}\delta_{jk}, \tag{3.4}$$

[2]These are not specific to turbulence correlation tensors, but generally valid for first- through fourth-order tensors.

where the fundamental isotropic tensor of rank 2 is the Kronecker delta,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

and of rank 3 is the Levi–Civita alternating tensor,

$$\epsilon_{ijk} = \begin{cases} 1 & \text{if } ijk \text{ is from the sequence 12312,} \\ -1 & \text{if } ijk \text{ is from the sequence 32132,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

As already alluded to, the implications of the small-scale isotropy postulate can be elucidated by writing the evolution equations (2.12), (2.15), and (2.19) in their most general isotropic forms using (3.1)–(3.4).

3.1. First-order velocity-temperature correlations. Let us first consider the equation governing the dissipation rate of turbulent heat flux. The isotropic form of (2.19) is obtained by substituting

$$\varepsilon_{i\theta}^{ISO} = 0, \tag{3.7}$$

$$\mathcal{F}_{i\theta}^{ISO} = 0, \tag{3.8}$$

$$\varepsilon_{ik}^{ISO} = \frac{1}{3}\varepsilon_{mm}\delta_{ik}, \tag{3.9}$$

$$\mathcal{E}_{km\theta i}^{ISO} \sim \epsilon_{kmi} = 0, \tag{3.10}$$

$$\mathcal{J}_{ikj}^{ISO} \sim \epsilon_{ikj} = 0, \tag{3.11}$$

which follows from (3.1)–(3.4). The last two results follow from the symmetry properties $\mathcal{E}_{km\theta i} = \mathcal{E}_{mk\theta i}$ and $\mathcal{J}_{ikj} = \mathcal{J}_{ijk}$, where the former only applies to homogeneous turbulence. The isotropic form of (2.19) then becomes

$$0 = -\frac{2}{3}\varepsilon\partial_i\Theta - \varepsilon_\theta g_i \beta Pr, \tag{3.12}$$

where $\varepsilon \equiv \frac{1}{2}\varepsilon_{mm}$ is the dissipation rate of turbulent kinetic energy. According to (3.12), isotropy would first require that the gravitation (g_i) must act in the direction of the mean temperature gradient $\partial_i\Theta$, which obviously not is generally true. Second, *if* the direction of the gravitational acceleration happens to coincide with the mean temperature gradient, e.g., $i = 2$, the resulting relationship $2\varepsilon\partial_2\Theta = -3\varepsilon_\theta\beta g_2 Pr$ seems far too stringent to be generally true. The implication of local isotropy, i.e., that $\varepsilon_{i\theta} = 0$, is therefore formally inconsistent with the Navier–Stokes equations. In fact, a closer examination of the evolution equation governing the third-order "generalized" dissipation tensor $\mathcal{E}_{km\theta i}$ yields the following additional constraints: $\varepsilon\partial_i\Theta = 0$ if $m \neq k = i$. For $m = k \neq i$, (3.12) is recovered. The terminology "generalized" alludes to the relation $(\kappa + \nu)\mathcal{E}_{mm\theta i} \equiv \varepsilon_{i\theta}$. Another interesting observation that can be made from (2.19) is that mean straining does not formally conflict with the assumption of local isotropy on this particular level of velocity-temperature correlation.

3.2. Second-order velocity-moments. Local isotropy on the second-order moment level requires (2.22) to balance in the isotropic limit (3.4). The terms in (2.22)

are replaced by their most general isotropic counterparts, and the result is

$$\varepsilon_{i\theta}^{ISO} = 0, \tag{3.13}$$

$$\varepsilon_{ij}^{ISO} = \frac{2}{3}\varepsilon\delta_{ij}, \tag{3.14}$$

$$\mathcal{H}_{ij}^{ISO} = \frac{1}{3}\mathcal{H}_{mm}\delta_{ij} = \frac{2}{3}\mathcal{H}\delta_{ij}, \tag{3.15}$$

$$\mathcal{E}_{mkij}^{ISO} = \varepsilon\left(\alpha_2\delta_{ij}\delta_{kl} + \alpha_3\delta_{ik}\delta_{jl} + \alpha_4\delta_{il}\delta_{jk}\right), \tag{3.16}$$

where the coefficients $\alpha_2 - \alpha_4$ are determined by imposing (i) homogeneity ($\mathcal{E}_{mkij} = \mathcal{E}_{kmij}$), (ii) continuity ($\mathcal{E}_{mkmj} = 0$), and (iii) the definition $\mathcal{E}_{mmij} = 2\varepsilon$. These constraints yield $\alpha_1 = 4/15$, $\alpha_2 = -1/15 = \alpha_3$. The resulting isotropic form of (2.12) can then be written as

$$d_t\varepsilon\delta_{ij} = \mathcal{H}\delta_{ij} - \frac{2}{5}\varepsilon S_{ij}, \tag{3.17}$$

where $\mathcal{H} \equiv \frac{1}{2}\mathcal{H}_{mm}$. This is the equation derived by Durbin and Speziale [2] which proves that the assumption of local isotropy is formally inconsistent with the Navier–Stokes equation on the second-order moment level when mean straining is imposed, i.e., when $i \neq j$. Clearly, the imposition of buoyancy does not render the local isotropy assumption formally invalid on the second-order velocity-moment level. It should further be noted that the implicit dependence on the stratification contained in the fluctuating pressure term in (2.23) does not contribute to the scalar $\mathcal{H}$ in incompressible flows.

Based on the theoretical arguments in the previous section, $\varepsilon_{i\theta}^{ISO} \neq 0$ in general. If we retain $\varepsilon_{i\theta}^{ISO} \neq 0$ and the assumption of local isotropy for rank 2 tensors, however, (3.17) becomes

$$d_t\varepsilon\delta_{ij} = \mathcal{H}\delta_{ij} - \frac{2}{5}\varepsilon S_{ij} - \underbrace{\frac{3}{2}\beta\left(g_i\varepsilon_{j\theta} + g_j\varepsilon_{i\theta}\right)}_{\mathcal{B}_{ij}}. \tag{3.18}$$

Equation (3.18) then provides us with another fact that strongly supports our assumption that $\varepsilon_{i\theta}^{ISO} \neq 0$ should be true; it implies that the rate of decay of ε, in the absence of mean shear ($S_{ij} = 0$), *should be unaffected by any imposed density stratification if the small-scale turbulence were truly isotropic*. However, there is no numerical or experimental evidence that this should be the case! On the contrary, it has been observed that even the slightest effect of buoyancy significantly alters the evolution of ε ($d_t\varepsilon$); see, e.g., Thoroddsen and Van Atta [16].

In order to provide an estimate of the nonlinear term $\mathcal{H}$ in (3.18), let us consider decaying grid turbulence unaffected by mean straining and stratification. The evolution equation for the turbulent time scale k/ε is readily obtained by combining (2.13) and (3.17). The results reads

$$d_t\left(\frac{k}{\varepsilon}\right) = -\left(1 + \frac{k}{\varepsilon^2}\mathcal{H}\right). \tag{3.19}$$

There exists experimental evidence that grid-generated turbulent kinetic energy exhibits a power-law decay, i.e., $k \sim t^{-n}$, where the decay exponent is $n \approx 1.3$ in a large number of measurements reported in the literature; cf., e.g., [7]. The value

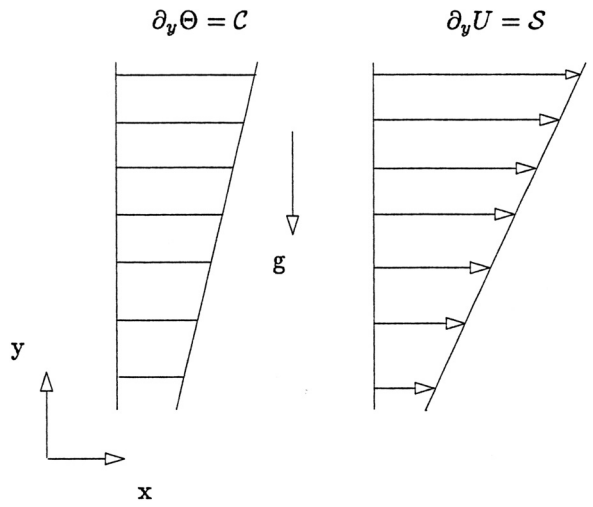


FIG. 3.1. *Homogeneous shear flow.*

of the decay exponent is reported to increase to $n \approx 5/2$ in the final period of decay. The power-law behavior of k requires $\varepsilon \sim t^{-(n+1)}$, and henceforth $k/\varepsilon \sim t$ and $d_t(k/\varepsilon) \sim O(1)$. With this, equation (3.19) provides the estimate

$$\mathcal{H} = \mathcal{O}\left(\frac{\varepsilon^2}{k}\right), \tag{3.20}$$

which is widely used, almost without exception, by turbulence modelers.

Let us now consider homogeneous shear flow with $\partial_y U = \mathcal{S} > 0$, $\partial_y \Theta = \mathcal{C} > 0$, and $\mathbf{g} = [0, -g, 0]$; see Figure 3.1. The assumption of local isotropy, in terms of an imposed density stratification, would then be a formally justified approximation if we can neglect $\mathcal{B}$ as compared to $\mathcal{H}$ in (3.18), i.e., if $||\mathcal{B}|| \ll ||\mathcal{H}||$ or equivalently if

$$k\mathcal{C} \gg |Ri_g \varepsilon_{2\theta}| \left(\frac{\mathcal{S}k}{\varepsilon}\right)^2 \tag{3.21}$$

by using the estimate (3.20). The gradient Richardson number $Ri_g \equiv \mathcal{N}^2/\mathcal{S}^2$, and $\mathcal{N}^2 \equiv \beta g \mathcal{C}$ is the Brunt–Väisala frequency. If we consider a flow close to equilibrium, it is reasonable to assume that $\mathcal{P}_{2\theta}/\varepsilon_{2\theta} = \mathcal{O}(1)$ in (2.16), where $\mathcal{P}_{2\theta} \equiv -\tau_{22}\mathcal{C} = -\frac{2}{3}k\mathcal{C}$. The last equality is obtained by substituting the isotropic value $\tau_{22} = \frac{2}{3}k$. Equation (3.21) can then be written as

$$|Ri_g| = \left|\frac{\mathcal{N}^2}{\mathcal{S}^2}\right| \ll \left(\frac{\varepsilon}{\mathcal{S}k}\right)^2 = \mathcal{O}(0.1). \tag{3.22}$$

The right-hand side of (3.22) has been evaluated using $\mathcal{S}k/\varepsilon \sim 6$, which typically is reached in physical and numerical experiments of homogeneous shear flows near equilibrium [15]. The constraint (3.22) thus implies that local isotropy constitutes a justifiable approximation only at very small Richardson numbers, in fact, so small that buoyancy effect cannot essentially be present in practice. The inequality also suggests that the imposition of density stratification exerts a significantly stronger effect on the dissipative scales than an imposed mean straining.

Durbin and Speziale [2] further demonstrated, in the absence of density stratification, that

$$\frac{\mathcal{S}k}{\varepsilon} \ll \mathcal{O}(1) \tag{3.23}$$

is a necessary condition for local isotropy to constitute a formally justified approximation in the absence of density stratification. This relation is readily obtained by requiring $||\mathcal{H}|| \gg ||\varepsilon\mathcal{S}||$ in (3.17). Using this and (3.22) yields the combined constraint

$$\left|\frac{\mathcal{N}^2k^2}{\varepsilon^2}\right| \ll \frac{\mathcal{S}^2k^2}{\varepsilon^2} \ll \mathcal{O}(1). \tag{3.24}$$

This result implies that the time scales associated with buoyancy and mean shear must be much larger than the integral turbulent time scale in order for the local isotropy hypothesis to constitute a formally justified approximation. In the absence of mean straining, the magnitude of the Brunt–Väisala frequency is thus required to be much smaller than that of the integral scale turbulent frequency in order for the hypothesis to constitute a physically plausible approximation. This is not feasible in homogeneous flows, at least for flows relatively close to equilibrium.

We can also recast (3.22) in terms of the buoyancy and shear Reynolds numbers frequently used in the literature,

$$Re_B = \left|\frac{\varepsilon}{\nu N^2}\right| = \frac{\varepsilon}{\nu|\beta g\mathcal{C}|} \qquad \text{and} \qquad Re_S = \frac{\varepsilon}{\nu\mathcal{S}^2}, \tag{3.25}$$

by noting that $Ri_g = Re_S/Re_B$. The result (3.24) can then be written as

$$\frac{1}{Re_B} \ll \frac{1}{Re_S} \ll \frac{1}{Re}, \tag{3.26}$$

where $Re \equiv k^2/(\varepsilon\nu)$ is the integral scale turbulent Re.

3.3. Third-order velocity-moments. Let us finally focus our attention on the evolution of the third-order dissipation rate tensor (ε_{ijk}). If small-scale turbulence on the third-order velocity-moment level were truly isotropic, the evolution equation (2.15) must be fully consistent with the mathematical properties (3.4). The terms $\mathcal{P}_{ijk}$, $\mathcal{U}_{ijk}$, and $\mathcal{N}_{ijk}$ are all third-rank tensors of the fluctuating velocity field. Since these terms must be symmetric for any permutation of indices, it follows from (3.6) that

$$\varepsilon_{ijk}^{ISO} = \mathcal{P}_{ijk}^{ISO} = \mathcal{U}_{ijk}^{ISO} = \mathcal{N}_{ijk}^{ISO} = 0 \qquad \forall \;\; i,j,k. \tag{3.27}$$

The term $\mathcal{G}_{ijk} \sim g_i d_{j\theta k}$ differs from the other terms in (2.24) in that it only compromises a second-rank tensor of fluctuating quantities, i.e., $d_{j\theta k}$. The most general isotropic form of $d_{j\theta k}$ can then be written as

$$d_{j\theta k}^{ISO} = \frac{1}{3}\overline{u_m\varepsilon'_{m\theta}}\delta_{jk} \neq 0, \tag{3.28}$$

according to (3.2). The result on the third-order velocity-moment level shows that the assumption of local isotropy is formally consistent with Navier–Stokes equations on the third-moment level if and only if $\overline{u_m\varepsilon'_{m\theta}} \equiv 0$. This requirement is, on the other hand, generally not fulfilled.

4. Concluding remarks. The present study has demonstrated that the hypothesis of local isotropy is formally inconsistent with the Navier–Stokes equations in homogeneous stratified turbulence, irrespective of whether the stratification is stable or not. The imposition of a mean temperature gradient is shown to essentially affect the small-scale turbulence in the same manner as an imposed mean shear, but with a significantly stronger impact.

George [4] has suggested, based on experimental findings, that the small-scale motion remains closely linked to the large-scale coherent motion. Anisotropies of the large scales would thus be reflected over the entire spectral range. These findings are consistent with the results presented herein.

The outcome of the present analysis is also very similar to the findings of Yeung, Brasseur, and Wang [22], although their approach is rather different. They considered the effect of anisotropic large-scale turbulence on the small-scale anisotropy, whereas the present study focuses on the imposition mean-flow anisotropy. Despite this difference, both cases reach the same conclusion, namely, that the imposition of large-scale anisotropy, be it related to turbulence or the mean flow, does not show up on all levels of small-scale velocity-moments.

In particular, density stratification does not formally conflict with the local isotropy hypothesis on the second-order level, whereas it shows up for the first- and third-order correlations examined here. Similarly, mean shear does not formally conflict with the isotropy assumption on the first- and third-order levels, whereas it is formally inconsistent on the second-order level. It is, however, sufficient to show anisotropy on *any* small-scale statistics in order for the local isotropy hypothesis to be violated. This was pointed out by Yeung, Brasseur, and Wang [22], who also argued that the converse is not true; a single statistical measure that displays a state of local isotropy is a necessary but not a sufficient condition to guarantee small-scale isotropy.

A qualitative analysis of the second-order dissipation rate transport equation has indicated that local isotropy constitutes a physically justifiable approximation, at this particular level of single-point moments, only if the imposed time scale associated with buoyancy, or mean straining, is much larger than the integral turbulent time scale. It can therefore be concluded that local isotropy does not seem to be a physically plausible argument in flows relatively close to equilibrium, since the imposed and the eddy turnover time scales usually are of the same order. A successful continuation in the development of predictive methods for turbulent flows relies heavily upon the ability to characterize small-scale turbulence in terms of the large scales. The theoretical outcome of this study has shown that it seems necessary to include information of the *mean* flow field in models for the small-scale turbulence in order to retain some consistency with the Navier–Stokes equations.

Acknowledgments. The authors wish to thank Prof. P. A. Durbin (Stanford University) and Dr. J. Werne (Colorado Research Associates, Div Boulder) for commenting on the draft.

REFERENCES

[1] J. Brasseur and C.-H. Wei, *Interscale dynamics and local isotropy in high Reynolds number turbulence within triadic interactions*, Phys. Fluids, 6 (1994), pp. 842–870.

[2] P. A. Durbin and C. G. Speziale, *Local anisotropy in strained turbulence at high Reynolds numbers*, ASME J. Fluids Engrg., 113 (1991), pp. 707–709.

[3] U. Frisch, *Turbulence: The Legacy of A.N. Kolmogorov*, Cambridge University Press, Cambridge, UK, 1995.

[4] W. K. George, *Self-preservation, and its relation to initial conditions, and coherent structures*, in Advances in Turbulence, W. K. George and R. Arndt, eds., Hemisphere, New York, 1998 pp. 1–23.
[5] W. K. George and H. J. Hussain, *Locally axisymmetric turbulence*, J. Fluid Mech., 233, pp. 1–23.
[6] A. N. Kolmogorov, *The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers*, C. R. Acad. Sci. URSS, 30 (1941), pp. 301–305.
[7] M. S. Mohamed and J. C. LaRue, *The decay power law in grid-generated turbulence*, J. Fluid Mech., 219 (1990), pp. 195–214.
[8] A. S. Monin and A. M. Yaglom, *Statistical fluid mechanics*, Vol. 2, MIT Press, Cambridge, MA, 1971.
[9] M. Nelkin, *Universiality and scaling in fully developed turbulence*, Adv. Phys., 43 (1994), pp. 143–181.
[10] A. M. Obukhov, *Structure of the temperature field in turbulent flow*, Isv. Akad. Nauk SSSR, Ser. Geogr. Geophys., 13 (1949), pp. 58–69.
[11] J. Rotta, *Statistische theorie nichhomogener turbulenz*, Z. Phys., 131 (1959), pp. 51–62.
[12] X. Shen and Z. Warhaft, *The anisotropy of the small scale structure in high Reynolds number ($R_\lambda \sim 1000$) turbulent shear flow*, Phys. Fluids, 12 (2000), pp. 2976–2989.
[13] K. R. Sreenivasan, *On local isotropy of passive scalars in turbulent shear flows*, Proc. Roy. Soc. London Ser. A, 434 (1991), pp. 165–182.
[14] K. R. Sreenivasan and R. A. Antonia, *The phenomenology of small-scale turbulence*, Annu. Rev. Fluid Mech., 29 (1997), pp. 435–472.
[15] S. Tavoularis and S. Corrsin, *Experiments in nearly homogeneous turbulent shear flow with a uniform mean temperature gradient. Part* 1, J. Fluid Mech., 104 (1981), pp. 311–347.
[16] S. Thoroddsen and C. W. Van Atta, *The influence of stable stratification on small-scale anisotropy and dissipation in turbulence*, J. Geophys. Res., 97 (1992), pp. 3647–3658.
[17] A. A. Townsend, *The uniform distortion of homogeneous turbulence*, Quart. J. Mech. Appl. Math., 28 (1959), pp. 104–127.
[18] M. S. Uberoi, *Equipartitioning of energy and local isotropy in turbulent flows*, J. Appl. Phys., 28 (1957), pp. 1165–1170.
[19] C. Van Atta, *Local isotropy of the smallest scales of turbulent scalar and velocity fields*, Proc. Roy. Soc. London Ser. A, 434 (1991), pp. 139–147.
[20] Z. Warhaft, *Passive scalars in turbulent flows*, Annu. Rev. Fluid Mech., 32 (2000), pp. 203–240.
[21] J. Werne and D. C. Fritts, *Anisotropy in stratified shear layer*, Phys. Chem. Earth (B), 26 (2001), pp. 263–268.
[22] P. K. Yeung, J. G. Brasseur, and Q. Wang, *Dynamics of direct large-small scale couplings in coherently forced turbulence: Concurrent physical- and Fourier-space views*, J. Fluid Mech., 283 (1995), pp. 43–95.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 322–342

RESONANCE AND BOUND STATES IN PHOTONIC CRYSTAL SLABS*

STEPHEN P. SHIPMAN† AND STEPHANOS VENAKIDES‡

Abstract. Using boundary-integral projections for time-harmonic electromagnetic (EM) fields, and their numerical implementation, we analyze EM resonance in slabs of two-phase dielectric photonic crystal materials. We characterize *resonant frequencies* by a complex Floquet–Bloch dispersion relation $\omega = W(\beta)$ defined by the existence of a nontrivial nullspace of a pair of boundary-integral projections parameterized by the wave number β and the time-frequency ω. At resonant frequencies, the crystal slab supports a source-free EM field. We link *complex* resonant frequencies, where the imaginary part is small, to resonant scattering behavior of incident source fields at nearby real frequencies and anomalous transmission of energy through the slab. At a *real* resonant frequency, the source-free field supported by the slab is a *bound state*. We present numerical examples which demonstrate the effects of structural defects on the resonant properties of a crystal slab and surface waves supported by a dielectric defect.

Key words. photonic crystal, boundary integral, Calderón's projection, resonance, dispersion relation, bound state, scattering, surface wave

AMS subject classifications. 78, 45, 65

DOI. 10.1137/S0036139902411120

1. Overview.

Photonic crystals are material structures with spatially periodic electromagnetic (EM) properties. A two-dimensional (2D) dielectric *photonic crystal slab* (Figure 1.1) has dielectric permittivity that does not vary in the z direction, is constant beyond some finite value of $|x|$, and is periodic in y. The photonic crystal slabs in our study consist of an array of circular homogeneous rods embedded in a matrix of a contrasting dielectric permittivity.

It is well known that photonic crystals may act as resonators. In previous work [1], [2], we investigated resonant behavior in photonic crystal slabs. In particular, a periodic channel defect in a slab whose period cell is shown in Figure 7.3(3b) below resulted in the appearance of narrow ranges of frequency values over which the steady-state field in the crystal exhibited amplitudes that were many times greater than the amplitude of the polarized, time and space harmonic incident EM source field [2]. The dielectric materials that we consider have no losses or gains; thus the phenomenon is due solely to resonant behavior in the scattering by the crystal. Over these narrow frequency ranges, the transmission of energy through the slab is either enhanced or inhibited, producing "spikes" in the transmission coefficient (as in Figures 7.1(5) and 7.3(3a) below). The exploitation of photonic crystal resonances in the engineering of photonic devices has received much attention in the literature in recent years. See, for example, [3] and [4] for applications to filters and transmission enhancement. The connection between the structure of transmission dips and properties of quasi-guided eigenmodes is investigated in [5] for square-patterned slabs on a substrate.

*Received by the editors July 10, 2002; accepted for publication (in revised form) March 25, 2003; published electronically November 19, 2003.
http://www.siam.org/journals/siap/64-1/41112.html
†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (shipman@math.lsu.edu).
‡Department of Mathematics, Box 90320, Duke University, Durham, NC 27708-0320 (ven@math.duke.edu). The research of this author was supported by grants ARO-DAAD19-99-1-0132 and NSF DMS-0207262.

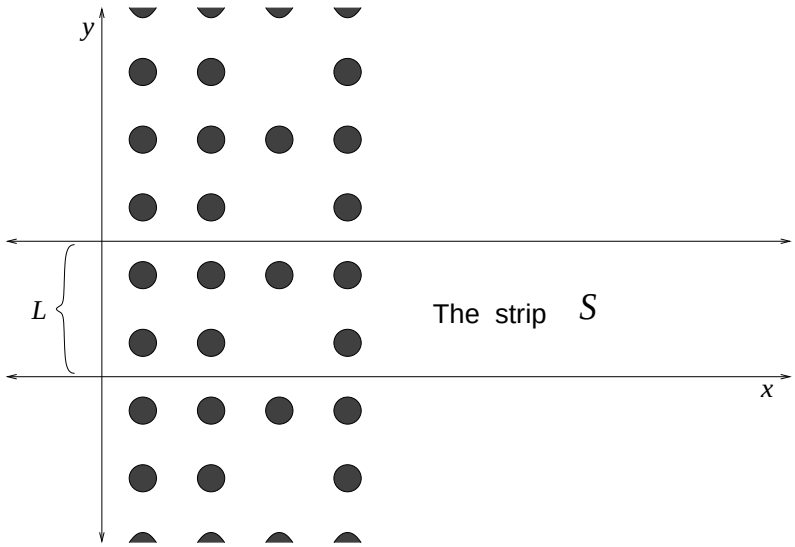


FIG. 1.1. *A cross section of a photonic crystal slab consisting of an array of homogeneous dielectric rods standing perpendicular to the xy-plane. The rod structure is periodic in the y-direction, with period L, and extends indefinitely as* $y \to \pm\infty$. *The rod structure is finite in the x-direction. Exterior to the rods is a homogeneous material of contrasting dielectric permittivity extending to infinity to the right and left. Pseudoperiodic fields in the plane can be analyzed in the strip* $\mathcal{S} = \{(x, y) : -\infty < x < \infty,\ 0 \le y \le L\}$ *consisting of a single period of the dielectric permittivity function.*

A connection between resonant frequencies and proper eigenvalues is known for Helmholtz resonators. Beale [6] shows that the (complex) resonant frequencies of a cavity in $\mathbb{R}^3$ with an opening converge to the eigenvalues (bound state frequencies) of the closed cavity as the opening disappears.

In the present study, we link resonant scattering behavior in dielectric photonic crystal slabs to certain complex frequencies with a small imaginary part at which the structure supports a source-free field (the term "source" is defined precisely in section 4). These are Bloch fields $\psi(x, y) = \tilde{\psi}(x, y)e^{i(\beta y - \omega t)}$, with $\tilde{\psi}(x, y)$ periodic in y and real wave number β. If a source-free field exists for a real value of ω, the structure sustains a traveling or standing wave along the slab that decays exponentially as $|x| \to \infty$ so that the slab acts as a waveguide. We call such a field a *localized field* or a *bound state*, localization being in the strip $\mathcal{S} = \{(x, y) : -\infty < x < \infty,\ 0 \le y \le L\}$, consisting of one period of the dielectric permittivity function (Figure 1.1). Frequencies at which the crystal slab supports a source-free field are called *resonant frequencies*, and they are described by a *dispersion relation* $\omega = W(\beta)$. We take $\Re(\omega) > 0$ and prove that $\Im(W(\beta)) \le 0$ with equality if and only if the corresponding source-free field is a bound state. If $\Im(W(\beta)) < 0$, then the field grows exponentially as $|x| \to \infty$, but decays in time.

In numerical experiments with several structures, we find isolated values of the wave number β for which the frequency $\omega = W(\beta)$ appears to be real, giving rise to an isolated bound state. At nearby values of β, $W(\beta)$ attains an imaginary part, and sources at real frequencies near $\Re(W(\beta))$ produce resonant scattering behavior. At these frequencies, the transmission coefficient exhibits anomalous behavior (see Figure 7.1). We also find a wave number range over which we show that the imaginary part

of the frequency is exactly zero (see Figure 7.2). Indeed, a slab nine rods thick yields a real branch of the dispersion relation; all but the first rod have equal radii, and the radius of the first rod is larger. The wave modes sustained by this structure are mainly supported on the larger rod, demonstrating that waves can exist on a defective surface of an otherwise perfect crystal.

Following the Floquet–Bloch theory, our mathematical investigation restricts the analysis to a single period of the dielectric permittivity considered as a function of x and y. Thus the problem is posed on the strip $\mathcal{S} = \{(x, y) : -\infty < x < \infty,\ 0 \le y \le L\}$, where L is the period in the y direction, as illustrated in Figure 1.1. Pseudoperiodic boundary conditions then apply to the fields: $\psi(x, L) = e^{i\beta L}\psi(x, 0)$ and $\partial_y\psi(x, L) = e^{i\beta L}\partial_y\psi(x, 0)$ for all values of x.

We use boundary-integral projections of Calderón's type with pseudoperiodic Green's functions and Green's identities on the strip $\mathcal{S}$. In sections 3 and 4 we show how these projections give rise to a system of two coupled integral equations that relate the trace of the steady-state field and its normal derivative on the boundaries of the rods to the trace of the source field and its normal derivative. The latter fields constitute the forcing. The system is Fredholm of the second kind in the former fields; that is, the integral operator involved is a compact perturbation of the identity.

The existence of a resonant frequency requires the existence of a nullspace of the boundary-integral operator; the latter depends parametrically on the dielectric structure and the parameters β (which we always assume to be real) and ω. To locate resonant frequencies in section 7, we discretize the integral operator and search for (β, ω) pairs for which it has an eigenvalue equal to zero. In this way, we calculate numerically the dispersion relations $\omega = W(\beta)$. In some cases, ω is a real function of β, so that the relation describes how the frequency of an x-localized wave traveling along the slab depends on its Bloch wave number. In other cases, ω is a complex function of β. We prove that, in this case, the corresponding fields become unbounded as $|x| \to \infty$ (they decay as $t \to \infty$), and therefore do not represent bound states. They do, however, force nearby real frequencies to exhibit resonant behavior. This phenomenon is examined also in [5] using a scattering matrix for square-patterned slabs.

2. Free pseudoperiodic Green's functions. We consider a lossless photonic crystal that consists of an array of dielectric rods, each with the same constant dielectric coefficient $\epsilon_1 > 0$, embedded in a matrix of a material with some other constant dielectric coefficient $\epsilon_0 > 0$. The rods stand perpendicular to the xy-plane and do not vary with z. The array is truncated to a finite width in the x-direction and extends periodically in the y-direction, with period 2π. Its planar cross section consists of a finite union D of planar domains D_j (the cross sections of the rods) with C^2 boundaries in the strip $\mathcal{S} = \{(x, y) : 0 < y < 2\pi\}$ that repeats periodically in the y-direction. We let $\partial\mathcal{S}$ have an inward-pointing normal vector, and we let $n(\mathbf{r})$ denote the outward-directed normal vector to the boundary ∂D of D.

Let $\psi(x, y)e^{-i\omega t}$ be the out-of-plane component (electric or magnetic) of a polarized time-harmonic electromagnetic field with nondimensionalized frequency[1] ω in the photonic crystal structure (it is constant in the z-direction). The Maxwell equations then reduce to the Helmholtz equation

$$\nabla^2\psi + \epsilon_1\omega^2\psi = 0 \quad (\text{in } D), \qquad \nabla^2\psi + \epsilon_0\omega^2\psi = 0 \quad (\text{in } \mathcal{S} \setminus \bar{D})$$

[1] We use ω for the reduced time-frequency in this paper; it is equal to our k in [1] and [2]. It relates to the physical frequency Ω (cycles per time) and period L by $\omega = \Omega L/c$, where c is the speed of light.

and the matching conditions

$$\lim_{h\to 0}\left(\psi(\mathbf{r}-hn(\mathbf{r}))-\psi(\mathbf{r}+hn(\mathbf{r}))\right)=0, \tag{2.1a}$$

$$\lim_{h\to 0}\left(\frac{\partial\psi(\mathbf{r}-hn(\mathbf{r}))}{\partial n(\mathbf{r})}-\nu\frac{\partial\psi(\mathbf{r}+hn(\mathbf{r}))}{\partial n(\mathbf{r})}\right)=0, \tag{2.1b}$$

on the boundary ($\mathbf{r}\in\partial D$). In these equations, $\nabla^2=\partial^2/\partial x^2+\partial^2/\partial y^2$, $\nu=1$ in the electric polarization case, and $\nu=\epsilon_1/\epsilon_0$ in the magnetic polarization case.

We will consider Helmholtz fields ψ that are pseudoperiodic in y. This means that, for some real number β, $\psi=e^{i\beta y}\tilde{\psi}$, where $\tilde{\psi}$ is periodic in y with the same period (2π) as the crystal.

First, we present the fundamental pseudoperiodic solutions of the Helmholtz equation.

THEOREM 2.1. *Let ϵ and β be real numbers and ω a complex number such that, for all integers m, $\epsilon\omega^2-(m+\beta)^2\neq 0$. For each integer m, let μ_m be defined by*

$$\mu_m^2-(m+\beta)^2+\epsilon\omega^2=0,$$

with $\Re(\mu_m)<0$ for all but a finite number of values of m. Then the series

$$G(\mathbf{r})=-\frac{1}{4\pi}\sum_{m=-\infty}^{\infty}\frac{1}{\mu_m}\exp\left(\mu_m|x|+i(m+\beta)y\right)$$

converges and is of class C^∞ for all $\mathbf{r}=(x,y)\in\mathbb{R}^2\setminus\{(0,2\pi n):n\in\mathbb{Z}\}$ and

$$\nabla^2G+\epsilon\omega^2G=-\sum_{n=-\infty}^{\infty}\delta(x,y-2\pi n)e^{2\pi ni\beta},$$

where δ is the Dirac delta-function in $\mathbb{R}^2$ with unit impulse at the origin.

Proof. We first consider the case in which $\Re(\mu_m)\leq 0$ for all $m\in\mathbb{Z}$. Then the series defining $G(\mathbf{r})$ converges to a tempered distribution on $\mathbb{R}^2$. This is seen as follows: Let $\phi(x,y)$ be a function of Schwartz class. Then, for m large enough so that $\Re(\mu_m)<0$, we have

$$\left|\iint\frac{1}{\mu_m}\exp(\mu_m|x|+i(m+\beta)y)\phi(x,y)\,dx\,dy\right| \leq\frac{1}{|\mu_m|}\int e^{\Re(\mu_m)|x|}\int|\phi(x,y)|\,dy\,dx\leq\frac{-2A}{|\mu_m|\Re(\mu_m)}, \tag{2.2}$$

where A is such that

$$\sup_{(x,y)\in\mathbb{R}^2}(1+y^2)|\phi(x,y)|<A\left(\int(1+z^2)^{-1}dz\right)^{-1}.$$

Since $\Re(\mu_m)/m=\mathcal{O}(1)$ as $m\to\infty$, the series for $G(\mathbf{r})$ converges to a tempered distribution.

Let $\widehat{G}(\mathbf{s})$, where $\mathbf{s}=(s_1,s_2)$, be the Fourier transform of $G(\mathbf{r})$:

$$\widehat{G}(\mathbf{s})=-\frac{1}{4\pi}\sum_{m=-\infty}^{\infty}\frac{-2}{s_1^2+\mu_m^2}\delta(s_2-(m+\beta));$$

$$[(\nabla^2+\epsilon\omega^2)G]\hat{}(\mathbf{s}) = \frac{1}{2\pi}\sum_{m=-\infty}^{\infty}\frac{-s_1^2-(m+\beta)^2+\epsilon\omega^2}{s_1^2+\mu_m^2}\delta(s_2-(m+\beta)) \tag{2.3}$$
$$= -\frac{1}{2\pi}\sum_{m=-\infty}^{\infty}\delta(s_2-(m+\beta))$$

$$\implies \quad (\nabla^2+\epsilon\omega^2)G = -\frac{1}{2\pi}\delta(x)\sum_{m=-\infty}^{\infty}e^{i(m+\beta)y}$$
$$= -\frac{1}{2\pi}\delta(x)e^{i\beta y}\sum_{m=-\infty}^{\infty}e^{imy} = -\delta(x)e^{i\beta y}\sum_{n=-\infty}^{\infty}\delta(y-2\pi n) \tag{2.4}$$
$$= -\sum_{n=-\infty}^{\infty}e^{2\pi n i\beta}\delta(x,y-2\pi n).$$

The result holds if we replace $1/\mu_m \exp(\mu_m|x| + i(m+\beta)y)$ in $G(\mathbf{r})$ with $-1/\mu_m \exp(-\mu_m|x| + i(m+\beta)y)$ for a finite number of integers m, because this amounts to adding a finite number of functions

$$\psi_m = \frac{1}{\mu_m}(e^{-\mu_m|x|}+e^{\mu_m|x|})e^{i(m+\beta)y} = \frac{1}{\mu_m}(e^{-\mu_m x}+e^{\mu_m x})e^{i(m+\beta)y},$$

which satisfy $(\nabla^2+\epsilon\omega^2)\psi_m = 0$.

Finally, we note that the ellipticity of $\nabla^2+\epsilon\omega^2$ implies that any distribution solution on a domain must be a function of class C^∞. Thus $G(\mathbf{r})$ is of class C^∞ on $\mathbb{R}^2 \setminus \{(0,2\pi n) : n \in \mathbb{Z}\}$. □

For our purposes, we make the following choices of the sign of μ_m. For real values of ω (such that $(m+\beta)^2 - \epsilon\omega^2 \neq 0$ for all integers m), we choose μ_m such that $G(\hat{\mathbf{r}})$ is a *radiating* Green's function. Thus, for the finite number of consecutive integers m such that $(m+\beta)^2 - \epsilon\omega^2 < 0$ (it is possible that there are no such values of m), we take μ_m to lie on the positive imaginary axis; these values of m give the finite number of outwardly *propagating* modes. For all other values of m, we take $\mu_m < 0$; these give the *decaying* modes.

In our investigations, we will consider continuous perturbations of ω into the lower-half complex plane, and we allow the values of μ_m to vary analytically with ω. As ω attains a negative imaginary part, the finite number of values of μ_m that gave the propagating modes now attain a positive real part and therefore grow as $|x| \to \infty$. For all other values of m, μ_m attains a negative imaginary part.

Figure 2.1 shows the number of propagating modes for real values of β and ω. Pairs (β,ω) for which there are no propagating modes and perturbations of these in the imaginary ω direction admit no scattering (extended) EM fields in the strip $\mathcal{S}$.

We will make use of the set of Helmholtz fields which, to the right and left of the scatterer, are equal to a superposition of modes that build the Green's functions G that we defined in Theorem 2.1.

DEFINITION 2.2. *We say that a function ψ is in the class $\mathcal{E}(\beta,\omega,\epsilon)$ if its domain contains $\{(x,y) : |x| > x_*\}$ for some $x_* > 0$ and*

$$\psi(x,y) = \sum_{m=-\infty}^{\infty} A_m^{\pm}\exp\left(\mu_m|x| + i(m+\beta)y\right) \quad \textit{for} \quad \pm x > x_*$$

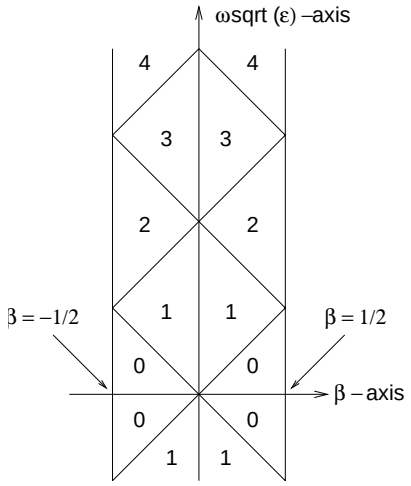


FIG. 2.1. *The number of propagating modes for real values of β and ω. The pattern is repeated periodically in β, with period 1. For fixed β and ω, there is a propagating mode for each m such that $|\sqrt{\epsilon}\omega| > |m + \beta|$.*

for some complex numbers $A_m^{\pm}$, where $\mu_m = \left((m+\beta)^2 - \epsilon\omega^2\right)^{1/2}$ with the sign determination described above.

Remark. The role of the class $\mathcal{E}(\beta, \omega, \epsilon)$ is to continue analytically into the complex ω-plane as $|x| \to \infty$ the standard outgoing radiation condition that applies when ω is real: $\mathcal{E}(\beta, \omega, \epsilon)$ contains the radiating fields when ω is real and their analytic continuations into the complex ω-plane.

3. The boundary-integral projections. Let $\xi \in H^1(\partial D)$ and $\eta \in L^2(\partial D)$ be given (H^1 is the linear space of functions on ∂D with square-integrable arclength derivatives), and denote $\boldsymbol{\xi} = (\xi, \eta)^t$. For any point $\hat{\mathbf{r}}$ in the strip $\mathcal{S}$ exterior to D, define

$$\psi(\hat{\mathbf{r}}) = \int_{\partial D} \left(\frac{\partial G(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\mathbf{r})} \xi(\mathbf{r}) - G(\hat{\mathbf{r}} - \mathbf{r}) \eta(\mathbf{r}) \right) ds(\mathbf{r}) \qquad (\hat{\mathbf{r}} \text{ exterior to } D). \tag{3.1}$$

This field is an element of $\mathcal{E}(\beta, \omega, \epsilon)$. For $\hat{\mathbf{r}} \in D$, define

$$\psi(\hat{\mathbf{r}}) = \int_{\partial D} \left(-\frac{\partial G(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\mathbf{r})} \xi(\mathbf{r}) + G(\hat{\mathbf{r}} - \mathbf{r}) \eta(\mathbf{r}) \right) ds(\mathbf{r}) \qquad (\hat{\mathbf{r}} \in D). \tag{3.2}$$

Both fields satisfy the Helmholtz equation (in the variable $\hat{\mathbf{r}}$) in their respective domains.

Let $\hat{\mathbf{r}}$ now be a point on ∂D, and consider the limits to $\hat{\mathbf{r}}$ of these exterior and interior fields and their normal derivatives:

$$\boldsymbol{\psi}_e(\hat{\mathbf{r}}) = \begin{bmatrix} \psi_e(\hat{\mathbf{r}}) \\ \\ \partial_n \psi_e(\hat{\mathbf{r}}) \end{bmatrix} := \begin{bmatrix} \lim\limits_{h \to 0^+} \psi(\hat{\mathbf{r}} + hn(\hat{\mathbf{r}})) \\ \\ \lim\limits_{h \to 0^+} \dfrac{\partial \psi}{\partial n(\hat{\mathbf{r}})}(\hat{\mathbf{r}} + hn(\hat{\mathbf{r}})) \end{bmatrix},$$

$$\boldsymbol{\psi}_i(\hat{\mathbf{r}}) = \begin{bmatrix} \psi_i(\hat{\mathbf{r}}) \\ \\ \partial_n \psi_i(\hat{\mathbf{r}}) \end{bmatrix} := \begin{bmatrix} \lim\limits_{h \to 0^+} \psi(\hat{\mathbf{r}} - hn(\hat{\mathbf{r}})) \\ \\ \lim\limits_{h \to 0^+} \dfrac{\partial \psi}{\partial n(\hat{\mathbf{r}})}(\hat{\mathbf{r}} - hn(\hat{\mathbf{r}})) \end{bmatrix}.$$

These limits are again in $H^1(\partial D) \times L^2(\partial D)$. See the appendix for a more detailed discussion of these limits.

The limit of the first-order normal derivatives of the Green's function produces singular contributions according to the Plemelj formula, and we obtain

$$\boldsymbol{\psi}_e = \tfrac{1}{2}(I + H)\boldsymbol{\xi},$$
$$\boldsymbol{\psi}_i = \tfrac{1}{2}(I - H)\boldsymbol{\xi},$$

in which H is an integral operator from $H^1(\partial D) \times L^2(\partial D)$ to itself defined by

(3.3)
$$(H\boldsymbol{\xi})(\hat{\mathbf{r}}) = 2 \begin{bmatrix} \displaystyle\int_{\partial D} \left(\frac{\partial G(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\mathbf{r})} \xi(\mathbf{r}) - G(\hat{\mathbf{r}} - \mathbf{r}) \eta(\mathbf{r}) \right) ds(\mathbf{r}) \\ \displaystyle\lim_{h \to 0} \int_{\partial D} \frac{\partial G(\hat{\mathbf{r}} + h\, n(\hat{\mathbf{r}}) - \mathbf{r})}{\partial n(\hat{\mathbf{r}}) \partial n(\mathbf{r})} \xi(\mathbf{r}) ds(\mathbf{r}) - \displaystyle\int_{\partial D} \frac{\partial G(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\hat{\mathbf{r}})} \eta(\mathbf{r}) ds(\mathbf{r}) \end{bmatrix}.$$

H has the form

$$H \begin{bmatrix} \xi \\ \eta \end{bmatrix} = 2 \begin{bmatrix} K & -J \\ L & -K' \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix},$$

in which the entries of the matrix are integral operators on ∂D. The integral kernels of J, K, and K' are $G(\hat{\mathbf{r}} - \mathbf{r})$, $\partial G(\hat{\mathbf{r}} - \mathbf{r})/\partial n(\mathbf{r})$, and $\partial G(\hat{\mathbf{r}} - \mathbf{r})/\partial n(\hat{\mathbf{r}})$, respectively. By integration by parts and using the identity

$$\frac{\partial^2 G(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\hat{\mathbf{r}}) \partial n(\mathbf{r})} + \frac{\partial^2 G(\hat{\mathbf{r}} - \mathbf{r})}{\partial s(\hat{\mathbf{r}}) \partial s(\mathbf{r})} = -n(\hat{\mathbf{r}}) \cdot n(\mathbf{r}) (\partial_x^2 + \partial_y^2) G(\hat{\mathbf{r}} - \mathbf{r}),$$

one can show that the integral kernel of L is

$$n(\hat{\mathbf{r}}) \cdot n(\mathbf{r}) \epsilon \omega^2 G(\hat{\mathbf{r}} - \mathbf{r}) + \frac{\partial G(\hat{\mathbf{r}} - \mathbf{r})}{\partial s(\hat{\mathbf{r}})} \frac{d}{ds},$$

where s is the arclength parameter and $\partial G/\partial s$ is a principal-value kernel (see the proof of Theorem 2.1 in [2]).

THEOREM 3.1. *H is a bounded linear operator from $H^1(\partial D) \times L^2(\partial D)$ into itself, the operators*

$$P_e = \tfrac{1}{2}(I + H),$$

$$P_i = \tfrac{1}{2}(I - H)$$

are complementary projections, and

$$H^2 = I.$$

Proof. Let us consider a solution ψ of $\nabla^2 \psi + \epsilon \omega^2 \psi = 0$ defined in D that has a continuous extension to $\overline{D}$ whose restriction to ∂D is in $H^1(\partial D)$ and such that $\lim_{h \to 0^-} \partial \psi / dn(\mathbf{r})(\mathbf{r} + hn(\mathbf{r}))$ exists and belongs to $L^2(\partial D)$. Then Green's identity holds: For $\hat{\mathbf{r}} \in D$,

$$\psi(\hat{\mathbf{r}}) = \int_{\partial D} \left(-\frac{\partial G(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\mathbf{r})} \psi(\mathbf{r}) + G(\hat{\mathbf{r}} - \mathbf{r}) \frac{\partial \psi(\mathbf{r})}{\partial n(\mathbf{r})} \right) ds(\mathbf{r}). \tag{3.4}$$

Green's identity says that (3.2) holds if we put ψ and $\partial_n\psi$ in place of ξ and η, and therefore

$$P_i^2(\boldsymbol{\xi}) = P_i(\boldsymbol{\psi_i}) = \boldsymbol{\psi_i} = P_i(\boldsymbol{\xi}),$$

and thus $P_i^2 = P_i$.

Now let a solution ψ of $\nabla^2\psi + \epsilon\omega^2\psi = 0$ such that $\psi \in \mathcal{E}(\beta, \omega, \epsilon)$ be defined in $\mathcal{S} \setminus \overline{D}$, and suppose that ψ has a continuous extension to $\overline{D}$ whose restriction to ∂D is in $H^1(\partial D)$ and such that $\lim_{h\to 0-} \partial\psi/dn(\mathbf{r})(\mathbf{r}+hn(\mathbf{r}))$ exists and belongs to $L^2(\partial D)$. Again, Green's identity holds: For $\hat{\mathbf{r}} \in \mathcal{S} \setminus \overline{D}$,

$$\psi(\hat{\mathbf{r}}) = \int_{\partial D} \left(\frac{\partial G(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\mathbf{r})} \psi(\mathbf{r}) - G(\hat{\mathbf{r}} - \mathbf{r}) \frac{\partial \psi(\mathbf{r})}{\partial n(\mathbf{r})} \right) ds(\mathbf{r}). \tag{3.5}$$

The contribution from the upper and lower sides of the strip $\mathcal{S}$ cancel because of the pseudoperiodicity of ψ and G^ϵ. Straightforward calculation shows that the contributions from vertical line segments truncating the strip on the left and right vanish identically as the points of truncation tend to infinity, because both G and ψ are in $\mathcal{E}(\beta, \omega, \epsilon)$. Again, we find that $P_e^2 = P_e$ by putting ψ and $\partial_n\psi$ in place of ξ and η in (3.1).

It is straightforward to calculate that $H^2 = I$. □

The operators in Theorem 3.1 are *boundary-integral projections of Calderón's type*. See Calderón [7], Seeley [8], and Ryaben'kii [9]. Nédélec [10] derives the Calderón's projections for the full harmonic Maxwell equations for a bounded domain in $\mathbb{R}^3$.

4. The Fredholm system of boundary-integral equations. We describe the EM scattering by a photonic crystal slab in terms of the decomposition given by the boundary-integral projections. We have seen that any pair of functions $(\xi, \eta)^t = \boldsymbol{\xi}$ in $H^1(\partial D) \times L^2(\partial D)$ can be expressed uniquely as the sum of the limiting values to ∂D of an interior Helmholtz field and its normal derivative and an exterior Helmholtz field in $\mathcal{E}(\beta, \omega, \epsilon)$ and its normal derivative.

DEFINITION 4.1. *An exterior-source field $\boldsymbol{\xi} = \boldsymbol{\phi}_i^{\epsilon_0}$ has zero exterior component in the above decomposition over the exterior medium ($\epsilon = \epsilon_0$); that is,*

$$P_e^{\epsilon_0}\boldsymbol{\phi}_i^{\epsilon_0} = 0, \quad P_i^{\epsilon_0}\boldsymbol{\phi}_i^{\epsilon_0} = \boldsymbol{\phi}_i^{\epsilon_0}.$$

An interior-source field $\boldsymbol{\xi} = \boldsymbol{\phi}_e^{\epsilon_1}$ has zero interior component in the above decomposition over the interior medium ($\epsilon = \epsilon_1$); that is,

$$P_i^{\epsilon_1}\boldsymbol{\phi}_e^{\epsilon_1} = 0, \quad P_e^{\epsilon_1}\boldsymbol{\phi}_e^{\epsilon_1} = \boldsymbol{\phi}_e^{\epsilon_1}.$$

In other words, when the sources are in the exterior, the source field extends from ∂D to a Helmholtz field over the medium ϵ_0 in the interior. Similarly, when the sources are in the interior, the source field extends from ∂D to a Helmholtz field of class $\mathcal{E}(\beta, \omega, \epsilon_1)$ over the medium ϵ_1 in the exterior.

Remark. In our numerical calculations of transmission in section 7, we use plane-wave source fields from the left at real frequencies ω; that is,

$$\psi_i^{\epsilon_0} = \exp\left(i\sqrt{\epsilon_0\omega^2 - (m+\beta)^2}\, x + i(m+\beta)y \right), \qquad \psi_e^{\epsilon_1} = 0.$$

A solution to the scattering problem has total exterior and interior fields ψ_{ext} and ψ_{int}, with traces (field and normal derivative) on ∂D given by

$$\boldsymbol{\psi}_{\text{ext}} = \boldsymbol{\psi}_e^{\epsilon_0} + \boldsymbol{\phi}_i^{\epsilon_0}, \qquad \boldsymbol{\psi}_{\text{int}} = \boldsymbol{\psi}_i^{\epsilon_1} + \boldsymbol{\phi}_e^{\epsilon_1}, \tag{4.1}$$

where $\boldsymbol{\psi}_i^{\epsilon_1}$ and $\boldsymbol{\psi}_e^{\epsilon_0}$ are the traces of the interior and exterior scattered fields. The field $\boldsymbol{\psi}_i^{\epsilon_1}$ extends to a Helmholtz field in the interior, and $\boldsymbol{\psi}_e^{\epsilon_0}$ extends to a Helmholtz field of class $\mathcal{E}(\beta,\omega,\epsilon_0)$ in the exterior. When ω is real, the exterior extension of $\boldsymbol{\psi}_e^{\epsilon_0}$ either decays or satisfies the outward radiation condition as $|x|\to\infty$.

The matching conditions at the interface of the two media are given by

$$\begin{bmatrix} \psi_{\rm int}(\hat{\mathbf{r}}) \\ \partial_n\psi_{\rm int}(\hat{\mathbf{r}}) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \nu \end{bmatrix} \begin{bmatrix} \psi_{\rm ext}(\hat{\mathbf{r}}) \\ \partial_n\psi_{\rm ext}(\hat{\mathbf{r}}) \end{bmatrix}. \tag{4.2}$$

We write the matching conditions in short form by defining $\Gamma = \left[\begin{smallmatrix} 1 & 0 \\ 0 & \nu \end{smallmatrix}\right]$ and $\boldsymbol{\psi} = \left[\begin{smallmatrix} \psi_{\rm ext}(\hat{\mathbf{r}}) \\ \partial_n\psi_{\rm ext}(\hat{\mathbf{r}}) \end{smallmatrix}\right]$ on ∂D and inserting them into (4.1) to obtain

$$\boldsymbol{\psi} = \boldsymbol{\psi}_e^{\epsilon_0} + \boldsymbol{\phi}_i^{\epsilon_0}, \qquad \Gamma\boldsymbol{\psi} = \boldsymbol{\psi}_i^{\epsilon_1} + \boldsymbol{\phi}_e^{\epsilon_1}. \tag{4.3}$$

We apply the projections $P_i^{\epsilon_0}$ and $P_e^{\epsilon_1}$ to the two equations, respectively,

$$P_i^{\epsilon_0}\boldsymbol{\psi} = \boldsymbol{\phi}_i^{\epsilon_0} \quad \text{and} \quad P_e^{\epsilon_1}\Gamma\boldsymbol{\psi} = \boldsymbol{\phi}_e^{\epsilon_1}. \tag{4.4}$$

DEFINITION 4.2. *By the* scattering problem *at* (β,ω)*, we mean the system of equations* (4.4) *with the source fields (see Definition* 4.1*) in the right-hand sides belonging to the space* $H^1(\partial D)\times L^2(\partial D)$*. A* scattering state *or* scattering field *is a Helmholtz field whose trace on* ∂D *is a solution of the system with a nonzero source. If there exists a nontrivial solution* $\boldsymbol{\psi}$ *in the absence of sources (zero right-hand sides), we call the frequency* ω *a* resonant frequency *(for the wave number* β*). The corresponding Helmholtz field is necessarily in* $\mathcal{E}(\beta,\omega,\epsilon_0)$*. If the field decays as* $|x|\to\infty$*, we call it a* bound state*.*

Remarks. 1. Our definition of scattering state analytically continues the traditional scattering states at real frequencies into the complex ω-plane. This is possible by our analytic continuation of the condition that defines the notion of an outgoing radiating field (see the Remark after Definition 2.2). A scattering state is an analytic function of ω, and a resonant frequency is a singularity of this function.

2. We say that a scattering field at a nonresonant real frequency near a resonant frequency exhibits *resonant behavior* if it exhibits amplitudes in the crystal structure that are large compared to the source amplitude.

3. We will need to consider only frequencies ω such that $\Re(\omega)>0$. We will prove in Theorems 5.1 and 5.2 that, in this case, a dielectric photonic crystal slab supports only Helmholtz fields with $\Im(\omega)\le 0$ and that such a field decays as $|x|\to\infty$ if $\Im(\omega)=0$, and is unbounded if $\Im(\omega)<0$.

We add the two equations (4.4) to obtain

$$(P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma)\boldsymbol{\psi} = \boldsymbol{\phi}_i^{\epsilon_0} + \boldsymbol{\phi}_e^{\epsilon_1}. \tag{4.5}$$

Inserting the expressions

$$P_i^{\epsilon_0} = \tfrac{1}{2}(I + H^{\epsilon_0}), \qquad P_e^{\epsilon_1} = \tfrac{1}{2}(I - H^{\epsilon_1})$$

from Theorem 3.1, and letting $\boldsymbol{\phi} = \boldsymbol{\phi}_i^{\epsilon_0} + \boldsymbol{\phi}_e^{\epsilon_1}$ represent the total source field, we rewrite (4.5) as

$$\left[\frac{1}{2}\,(I+\Gamma) + \frac{1}{2}\,(H^{\epsilon_0} - H^{\epsilon_1}\Gamma)\right]\boldsymbol{\psi} = \boldsymbol{\phi}.$$

Finally, we insert the expression (3.3) for H^{ϵ_0} and H^{ϵ_1} to obtain the following expanded version of (4.5) (recall $\boldsymbol{\psi} = \boldsymbol{\psi}_{ext}$):

$$\psi_{\rm ext}(\hat{\mathbf{r}}) + \int_{\partial D} \left[\frac{\partial \left(G^{\epsilon_1} - G^{\epsilon_0}\right)(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\mathbf{r})} \psi_{\rm ext}(\mathbf{r}) \right.$$
$$\left. - (\nu G^{\epsilon_1} - G^{\epsilon_0})(\hat{\mathbf{r}} - \mathbf{r}) \frac{\partial \psi_{\rm ext}}{\partial n(\mathbf{r})}(\mathbf{r}) \right] ds(\mathbf{r}) \;=\; \phi(\hat{\mathbf{r}}), \tag{4.6a}$$

$$\frac{1+\nu}{2} \frac{\partial \psi_{\rm ext}}{\partial n(\hat{\mathbf{r}})}(\hat{\mathbf{r}}) + \int_{\partial D} \left[\frac{\partial^2 \left(G^{\epsilon_1} - G^{\epsilon_0}\right)(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\hat{\mathbf{r}}) \partial n(\mathbf{r})} \psi_{\rm ext}(\mathbf{r}) \right.$$
$$\left. - \frac{\partial \left(\nu G^{\epsilon_1} - G^{\epsilon_0}\right)(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\hat{\mathbf{r}})} \frac{\partial \psi_{\rm ext}}{\partial n(\mathbf{r})}(\mathbf{r}) \right] ds(\mathbf{r}) \;=\; \frac{\partial \phi}{\partial n(\hat{\mathbf{r}})}(\hat{\mathbf{r}}). \tag{4.6b}$$

This is a Fredholm equation of the second kind; cancellation in the difference of the Green's functions reduces the leading singularities of all the kernels to at most logarithmic, making the corresponding operators of Hilbert–Schmidt class in $L^2(\partial D) \times L^2(\partial D)$. Equation (4.6) can thus be solved for $\boldsymbol{\psi}$ in $L^2(\partial D) \times L^2(\partial D)$ whenever the nullspace is trivial. The solution $\boldsymbol{\psi}$ necessarily belongs to $H^1(\partial D) \times L^2(\partial D)$; indeed, (4.6) can be written as $\boldsymbol{\psi} = M\boldsymbol{\psi} + 2(I + \Gamma)^{-1}\boldsymbol{\phi}$, where M consists of kernels with at most logarithmic singularities, and is thus bounded from $L^2(\partial D) \times L^2(\partial D)$ to $H^1(\partial D) \times H^1(\partial D)$. In fact, M is bounded from $H^s(\partial D) \times H^s(\partial D)$ to $H^{s+1}(\partial D) \times H^{s+1}(\partial D)$, and so $\boldsymbol{\psi}$ is C^∞ if $\boldsymbol{\phi} = 0$, i.e., if $\boldsymbol{\psi}$ is a nullfield. In summary, if the source field $\boldsymbol{\phi}$ belongs to $H^1(\partial D) \times L^2(\partial D)$, then so does the field $\boldsymbol{\psi}$.

Boundary integral equations of the second kind for EM fields have also been derived by Müller [11], Colton and Kress [12], [13], and Nédélec [10].

THEOREM 4.3. *If the system*

$$P_e^{\epsilon_0} \boldsymbol{f} = 0, \qquad P_i^{\epsilon_1} \boldsymbol{f} = 0$$

has only the trivial solution, then the scattering problem (4.4) *is equivalent to the Fredholm system* (4.5) *(equivalently,* (4.6)*).*

Proof. Given (4.5), we may write

$$P_i^{\epsilon_0} \boldsymbol{\psi} = \boldsymbol{\phi}_i^{\epsilon_0} + \boldsymbol{f}, \qquad P_e^{\epsilon_1} \Gamma \boldsymbol{\psi} = \boldsymbol{\phi}_e^{\epsilon_1} - \boldsymbol{f}.$$

Applying $P_e^{\epsilon_0}$ to the first relation and $P_i^{\epsilon_1}$ to the second relation, we obtain

$$P_e^{\epsilon_0} \boldsymbol{f} = 0, \qquad P_i^{\epsilon_1} \boldsymbol{f} = 0.$$

These equations imply $\boldsymbol{f} = 0$ by the hypothesis of the theorem. ∎

A comparison of the conditions of the theorem with (4.4) allows us to reformulate the theorem in a more physical way, as follows.

THEOREM 4.4 (reformulated Theorem (4.3)). *If the inverse dielectric structure (physics terminology meaning the original geometry with the two dielectric materials interchanged) does not support an electrically polarized field, then our scattering problem* (4.4) *is equivalent to the Fredholm system* (4.5) *(equivalently,* (4.6)*).*

5. Resonant frequencies and bound states. We have seen that the dielectric structure supports a nonzero source-free electromagnetic field ψ at (β,ω) if and only if the field satisfies

$$P_i^{\epsilon_0} \boldsymbol{\psi} = 0 \quad \text{and} \quad P_e^{\epsilon_1} \Gamma \boldsymbol{\psi} = 0, \tag{5.1}$$

where $\boldsymbol{\psi}(\mathbf{r}) = \left(\lim_{h\to 0}\psi(\mathbf{r}+hn(\mathbf{r})), \lim_{h\to 0}\partial_n\psi(\mathbf{r}+hn(\mathbf{r}))\right)^t$ on ∂D. This implies that

$$(P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma)\boldsymbol{\psi} = 0, \tag{5.2}$$

so that $(P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma)$ has a nontrivial nullspace at the resonant frequency ω for wave number β.

We now present a theorem that provides conditions for the existence of resonance frequencies.

THEOREM 5.1. *Let the frequency ω and a real value of β be given, and assume that $(m+\beta)^2 - \epsilon\omega^2 \neq 0$ for all integers m. If $P_i^{\epsilon_0}+P_e^{\epsilon_1}\Gamma$ has a nontrivial nullspace, then at least one of the following holds:*

1. *ω is a resonant frequency for the structure, with interior dielectric coefficient ϵ_1 and exterior coefficient ϵ_0 for the wave number β, according to whether $\nu = 1$ or $\nu = \epsilon_1/\epsilon_0$ in Γ, or*
2. *ω is a resonant frequency for the inverse structure, with ϵ_1 and ϵ_0 switched, for the wave number β.*

If ω^2 is real, then the corresponding source-free Helmholtz field ψ in $\mathcal{E}(\beta,\omega,\epsilon_0)$ or $\mathcal{E}(\beta,\omega,\epsilon_1)$ decays to zero as $|x|\to\infty$ and is therefore an x-localized field (a bound state in the strip $\mathcal{S}$). Otherwise, $\Im(\omega^2) < 0$ and ψ becomes unbounded as $|x|\to\infty$.

Remarks. 1. In our numerical studies, we will be interested only in the case in which $\Re(\omega) > 0$ and $\Im(\omega)$ is small and negative. Assuming $\Re(\omega) > 0$, the condition $\Im(\omega^2) < 0$ is equivalent to $\Im(\omega) < 0$.

2. Suppose that the pair (β_0,ω_0) with $\omega_0 > 0$ admits a bound state and that the Green's function has no propagating modes (see Figure 2.1). Then, for (β,ω) in a neighborhood of (β_0,ω_0), the Green's function has only decaying modes, so that all functions in $\mathcal{E}(\beta,\omega,\epsilon_0)$ are decaying. This implies that, in a vicinity of β_0 (we always assume β is real), a relation $\omega = W(\beta)$ describing resonant frequencies must be real-valued and must therefore be a dispersion relation for bound states.

Proof. We assume that (5.2) holds for some nonzero $\boldsymbol{\psi}$. Theorem 4.3, with both source fields in (4.4) taken to be zero, implies that equations (5.1) hold (giving statement (1) in Theorem 5.1) *or* that there exists a nonzero $\boldsymbol{f}$ such that $P_e^{\epsilon_0}\boldsymbol{f} = 0$ and $P_i^{\epsilon_1}\boldsymbol{f} = 0$ (giving statement (2) in Theorem 5.1).

We defer the proof of the condition on ω^2 and the behavior of the fields as $|x|\to\infty$ to the proof of Theorem 5.2. ∎

THEOREM 5.2. *Suppose that $\psi \in \mathcal{E}(\beta,\omega,\epsilon_0)$ is pseudoperiodic in y, is not identically zero, and satisfies $\nabla^2\psi + \epsilon\omega^2\psi = 0$ in $\mathcal{S}\setminus\partial D$ (where $\epsilon = \epsilon_1$ in D and ϵ_0 otherwise) and the matching conditions (4.2) on ∂D. Then $\Im(\omega^2) \le 0$. In addition, $|\psi|\to 0$ as $|x|\to\infty$ if and only if ω^2 is real-valued.*

Proof. Let T denote the finite strip $\{(x,y) : -x_0 \le x \le x_0, 0 \le y \le 2\pi\}$, where $x_0 > x_* > 0$ and x_* is given in Definition 2.2, and let ∂T be its boundary with outward-pointing normal vector n. We also take n pointing outward on ∂D. The divergence theorem gives

$$\int_{\partial T}\overline{\psi}\partial_n\psi ds + \int_{\partial D}\left(\overline{\psi}_{\rm int}\partial_n\psi_{\rm int} - \overline{\psi}_{\rm ext}\partial_n\psi_{\rm ext}\right)ds = \iint_T\left(\nabla\overline{\psi}\cdot\nabla\psi + \overline{\psi}\nabla^2\psi\right)dA.$$

Using the relation $\overline{\psi}\partial_n\psi = \overline{\tilde{\psi}}\partial_n\tilde{\psi} + i\beta n_y|\tilde{\psi}|^2$ (n_y is the y-component of the normal vector n), we see that the integrals over the top and bottom parts of ∂T cancel, and the integral over ∂T becomes an integral over $\Gamma_L\cup\Gamma_R$, where Γ_L and Γ_R are the left

and right sides of ∂T. On the right-hand side, we use the Helmholtz equation. We now have

(5.3)
$$\int_{\Gamma_L\cup\Gamma_R} \overline{\psi}\partial_n\psi ds + \int_{\partial D}\left(\overline{\psi}_{\rm int}\partial_n\psi_{\rm int} - \overline{\psi}_{\rm ext}\partial_n\psi_{\rm ext}\right) ds = \iint_T \left(|\nabla\psi|^2 - \epsilon\omega^2|\psi|^2\right) dA.$$

Using the conjugate equation for the interior of D, $\nabla^2\overline{\psi}_{\rm int}+\epsilon_1\overline{\omega}^2\overline{\psi}_{\rm int}=0$, we compute

$$\begin{aligned} -\iint_D |\psi_{\rm int}|^2(\epsilon_1\omega^2-\epsilon_1\overline{\omega}^2)dA &= -\iint_D \left(\overline{\psi}_{\rm int}\epsilon_1\omega^2\psi_{\rm int} - \psi_{\rm int}\epsilon_1\overline{\omega}^2\overline{\psi}_{\rm int}\right) dA \\ &= \iint_D \left(\overline{\psi}_{\rm int}\nabla^2\psi_{\rm int} - \psi_{\rm int}\nabla^2\overline{\psi}_{\rm int}\right) dA = \int_{\partial D}\left(\overline{\psi}_{\rm int}\partial_n\psi_{\rm int} - \psi_{\rm int}\partial_n\overline{\psi}_{\rm int}\right) ds. \end{aligned}$$

Therefore,

$$\Im\int_{\partial D}\overline{\psi}_{\rm int}\partial_n\psi_{\rm int}ds = -\epsilon_1\Im(\omega^2)\iint_D|\psi_{\rm int}|^2dA.$$

Using this and the matching conditions $\psi_{\rm int}=\psi_{\rm ext}$ and $\partial_n\psi_{\rm int}=\nu\partial_n\psi_{\rm ext}$ on ∂D, we can write the imaginary part of (5.3):

$$\Im\int_{\Gamma_L\cup\Gamma_R}\overline{\psi}\partial_n\psi ds \;-\; \epsilon_1(1-\nu^{-1})\Im(\omega^2)\iint_D|\psi|^2dA \;=\; -\Im(\omega^2)\iint_T\epsilon|\psi|^2dA. \tag{5.4}$$

By Definition 2.2, since $\psi\in\mathcal{E}(\beta,\omega,\epsilon_0)$, there exist complex numbers $A_m^{\pm}$ such that

$$\psi(x,y) = \sum_{m=-\infty}^{\infty} A_m^{\pm}\exp\left(\mu_m|x| + i(m+\beta)y\right) \quad \text{for} \quad \pm x > x_0.$$

Straightforward computation yields

$$\int_{\Gamma_L\cup\Gamma_R}\overline{\psi}\partial_n\psi ds \;=\; 2\pi\sum_{m=-\infty}^{\infty}\mu_m\left(|A_m^-|^2+|A_m^+|^2\right)e^{2\Re(\mu_m)|x_0|},$$

and, after splitting the right-hand side of (5.4) into an interior and an exterior integral, we obtain

$$\begin{aligned} &2\pi\sum_{m=-\infty}^{\infty}\Im(\mu_m)\left(|A_m^-|^2+|A_m^+|^2\right)e^{2\Re(\mu_m)|x_0|} \\ &\qquad = -\Im(\omega^2)\left(\epsilon_1\nu^{-1}\iint_D|\psi|^2dA + \epsilon_0\iint_{T\setminus D}|\psi|^2dA\right). \end{aligned} \tag{5.5}$$

If $\Im(\omega^2)>0$, all modes are decaying in x ($\Re(\mu_m)<0$ for all m), and we obtain a contradiction by letting x_0 tend to ∞; therefore, $\Im(\omega^2)\le 0$. If $\Im(\omega^2)=0$, then $\Im(\mu_m)>0$ for all propagating modes and $\Im(\mu_m)=0$ for all decaying modes; therefore, $A_m^{\pm}=0$ for all propagating modes, so that $|\psi|\to 0$ as $|x|\to\infty$. Conversely, if $|\psi|\to 0$ as $|x|\to\infty$, then $A_m^{\pm}=0$ for all nondecaying modes (those for which $\Re(\mu_m)\ge 0$), and thus the left-hand side of 5.5 decays exponentially as $x_0\to\infty$. Letting x_0 tend to ∞ shows that $\Im(\omega^2)=0$. ∎

Remarks. 1. The quantity $\Im\int_\Gamma\overline{\psi}\partial_n\psi\,ds$ appearing in the proof of Theorem 5.2 is the time-averaged energy flow carried by ψ through Γ.

2. If $\Im(\omega^2)<0$, then ψ does not decay as $|x|\to\infty$, and thus by the definition of $\mathcal{E}(\beta,\omega,\epsilon)$, ψ becomes unbounded as $|x|\to\infty$, and this completes the proof of Theorem 5.1.

6. Dispersion relations. We define a *dispersion relation* for a photonic crystal slab to be a multivalued function $\omega = W(\beta)$ describing pairs (β, ω) for which ω is a resonant frequency for wave number β. This means that the pair (5.1) is satisfied for a nonzero field $\boldsymbol{\psi}(\mathbf{r}) = \left(\lim_{h\to 0} \psi(\mathbf{r} + hn(\mathbf{r})), \lim_{h\to 0} \partial_n \psi(\mathbf{r} + hn(\mathbf{r}))\right)^t$ on ∂D, where ψ is a pseudoperiodic source-free Helmholtz field in the class $\mathcal{E}(\beta, \omega, \epsilon_0)$. If we are able to eliminate the second alternative in Theorem 5.1, then the single equation (5.2) is sufficient for defining a dispersion relation, as it would then imply the pair (5.1). In the numerical examples below, we are able computationally to eliminate the second alternative.

Because our operator is of the Hilbert–Schmidt class from $L^2(\partial D) \times L^2(\partial D)$, its determinant is defined, and it depends analytically on β and ω. Thus, a *necessary* condition for the pair (β, ω) to support a source-free field in the given crystal is

$$D(\beta, \omega) := \det(P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma) = 0. \tag{6.1}$$

The dispersion relation is therefore given by branches of $D(\beta, \omega) = 0$. Whether this condition is also *sufficient* depends on the inverse dielectric structure. Indeed, we know from Theorem 5.1 that if the inverse structure does not support an electrically polarized source-free field, then (6.1) does define a dispersion relation for fields in the original structure.

In practice, we obtain dispersion relations by computing the curves $\lambda(\beta, \omega) = 0$ numerically, where λ is an eigenvalue of $P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma$. Suppose that zero is a simple eigenvalue at the pair (β_0, ω_0). Then the smallest eigenvalue λ is an analytic function of β and ω in a neighborhood of (β_0, ω_0), say $\lambda = \lambda(\beta, \omega)$. Suppose also that $\lambda(\beta_0, \omega) \not\equiv 0$ near ω_0. By the Weierstraß preparation theorem, there exists an integer $n \geq 1$ such that

$$\lambda(\beta, \omega) = h(\beta, \omega)(\omega^n + W_{n-1}(\beta)\omega^{n-1} + \cdots + W_1(\beta)\omega + W_0(\beta)) \tag{6.2}$$

near (β_0, ω_0), where h and W_i $(i = 0, \ldots, n-1)$ are analytic functions and $h \neq 0$ near (β_0, ω_0). Thus, $\lambda = 0$ is equivalent to $\omega^n + W_{n-1}(\beta)\omega^{n-1} + \cdots + W_1(\beta)\omega + W_0(\beta) = 0$. Let us consider the case in which $n = 1$ (this is when $\partial\lambda/\partial\omega \neq 0$ near (β_0, ω_0)). Then we have a relation

$$\omega = W(\beta)$$

that describes the locus of (β, ω)-pairs for which $\lambda = 0$. For real values of β near β_0, the curve $\omega = W(\beta)$ in the complex ω-plane gives a dispersion relation; it is periodic in β with period 1.

We made two assumptions in the preceding paragraph: that zero is a *simple* eigenvalue at (β_0, ω_0), and that $\partial\lambda/\partial\omega \neq 0$ there. Numerical calculations show that both assumptions are true generically, giving rise to dispersion relations for simple eigenvalues. When two branches cross, as in Figure 7.3(1a) and (2a) below, we see an eigenvalue of multiplicity 2. We have not encountered a situation in which there is a simple eigenvalue at some point (β_0, ω_0) and $\partial\lambda/\partial\omega = 0$ there, that is, $\partial^k\lambda/\partial\omega^k = 0$ for $k < n$ and $\partial^n\lambda/\partial\omega^n \neq 0$ for some $n > 1$ in (6.2). In this situation, (6.2) shows that there would exist a dispersion relation defined by an equation that is algebraic in ω and may have several branches emanating from (β_0, ω_0).

If $\omega = W(\beta)$ lies on the real ω-axis over a range of β-values, then we have a dispersion relation for x-localized Helmholtz fields ψ in the crystal (bound states in the strip $\mathcal{S}$). The complex time-dependent electric or magnetic fields associated with

ψ are $\psi(x,y)e^{-i\omega t} = \tilde{\psi}(x,y)e^{i(\beta y-\omega t)}$ (with $\tilde{\psi}$ periodic in y), which are Bloch waves traveling along the photonic crystal slab. If $\omega_0 = W(\beta_0)$ is real but the dispersion relation goes into the lower-half ω-plane for $\beta \neq \beta_0$, then there exists a bound state just for the isolated pair (β_0, ω_0), and nearby pairs $(\beta, W(\beta))$ are complex resonant frequencies. We demonstrate numerically below that both situations do occur.

One of the main phenomena observed in this study is that *complex resonant frequencies are linked to resonant scattering behavior and transmission anomalies at nearby* real *frequencies.* See also [5].

7. Numerical studies: Bound states, surface waves, and resonances. In this section we present three examples that illustrate the connections between complex dispersion relations, bound states and resonant frequencies, transmission anomalies, and resonant behavior of real frequencies. In particular, we calculate a dispersion relation for surface waves at bandgap frequencies on a thick structure approximating a semi-infinite crystal, and we provide a mathematical context for understanding resonant phenomena produced by a channel defect in a crystal slab.

We calculate the dispersion relations numerically as follows. We first search for a pair (β_0, ω_0) at which $P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma$ has an eigenvalue that is practically zero: $\lambda(\beta_0, \omega_0) \approx 0$. Then we increment β to β_1 and search in a complex vicinity of ω_0 for a value ω_1 such that $\lambda(\beta_1, \omega_1) \approx 0$. We continue to increment β, and in this way trace out a curve $\omega(\beta)$ represented by the computed points (β_n, ω_n) such that $\lambda(\beta_n, \omega_n) \approx 0$. To find the value ω_n, we simply compute the minimum of the smallest eigenvalue $\lambda(\beta, \omega)$ of $P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma$ as ω varies over a grid about ω_{n-1}, keeping β fixed at β_n, and then refine the search if necessary. We intend to develop a more efficient gradient search method for future investigations. In the examples in subsections 7.1 and 7.2, we find the initial pair (β_0, ω_0) by taking $\beta_0 = 0$ and computing the minimum eigenvalue of $P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma$, using MATLAB, on a grid of real values of ω in the interval $(0, 1)$. An initial bound state was easy to find. In the example of subsection 7.3, we knew to search in the vicinity of a spike in the transmission graph that we had computed in [2].

To show that $\lambda(\beta, \omega)$ actually achieves a value of zero, it suffices to fix β and compute (numerically) a positive winding number of λ as an analytic function of ω about some small closed curve in the complex ω-plane. We perform this verification at selected points on the dispersion relations. We also check numerically that the alternative (part (2)) in Theorem 5.1 does not hold.

To compute the eigenvalues of the boundary-integral operator $P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma$, we discretize the integral system using quadratic basis elements for the fields and point-sampling of the equations. Complete details of these calculations are presented in [2]. Once we have obtained a numerical solution to $(P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma)\boldsymbol{\psi} = \boldsymbol{\phi}$, we compute the scattering and bound ($\phi = 0$) states using Green's identities (3.4) and (3.5).

In all three examples, we study electrically polarized fields with $\epsilon_0 = 1$ and $\epsilon_1 = 12$.

In the figures, the fields in the crystals are represented by contour plots of their magnitudes. White represents the maximal amplitude, and black represents an amplitude of zero. One y-period is shown, with the x-direction truncated outside the support of the crystal slab.

7.1. A single string of rods. (Figure 7.1.) Our first example is a good illustration of the connection between dispersion relations, bound states, and resonant scattering phenomena. The period of our dielectric structure consists of a single rod in air, so that the crystal slab degenerates into a string of rods running in the y-direction.

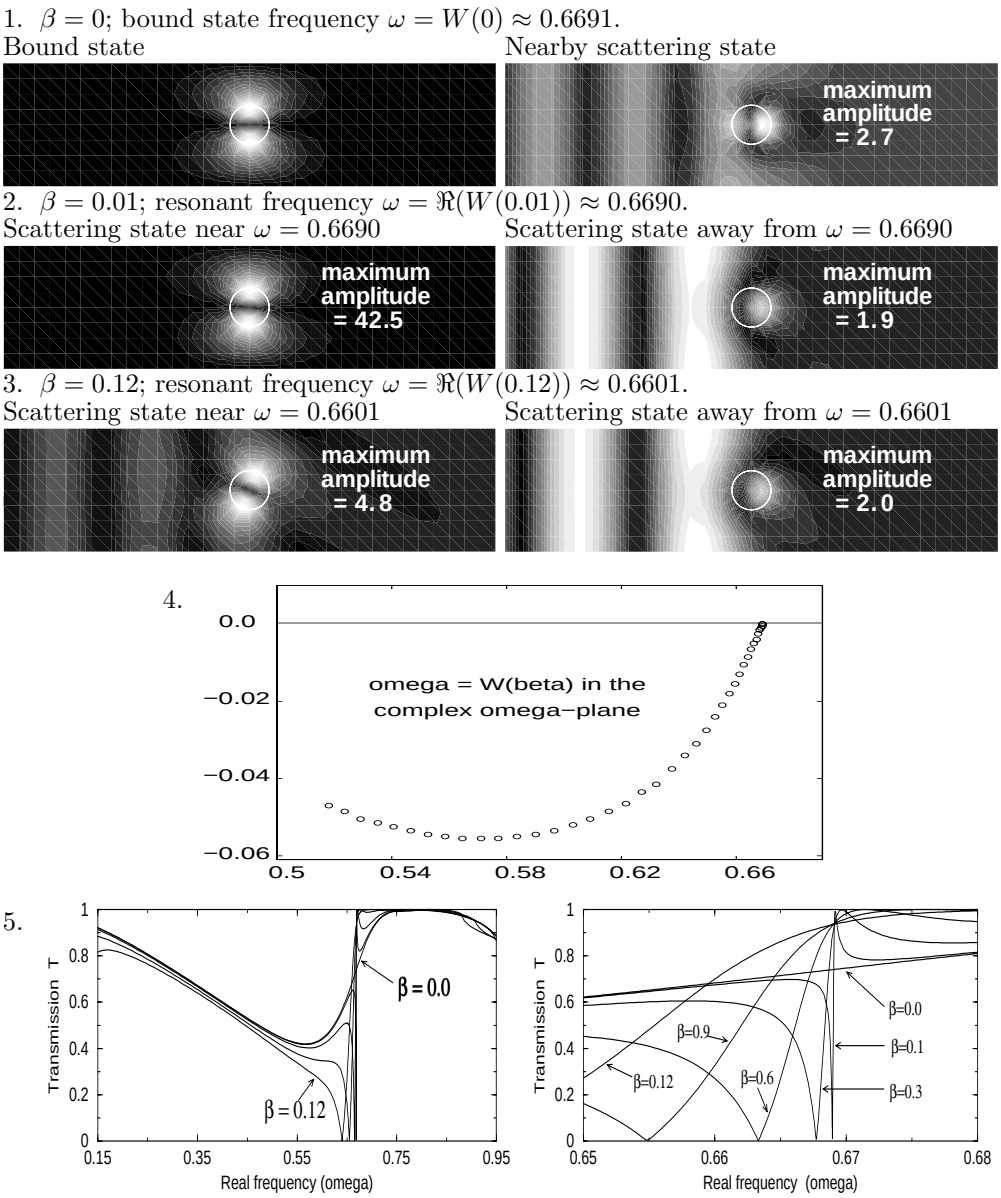


FIG. 7.1. *In the contour plots, one period in y of the crystal is shown, with the magnitude of the electrically polarized fields plotted.* 1. *There is a bound state at* $(\beta = 0, \omega = W(0) \approx 0.6691)$. *There is no unusually high amplification of an incident plane-wave source (of amplitude* 1*) for* $\beta = 0$ *at frequencies* ω *near the bound-state frequency.* 2,3. *When* $\Im(W(\beta)) < 0$, *incident plane-wave sources are amplified in the rod at frequencies near the real part of* $W(\beta)$. *A smaller imaginary part corresponds to greater amplification. The field structure at resonant frequencies is similar to that of the bound state.* 4. *The dispersion relation* $W(\beta)$ *plotted in the complex* ω *plane for real values of beta from* $\beta = 0$ *to* $\beta = 0.38$. 5. *Spikes in the transmission coefficient near* $\omega = \Re(W(\beta))$ *for* $\beta > 0$, *where* $\Im(W(\beta)) < 0$.

We find numerically a y-periodic (this means $\beta = 0$) bound state at $\omega \approx 0.6691$. This is a standing wave that exists in the absence of any sources. However, as β moves away from zero, the bound state disappears and the dispersion relation $\omega = W(\beta)$ enters the lower-half complex ω-plane (Figure 7.1(4)). In place of a bound state, we find instead resonant scattering states at *real* frequencies near the real part $\Re(W(\beta))$ of the complex resonant frequency on the dispersion relation. These states are sustained by plane-wave source fields that are amplified as they experience resonant scattering within the dielectric structure. Coinciding with these large fields are anomalies (spikes) in the transmission coefficient[2] $T(\beta, \omega)$ for real values of ω near $\Re(W(\beta))$, for a fixed value of β (see Figure 7.1(5)). These spikes consist of a drop to 0% transmission ($T=0$) to the left of the frequency of the periodic bound state that exists at $\beta = 0$, followed by a sharp increase to 100% transmission ($T = 1$) to the right. As $|\beta|$ decreases to zero, the width of the spike decreases and the resonant amplification of the scattering fields increases. The phenomena become more and more localized about the bound-state frequency. At $\beta = 0$, the transmission anomaly and resonant behavior disappear, and we have in their place the bound state.

We find that, for $\omega_0 \approx 0.6691$, $T(\beta, \omega_0) \approx 0.935$ for values of β near but not equal to zero (see Figure 7.1(5)). However, continuing our calculation of the curve $T(0, \omega)$ through $\omega = \omega_0$ gives $T(0, \omega_0) \approx 0.739$. Thus we demonstrate numerically that $\lim_{\beta \to 0} T(\beta, \omega_0) \neq \lim_{\omega \to \omega_0} T(0, \omega)$, so that we cannot define the transmission coefficient continuously at $(0, \omega_0)$. This observation strengthens our belief that there is a bound state at $(0, \omega_0)$, that is, that $\Im(W(0))$ is indeed *exactly* zero.

From Figure 7.1(4), the dispersion relation evidently has the form

$$\omega = \omega_0 + a(\beta)\beta^2,$$

where a is analytic and $a(0) \neq 0$. ($W(\beta)$ is symmetric about $\beta = 0$, as the Green's functions at $\pm\beta$ have the same set of modes, and the structure is symmetric in y.)

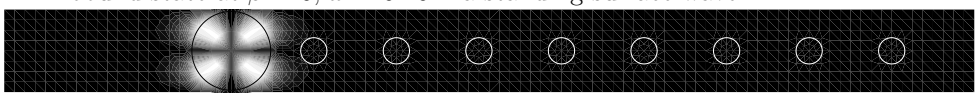
7.2. Surface waves. (Figure 7.2.) We investigate waves on the surface of a semi-infinite photonic crystal. We seek waves at the interface between the left half-plane containing air and the right half-plane filled with a square lattice of circular rods. If we consider frequencies that cannot propagate through the infinite crystal, or bandgap frequencies, it is reasonable to approximate the semi-infinite crystal by a finite slab several rods thick (truncated in the x-direction but still periodic in y). To capture surface waves at bandgap frequencies, we place a defect on the left surface of the slab by making the first rod much larger than the others.

There appears to be a periodic ($\beta = 0$) bound state at $\omega \approx 0.401$, as the smallest eigenvalue of $P_i^{\epsilon_0} + P_e^{\epsilon_1}\Gamma$ is practically zero there. The field is localized at the defective surface of the slab (see Figure 7.2(1))—it is a standing surface wave.

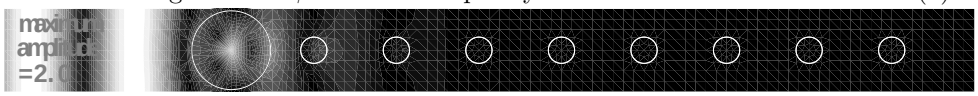
As β moves away from zero, the surface wave disappears, and we obtain a dispersion relation $\omega = W(\beta)$ with a very small negative imaginary part. At about $\beta \approx 0.345$, however, the relation enters a regime in which $|W(\beta)| < \min_{m \in \mathbb{Z}}(|\beta + m|)$, which is the scenario in which *all* modes of the Green's function decay in $|x|$ and the exterior medium admits no traveling waves. We deduce from Theorem 5.1 (or Theorem 5.2) that this part of the dispersion relation is necessarily real and is therefore a

[2]Our transmission coefficient is the square root of the ratio of the energy transmitted through the slab on the right to the energy of a plane-wave source field $\exp(i\sqrt{\epsilon_0\omega^2 - (m+\beta)^2}\,x + i(m+\beta)y)$ incident upon the slab from the left. Details of how we calculate the transmission coefficient are given in [2].

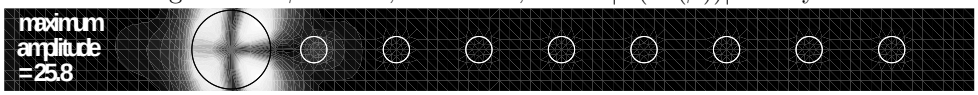
1. A bound state at $\beta = 0$, $\omega \approx 0.401$: a standing surface wave.



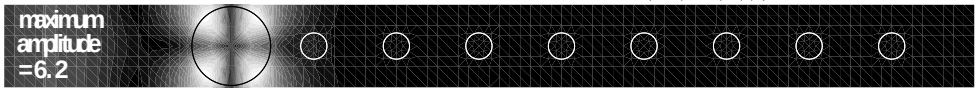
2. A scattering state for $\beta = 0$ at a frequency near that of the bound state in (1).



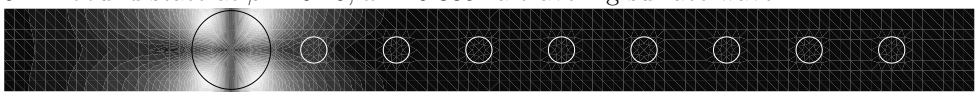
3. A scattering state at $\beta = 0.23$, $\omega \approx 0.368$, where $|\Im(W(\beta))|$ is very small.



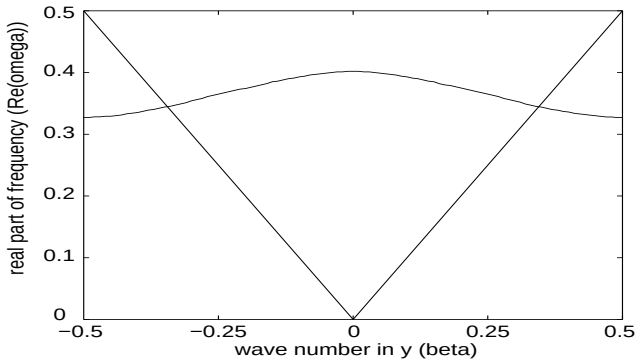
4. A scattering state at $\beta = 0.28$, $\omega \approx 0.358$, near $\max |\Im(W(\beta))|$.



5. A bound state at $\beta = 0.40$, $\omega \approx 0.335$: a traveling surface wave.



6.



7.

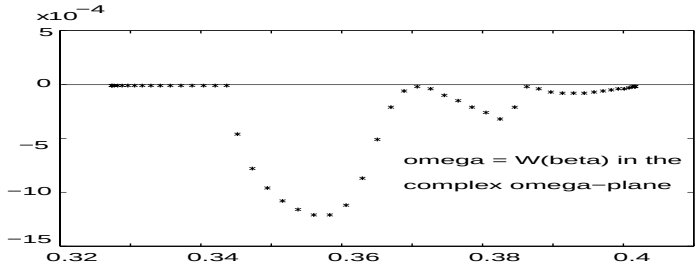


FIG. 7.2. *A crystal slab with a defect on the left surface, where one rod is bigger than the rest. In the contour plots, one period in y of the crystal is shown, with the magnitude of the electrically polarized fields plotted.* 1,2. *There is a periodic ($\beta = 0$) bound state, which is a standing surface wave localized in the large rod at the surface of the crystal slab. At frequencies near that of the surface wave, incident plane-wave sources are not amplified in the rod.* 3,4. *When $\Im(W(\beta)) < 0$, incident plane-wave sources (with amplitude 1) at frequencies near $\Re(W(\beta))$ resonate in the first rod.* 5. *A bound state at parameter values on the dispersion relation, where $\omega = W(\beta)$ is necessarily real and the strip admits no scattering (extended) states.* 6. *$\Re(W(\beta))$ plotted against β. $\Im(W(\beta)) = 0$ when $\Re(W(\beta)) < |\beta|$ ($0.345 < |\beta| < 0.5$).* 7. *$\omega = W(\beta)$ plotted in the complex ω-plane.*

dispersion relation for x-localized fields. In fact, in this regime, there are no scattering (extended) states in the strip (see Remark 2 following Theorem 5.1).

We therefore demonstrate the existence of a dispersion relation for waves traveling along the surface of a photonic crystal slab. Our choice of eight rods to approximate the semi-infinite crystal appears to be sufficient: Increasing the number of rods results in no appreciable difference in the calculated curve. Thus, we are led to believe that the dispersion relation for the semi-infinite crystal exists and that the dispersion relation for the slab is a good approximation.

As in the case of the single string of rods in the previous example, we see again how the complex part of the dispersion relation, which is very close to the real axis, affects the scattering states at nearby real frequencies. We do not find anomalies in the transmission as we did in that example; however, we believe that they are there but are too sharp to be detected numerically because of the small size of the imaginary part of the dispersion relation and the thickness of the slab.

7.3. A channel defect. (Figure 7.3.) In [2], we studied the effect that a periodic channel through a photonic crystal slab has on the transmission coefficient and the structure of the scattering states. Our intention was to study resonant behavior at bandgap frequencies and modes propagating through the channel; however, we also found intriguing sharp transmission spikes and resonant scattering fields in near–full-transmission frequency regions. The phenomena are similar in behavior to those in our previous two examples, and a similar analysis provides us with a better mathematical understanding of them.

The (β,ω) pairs at which resonant behavior and transmission anomalies occur for the crystal with a periodic channel are described by a dispersion relation $\omega = W(\beta)$, which we calculate numerically.[3] We demonstrate in [2] that both phenomena disappear as the width of the channel decreases and the slab returns to its perfect structure. In the present study, we find that, in place of resonant scattering states, the perfect slab admits bound states, also described by a dispersion relation.

Our numerical calculations give us dispersion relations that lie practically on the real ω-axis. However, they may have a small negative imaginary part that we cannot resolve numerically. Based on our findings in the previous two examples, we conjecture that the relation is identically real for the perfect slab (there is no resonant amplification, and there are no transmission spikes) and that it has a very small negative imaginary part for the slab with a periodic channel (there are high-amplitude resonances and very sharp transmission spikes).

The structure of the resonant fields near $\omega = \Re(W(\beta))$ when $\Im(W(\beta))$ is small resembles the field of a nearby bound state, if such a bound state exists. See Figure 7.3(1) and (3), which shows the y-directional structure of a bound state and a resonant scattering state. A typical scattering state at small values of β exhibits the structure of an x-directional interference pattern, a point we discuss in [2].

In summary, we have a plausible explanation for the channel-induced resonant behavior that we observed in [2]: There is a dispersion relation $\omega = W(\beta)$ for bound states traveling along the perfect crystal slab. When a periodic channel is introduced, the new (perturbed) relation $\omega = W(\beta)$ gives resonant frequencies with a small imaginary part. These are responsible for the observed buildup of large fields in the slab and for the transmission anomaly near $\omega = \Re(W(\beta))$. In the limit of zero imaginary

[3] In [2], we scaled the frequency in this example by 1/6 to compare with the scale of the finer period of one rod. We did not make that scaling here.

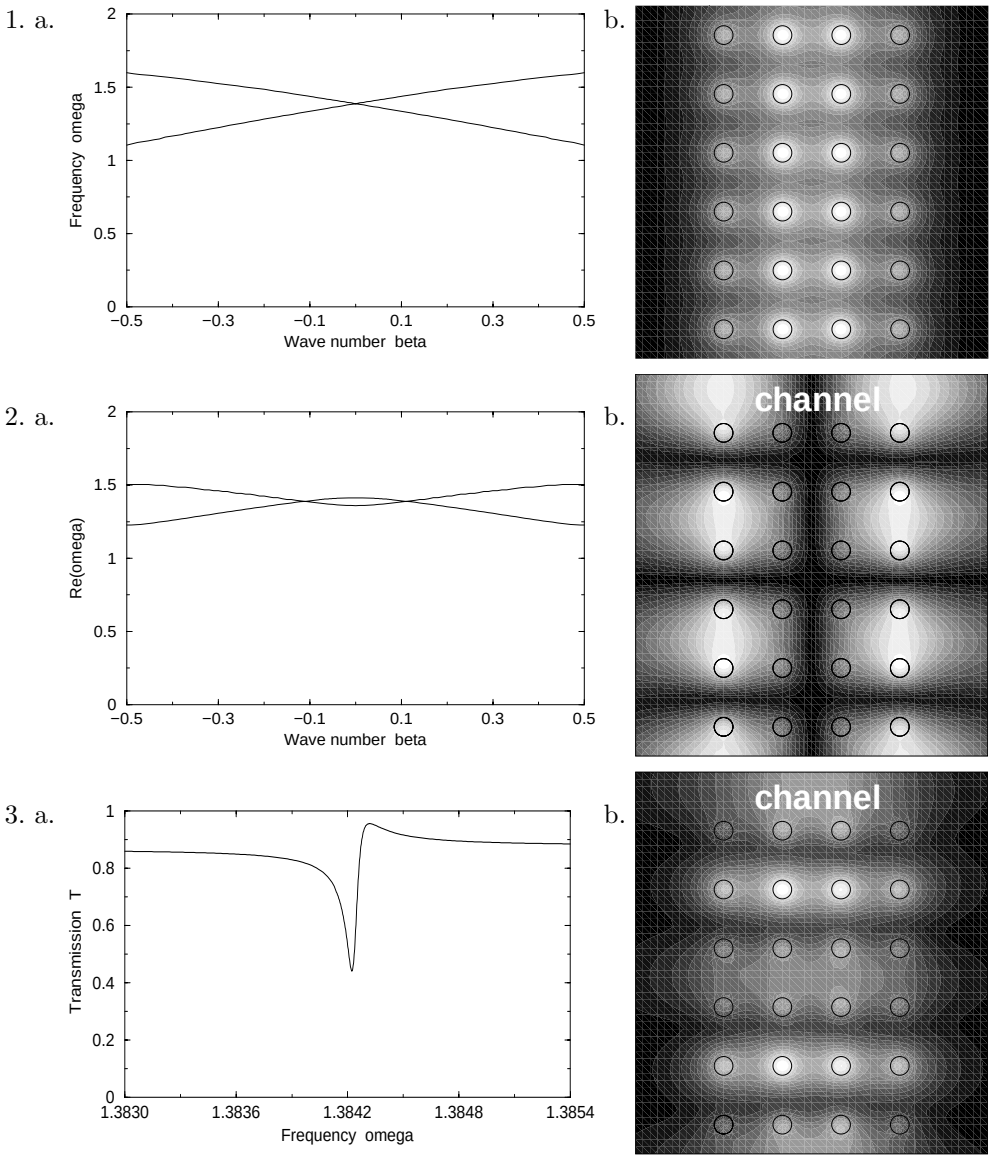


FIG. 7.3. *In the contour plots, a period in y of the crystal is shown, and the strip is truncated in the x-direction. The magnitude of the electrically polarized fields is plotted.* 1.a. *Branches of the dispersion relation for a perfect square-lattice crystal four rods thick.* 1.b. *The bound state at* $\beta = 0.1$ *and* $\omega \approx 1.434$. 2.a. *Branches of the real part of the dispersion relation for the crystal with a periodic channel containing an extra half-period of space after every six rows of rods.* 2.b. *A resonant scattering state at* $\beta = 0.5$ *and* $\omega \approx 1.224$. 3.a. *An anomaly in the transmission coefficient for* $\beta = 0.1$. 3.b. *A resonant scattering state at* $\beta = 0.1$ *and* $\omega \approx 1.3842$ *(slightly to the left of where the two branches of the dispersion relation cross, on the lower branch). The amplitude of the field in the crystal reaches about* 49 *times the amplitude of the incident field.*

part, the anomaly occurs at a single point (the point of the bound state) and is not observed through the numerical scattering experiment.

The numerical evidence indicates that the slab with a periodic channel supports resonant frequencies that converge to bound state frequencies as the channel closes up. A similar phenomenon is known for Helmholtz resonators, in which resonant frequencies of a cavity with an opening converge to the bound state frequencies of the closed cavity as the opening disappears (Beale [6]).

8. Appendix. Let D be a domain whose boundary $C = \partial D$ is a closed curve of class C^2 with arclength parameterization $\gamma(t)$, $t \in [a, b]$, and outward-pointing normal vector $n(t)$ at $\gamma(t)$. For sufficiently small values of $\rho > 0$, the function $\gamma_\rho(t) = \gamma(t) - \rho n(t)$ parameterizes a C^1 closed curve C_ρ contained in the interior of D.

The Green's kernel $G(\mathbf{r})$ (or its derivatives) takes a function ϕ on C to a function ψ_ρ on C_ρ by

$$\psi_\rho(\mathbf{r}) = \int_C G(\hat{\mathbf{r}} - \mathbf{r})\phi(\mathbf{r}) ds(\mathbf{r}),$$

where $\hat{\mathbf{r}} \in C_\rho$. This map can be realized as a map taking a function on $[a, b]$ to another by recycling notation and writing

$$\psi_\rho(t) = \int_a^b G(\gamma_\rho(t) - \gamma(s))\phi(s)\, ds.$$

We make the identifications $H^1(C) \equiv H^1([a, b])$ and $L^2(C) \equiv L^2([a, b])$.

1. The singular part of the integral kernel $G(\gamma(t) - \gamma(s))$ is $\log|t - s|$, and the singular part of its derivative, $d/dt \log|t-s|$, which has the Hilbert-transform singularity, is a principal-value integral. Both are bounded operators from L^2 to L^2, so $G(\gamma(t) - \gamma(s))$ is a bounded operator from L^2 to H^1. $G(\gamma_\rho(t) - \gamma(s))$ and $d/dt\, G(\gamma_\rho(t) - \gamma(s))$ are regularizations of these singular kernels, so $G(\gamma_\rho(t) - \gamma(s))$ converges to $G(\gamma(t) - \gamma(s))$ as $\rho \to 0$ as bounded operators from L^2 to H^1.
2. By the theory of the double-layer potential and the Plemelj formula, the integral kernels

$$\frac{\partial G(\gamma_\rho(t) - \gamma(s))}{\partial n(\gamma(s))}, \qquad -\frac{\partial G(\gamma_\rho(t) - \gamma(s))}{\partial n(\gamma(t))},$$

as applied to L^2, are a regularization of their limiting form as $\rho \to 0$, which is $1/2$ the identity operator plus a weakly singular integral kernel.
3. Using the identity

$$\frac{\partial^2 G(\hat{\mathbf{r}} - \mathbf{r})}{\partial n(\hat{\mathbf{r}})\partial n(\mathbf{r})} + \frac{\partial^2 G(\hat{\mathbf{r}} - \mathbf{r})}{\partial t(\hat{\mathbf{r}})\partial t(\mathbf{r})} = -n(\hat{\mathbf{r}}) \cdot n(\mathbf{r})(\partial_x^2 + \partial_y^2) G(\hat{\mathbf{r}} - \mathbf{r})$$

and the Helmholtz equation

$$(\partial_x^2 + \partial_y^2) G(\hat{\mathbf{r}} - \mathbf{r}) + \epsilon\omega^2 G(\hat{\mathbf{r}} - \mathbf{r}) = 0 \quad \text{for} \quad \hat{\mathbf{r}} \neq \mathbf{r},$$

we can rewrite the kernel $\partial^2 G(\hat{\mathbf{r}} - \mathbf{r})/\partial n(\hat{\mathbf{r}})\partial n(\mathbf{r})$ for $\hat{\mathbf{r}} \notin \partial D$, applied to a

function $\xi \in H^1$:

$$\int_{\partial D} \frac{\partial^2 G(\hat{\mathbf{r}}-\mathbf{r})}{\partial n(\hat{\mathbf{r}})\partial n(\mathbf{r})}\xi(\mathbf{r})\,ds(\mathbf{r}) = \int_{\partial D}\left(n(\hat{\mathbf{r}})\cdot n(\mathbf{r})\epsilon\omega^2 G(\hat{\mathbf{r}}-\mathbf{r})\xi(\mathbf{r}) + \frac{\partial G(\hat{\mathbf{r}}-\mathbf{r})}{\partial t(\hat{\mathbf{r}})}\frac{d\xi}{dt(\mathbf{r})}(\mathbf{r})\right) ds(\mathbf{r}). \tag{8.1}$$

The kernel $\partial G/\partial t$ converges to a principal-value kernel as $\hat{\mathbf{r}} \to \partial D$, and, since $d\xi/dt(\mathbf{r}) \in L^2$, we see that the operator with kernel $\partial^2 G(\hat{\mathbf{r}}-\mathbf{r})/\partial n(\hat{\mathbf{r}})\partial n(\mathbf{r})$ converges to a bounded operator from H^1 to L^2 as $\hat{\mathbf{r}} \to \partial D$.

REFERENCES

[1] S. Venakides, M. A. Haider, and V. Papanicolaou, *Boundary integral calculations of two-dimensional electromagnetic scattering by photonic crystal Fabry–Perot structures*, SIAM J. Appl. Math., 60 (2000), pp. 1686–1706.

[2] M. A. Haider, S. P. Shipman, and S. Venakides, *Boundary-integral calculations of two-dimensional electromagnetic scattering in infinite photonic crystal slabs: Channel defects and resonances*, SIAM J. Appl. Math., 62 (2002), pp. 2129–2148.

[3] S. Fan, P. R. Villeneuve, and J. D. Joannopoulos, *Channel-drop filters in photonic crystals*, Optics Express, 3 (1998), pp. 4–11.

[4] A. Krishnan, T. Thio, T. J. Kim, H. J. Lezec, T. W. Ebbessen, P. A. Wolff, J. Pendry, L. Martin-Moreno, and F. J. Garcia-Vidal, *Evanescently coupled resonance in surface plasmon enhanced transmission*, Optics Commun., 200 (2001), pp. 1–7.

[5] S. G. Tikhodeev, A. L. Yablonskii, E. A. Muljarov, N. A. Gippius, and T. Ishihara, *Quasiguided modes and optical properties of photonic crystal slabs*, Phys. Rev. B, 66 (2002), 045102.

[6] J. T. Beale, *Scattering frequencies of resonators*, Comm. Pure Appl. Math., 26 (1973), pp. 549–563.

[7] A. P. Calderón, *Boundary-value problems for elliptic equations*, in Outlines Joint Sympos. Partial Differential Equations (Novosibirsk, Russia), Acad. Sci. USSR Siberian Branch, Moscow, 1963, pp. 303–304.

[8] R. T. Seeley, *Singular integrals and boundary value problems*, Amer. J. Math., 88 (1966), pp. 781–809.

[9] V. S. Ryaben'kii, *Boundary equations with projections*, Russian Math. Surveys, 40 (1985), pp. 147–183.

[10] J.-C. Nédélec, *Acoustic and Electromagnetic Equations*, Springer-Verlag, New York, 2001.

[11] C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer-Verlag, New York, 1969.

[12] D. L. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, New York, 1998.

[13] D. L. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, Wiley-Interscience, New York, 1983.

SIAM J. APPL. MATH.
Vol. 64, No. 1, pp. 343–368

HOPF BIFURCATION IN QUASI-GEOSTROPHIC CHANNEL FLOW*

ZHI-MIN CHEN†, MICHAEL GHIL‡, ERIC SIMONNET§, AND SHOUHONG WANG¶

This paper is dedicated to the memory of Jacques-Louis Lions

Abstract. In this article, we conduct a rigorous stability and bifurcation analysis for a highly idealized model of planetary-scale atmospheric and oceanic flows. The model is governed by the two-dimensional, quasi-geostrophic equation for the conservation of vorticity in an east-west oriented, periodic channel. The main result is the existence of Hopf bifurcation of the flow as the Reynolds number crosses a critical value.

The key idea in proving this result is translating the eigenvalue problem into a difference equation and treating the latter by continued-fraction methods. Numerical results are obtained by using a finite-difference scheme with high spatial resolution and these results agree closely with the theoretical predictions. The spatio-temporal structure of the limit cycle corresponds to a wave that propagates slowly westward and is symmetric about the midaxis of the channel. For plausible paramater values that correspond to midlatitude atmospheric flows, the period of this wave is 20–25 days.

Key words. quasi-geostrophic channel flow, Hopf bifurcation, atmospheric and oceanic waves

AMS subject classifications. 35Q30, 86A05, 86A10, 76E20, 58C40

DOI. 10.1137/S0036139902406164

1. Introduction.

A key problem in the study of climate dynamics is to understand and predict the periodic, quasi-periodic, aperiodic, and fully turbulent characteristics of large-scale atmospheric and oceanic flows. Bifurcation theory enables one to determine how qualitatively different flow regimes appear and disappear as control parameters vary; it provides us, therefore, with an important method to explore the theoretical limits of predicting these flow regimes. In the present paper, we study bifurcations of the original partial differential equations (PDEs) that govern geophysical flows, whereas most studies so far have only considered systems of ordinary differential equations (ODEs) that are obtained by projecting the PDEs onto a finite-dimensional solution space, either by finite differencing or by truncating a Fourier expansion (see Ghil and Childress [10] and further references there). The present approach should allow us to overcome some of the inherent limitations of the numerical bifurcation results that dominate the climate dynamics literature up to this point, and to capture the essential dynamics of the governing PDE systems.

*Received by the editors April 23, 2002; accepted for publication (in revised form) June 20, 2003; published electronically December 17, 2003.

http://www.siam.org/journals/siap/64-1/40616.html

†School of Mathematics, Nankai University, Tianjin 300071, People's Republic of China (zhimin@nankai.edu.cn). The research of this author was supported in part by the National Natural Science Foundation of China.

‡Department of Atmospheric Sciences and Institute of Geophysics and Planetary Physics, University of California Los Angeles, Los Angeles, CA 90095-1567 and Département Terre-Atmosphère-Océan, Ecole Normale Supérieure, Paris, France (ghil@atmos.ucla.edu, http://www.atmos.ucla.edu/tcd). The research of this author was supported in part by the National Science Foundation under grant ATM-0082131.

§Institut Nonlinéaire de Nice, UMR 6618 CNRS 1361, route des Lucioles 06560 Valbonne, France (eric.simonnet@inln.cnrs.fr). The research of this author was supported in part by the Department of Energy under grant DE-FG02-01ER63251 to Michael Ghil.

¶Department of Mathematics, Indiana University, Bloomington, IN 47405 (showang@indiana.edu, http://www.indiana.edu/~fluid). The research of this author was supported in part by the Office of Naval Research under grant N00014-96-1-0425 and by the National Science Foundation under grant DMS-0306447.

The basic equations of large-scale atmospheric and oceanic circulation are the primitive equations. These equations can be derived from the full Navier–Stokes equations *with gravity, rotation, and variable density* by neglecting vertical accelerations (the so-called hydrostatic approximation) and compressibility effects (i.e., sound waves); see Ghil and Childress [10], Kalnay [21], Lions, Temam, and Wang [24, 25], and Pedlosky [36]. One philosophy in the geosciences is to study in great detail simplified models that approximate well the dominant balance of forces on the planetary-scale atmospheric and oceanic flows before addressing the more complete PDE systems that govern these flows in all their complexity. By starting with models that incorporate only the most important effects, and by gradually bringing in others, one is able to proceed inductively and thereby avoid the pitfalls inevitably encountered when a great many poorly understood factors are introduced all at once.

The ideas of dynamical systems theory and nonlinear functional analysis have been applied so far to climate dynamics mainly by careful numerical studies. These were pioneered by Lorenz [26, 27], Stommel [43], and Veronis [44, 45] among others, who explored the bifurcation structure of low-order models of atmospheric and oceanic flows. More recently, pseudoarclength continuation methods have been applied to atmospheric (Legras and Ghil [23]) and oceanic (Speich, Dijkstra, and Ghil [42] and Dijkstra [9]) models with increasing horizontal resolution. These numerical bifurcation studies have so far produced fairly reliable results for two classes of geophysical flows: (i) atmospheric flows in a periodic midlatitude channel, in the presence of bottom topography and a forcing jet; and (ii) oceanic flows in a rectangular midlatitude basin, subject to wind stress on its upper surface. In both cases, the symmetry properties of the forcing have a decisive effect on the bifurcations that arise—saddle-node (Charney and DeVore [6] and Pedlosky [35]) or Hopf (Legras and Ghil [23] and Jin and Ghil [19]) in the atmospheric channel and saddle-node, pitchfork, or Hopf [2, 4, 14, 16, 18, 32, 37, 38, 42] in the oceanic basin.

More recently, the role of global bifurcations, via homoclinic and heteroclinic orbits, has been demonstrated numerically in the wind-driven ocean circulation problem, for both shallow-water (Chang et al. [5], Simonnet et al. [40, 41]) and quasi-geostrophic (QG) (Meacham [31] and Nadiga and Luce [34]) models. Both of these models represent further simplifications of the primitive equations [10, 36]. Still, the only mathematically rigorous proof of a bifurcation in either the atmospheric or the oceanic problem outlined here appears, as far as we know, in the work of Wolansky [47, 48], and it extends solely to the existence of asymmetric stationary solutions.

The present paper addresses the somewhat more difficult problem of proving the existence of Hopf bifurcation in a QG flow in two dimensions. The main difficulty in solving this problem is in estimating the crucial information on the spectrum of the problem linearized around the basic flow. The main objective of this article is to overcome this difficulty and bring a new set of tools to the rigorous study of successive bifurcations in geophysical fluid dynamics problems.

More precisely, we conduct a bifurcation analysis of the following idealized two-dimensional (2-D) QG flow problem. The governing equation dictates the conservation of vorticity, as modified by forcing and dissipation:

$$\partial_t \Delta\psi + \varepsilon J(\psi, \Delta\psi) + \partial_x \psi = E\Delta^2\psi - \tau_0 \sin \pi y, \tag{1.1}$$

where $\psi = \psi(x, y, t)$ is a streamfunction and $J(\psi, \phi) = \partial_x\psi\partial_y\phi - \partial_y\psi\partial_x\phi$ is the advection operator. The x-axis is directed to the east and the y-axis to the north. The zonal

and meridional velocity components u and v are obtained from the streamfunction by

$$u = -\psi_y, \qquad v = \psi_x.$$

The relative vorticity ξ and the streamfunction ψ are related by the Poisson equation

$$\Delta\psi = \xi.$$

Equation (1.1) is derived from either the shallow-water equations with rotation or the primitive equations by the so-called QG approximation, which assumes that the balance between the Coriolis force and the pressure gradient dominates the flow. This approximation corresponds to a singular perturbation that filters out the Poincaré waves, also called inertia-gravity waves (i.e., gravity waves modified by the presence of rotation). The QG equation (1.1) only supports Rossby waves, whose phase velocity—in the absence of forcing and dissipation, i.e., with a zero right-hand side—is comparable to the characteristic particle velocity; see Ghil and Childress [10] and Pedlosky [36]. Wolansky [48] studied the so-called barotropic, 2-D version of the QG model (1.1), while Wang [46] obtained results on existence, uniqueness, and long-time dynamics of the so-called baroclinic, three-dimensional (3-D) version.

The flow domain is a rectangular region $\Omega = \{(x, y);\ 0 \le x \le 2/a;\ 0 \le y \le 2\}$. For simplicity, we use here only the zonal component of the forcing. In an atmospheric model, this forcing represents—in the QG vorticity equation (1.1), in which there are no explicit thermodynamic effects—the transfer of angular momentum into midlatitudes due to the tropical Hadley cell (Lorenz [28]). Alternatively, one can think about a zonal forcing jet that would be in perfect geostrophic equilibrium with the pole-to-equator temperature gradient (Lorenz [27] and Ghil and Childress [10]).

In an oceanic model, the time-and-longitude independent forcing on the right-hand side of (1.1) is the curl of the wind stress

$$\nabla \times \tau = -\tau_0 \sin \pi y;$$

a wind stress of the form $\tau = -\tau_0(\cos \pi y, 0)$ mimics the annually averaged zonal wind distribution over the North Atlantic or North Pacific, with westerly (i.e., eastward) winds over the midlatitudes and easterlies in the tropics and polar latitudes.

The parameters ε and E are positive constants, called the Rossby and Ekman numbers, respectively. They measure the relative importance of nonlinearity and lateral diffusion. The effect of the bottom friction is neglected in this article. The Reynolds number is defined here as

$$R = \frac{\varepsilon}{E}.$$

The unknown streamfunction ψ satisfies periodic boundary conditions at $x = 0, 2/a$ and free-slip boundary conditions at $y = 0, 2$:

$$\begin{cases} \psi(t, 2/a, y) = \psi(t, 0, y), \\ \psi(t, x, 0) = \psi(t, x, 2) = 0;\ \partial_y^2\psi(t, x, 0) = \partial_y^2\psi(t, x, 2) = 0. \end{cases} \tag{1.2}$$

Along the meridional boundaries $y = 0, 2$, $\Delta\psi = \partial_y^2\psi = -\partial_y u$, and $u_y = 0$ corresponds to free slip along these boundaries. More general, "partial-slip" boundary conditions—intermediate between free slip and no slip ($u = -\psi_y = 0$)—are discussed in Appendix A of [18] for the 2-D shallow-water equations and in [14] for a 3-D version of the QG equations.

It is readily seen that (1.1) with (1.2) admits the steady-state solution $\psi_0 = \psi_0(y)$, where

$$\psi_0 = \frac{\tau_0}{\pi^4 E} \sin \pi y. \tag{1.3}$$

We shall take $\tau_0 = 1$ below for simplicity.

This midlatitude channel with zonal periodicity is a better model for the atmospheric problem that we outlined above than for the oceanic one. Still, the methods we apply might eventually be extended to the latter. As mentioned already, Wolansky [47] studied existence, uniqueness, and stability of stationary solutions of (1.1) in the presence of topography and free-surface effects, which are omitted here. Wolansky [48] showed, for a domain bounded by a closed streamline and nondivergent forcing, that sufficient conditions exist under which a branch of asymmetric stationary solutions bifurcates from the symmetric branch obtained when the domain, as well as the forcing, admits a symmetry group. His results were shown to apply in an infinite channel with the symmetry group of zonal translations.

The main objective of this article is to prove the following theorem.

THEOREM 1.1.

(i) *Let $a \geq \sqrt{3}/2$ and $E > 0$. Then* (1.1) *and* (1.2) *are linearly stable around ψ_0 for any Reynolds number $R > 0$, where $\psi_0 = \psi_0(y)$ is given by* (1.3).

(ii) *Let $\sqrt{3}/4 \leq a \leq \alpha_0$ and $E > 0$. Then there exists a critical Reynolds number $R_0 > 0$ for* (1.1) *and* (1.2). *Moreover* (1.1) *and* (1.2) *admit a nontrivial, time-periodic, classical solution ψ_R branching off ψ_0 as the Reynolds number R crosses R_0, provided that $E > c_0$ for some constant $c_0 > 0$.*

This theorem is based essentially on the eigenvalue analysis of the spectral problem with respect to the linearization of (1.1) and (1.2) around ψ_0, by using the continued-fraction method first introduced by Meshalkin and Sinai [33]. For the 2-D Navier–Stokes equations, without the Coriolis term and with periodic boundary conditions in both the x and y directions, stability and bifurcation were studied in [7, 17, 33]. For the 3-D Navier–Stokes equations, without the Coriolis term and with periodic boundary conditions in three directions, pitchfork bifurcation was studied by Chen and Wang [8]. They showed that the bifurcated branches exist for all Reynolds number values past the critical one and that the stationary solutions on these branches stay bounded as $R \to \infty$.

For 2-D incompressible viscous flows with periodic boundary conditions, Meshalkin and Sinai [33] deduced the linear stability of the steady state (1.3) when $a = 1$ with respect to all Reynolds numbers, Iudovich [17] proved the existence of steady-state bifurcation when $0 < a < 1$, and Chen and Price [7] obtained the existence of Hopf bifurcation for some a with $0 < a < \sqrt{3}/2$. Steady-state bifurcation, however, no longer occurs for any a under the free-slip boundary conditions described by (1.2).

In the present paper, the constant a is bounded from below by $\sqrt{3}/4$ for Hopf bifurcation to occur. In fact, for $0 < a < \sqrt{3}/4$ multiple pairs of eigenvalues crossing the imaginary line do occur, and by applying our approach here in a more sophisticated manner, the existence of time-periodic solutions can also be proven rigorously in this case; this will be reported elsewhere. Numerically, the multiple periodic solutions are found in section 6 here (Figures 6.2 and 6.3).

To prove assertion (ii) of Theorem 1.1 requires verifying a transversal crossing condition; see (1.7) below. This verification is rendered more difficult by the presence of the so-called β-term $\partial_x \psi$ in (1.1), which arises due to the meridional gradient of

the planetary vorticity [10, 36]. In order to prove the validity of this condition, we have to assume that $E > c_0$ for some constant c_0, although numerical experiments reveal the occurrence of Hopf bifurcation for large τ_0 but small E [14, 18, 40, 41, 42]. Technical difficulties prevent us from covering in Theorem 1.1 the range of x-periods given by $\alpha_0 < a < \sqrt{3}/2$, with $3/4 < \alpha_0 \simeq 0.8$. The overall situation for our QG channel flow, governed by (1.1), (1.2), is illustrated in Figure 1.1.

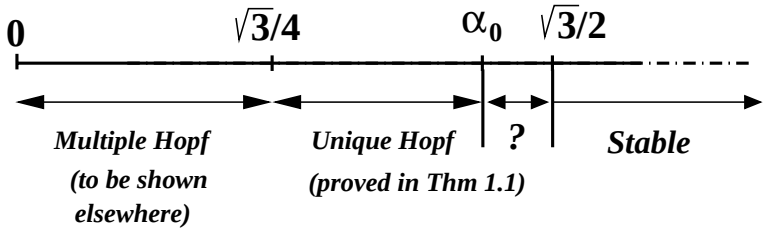


FIG. 1.1. *Range of values of the channel aspect ratio a for which various assertions hold. The equivalence between the spectral problem* (1.5) *with boundary conditions* (1.2) *and the difference equation of* (2.5)–(2.10) *holds for* $\sqrt{3}/4 \le a \le \sqrt{3}/2$ *(see Lemma* 2.2 *below). The numerical results in Figure* 6.2 *indicate that Hopf bifurcation also occurs for* $\alpha_0 \le a < \sqrt{3}/2$.

To prove Theorem 1.1, we decompose the streamfunction into a stationary part $\psi_0(x, y)$ given by (1.3) and a perturbation with exponential time dependence

$$\psi(x, y, t) = e^{\rho t}\phi(x, y), \tag{1.4}$$

where ϕ satisfies the boundary conditions (1.2). Introducing this solution into the governing equation and linearizing the latter with respect to the amplitude of the perturbation, we obtain the spectral problem

$$\rho\Delta\phi + \varepsilon\left(\partial_x\phi\,\partial_y\Delta\psi_0 - \partial_y\psi_0\,\partial_x\Delta\phi\right) + \partial_x\phi = E\Delta^2\phi.$$

Substituting the basic solution ψ_0 from (1.3) yields

$$L(\rho)\phi \doteq \left(E\Delta^2 - \partial_x + \frac{R}{\pi^3}\cos\pi y(\pi^2 + \Delta)\partial_x - \rho\Delta\right)\phi = 0. \tag{1.5}$$

The stability assertion (i) in the main theorem, i.e., the nonexistence of an eigenvalue ρ with $\operatorname{Re}\rho \ge 0$ for all $R > 0$, will be obtained by using the argument that Meshalkin and Sinai [33] first applied to the linear stability analysis of the 2-D Navier–Stokes equations.

The main effort is devoted to the proof of the bifurcation assertion (ii) of Theorem 1.1. Using the functional analysis framework of the Hopf bifurcation theorem in an infinite-dimensional setting (Joseph and Sattinger [20], Marsden and McCracken [30]), assertion (ii) of this theorem will follow if the following assertions can be shown to hold:

(a) There exists a critical Reynolds number $R_0 > 0$ and an eigenvalue $\rho = \rho(R_0)$ of (1.2) and (1.5) such that $\operatorname{Re}\rho(R_0) = 0$ and $\operatorname{Im}\rho(R_0) \ne 0$;

(b) this eigenvalue is simple, i.e.,

$$1 = \dim \bigcup_{n\ge 1} \left\{\phi \in H^4; L^n(\rho)\phi = 0\right\}; \tag{1.6}$$

(c) the transversal crossing condition

$$\text{Re}\,\frac{d\rho(R_0)}{dR} > 0. \tag{1.7}$$

Here H^4 denotes the complex space

$$H^4 = \{\phi \in L_2(\Omega);\ \Delta^2\phi \in L_2(\Omega),\ \phi \text{ satisfies (1.2)}\},$$

which contains the solution space of (1.1) and (1.2). This is a Hilbert space in the norm

$$\|\phi\|_{H^4} = \|\Delta^2\phi\|_{L_2(\Omega)}.$$

It is readily seen from the definition of H^4 that the bifurcating solution can be represented in the form of the Fourier expansion

$$\psi(x,y,t) = \sum_{m=-\infty}^{\infty}\sum_{n\geq 1-k}\sum_{k=0}^{1}(X_{m,n,k}(t)\cos ma\pi x + Y_{m,n,k}(t)\sin ma\pi x)\sin(n+k/2)\pi y. \tag{1.8}$$

This paper is organized as follows. The rigorous stability and bifurcation analysis of the problem is carried out in sections 2–5. Section 2 contains the proof of assertion (i) and a basic lemma on formulating the spectral problem. The proof of assertion (ii) is completed by combining section 3 on the existence of R_0 and $\rho(R_0)$, section 4 on the simplicity of the eigenvalue $\rho = \rho(R_0)$, and section 5 on the transversal crossing condition (1.7).

These analytical results are verified and complemented by numerical results in section 6. This section also contains comments on the geophysical significance of the periodic solutions. Brief remarks on the role of symmetry breaking in the bifurcation of time-periodic vs. stationary solutions follow in section 7.

2. Stability and equivalent formulation of the spectral problem. To prove the linear stability result of Theorem 1.1 and to obtain an equivalent formulation of the spectral problem, we follow the argument of Meshalkin and Sinai [33] by transforming the spectral problem into a difference equation, which is solved by continued-fraction methods.

As stated already in section 1, the free-slip boundary condition in (1.2) is equivalent to the condition $\phi = \Delta\phi = 0$ at $y = 0$ and $y = 2$. An application of this condition to (1.5) yields the generalized boundary condition

$$\Delta^n\phi = 0 \text{ at } y = 0 \text{ and } y = 2 \quad (n = 0, 1, 2, \ldots).$$

Thus the general expansion of the unknown function $\phi \in H^4$ for the spectral problem represented by (1.2) and (1.5) takes the form

$$\phi(x,y) = \sum_{m=-\infty}^{\infty}\sum_{n\geq 1-k}\sum_{k=0}^{1} e^{ima\pi x} i^n \phi_{n,m,k}\sin(n+k/2)\pi y, \quad i = \sqrt{-1}\ . \tag{2.1}$$

The fact that i^n appears explicitly in the coefficients of this expansion is only for convenience of notation in the derivation.

Note that no bifurcation can occur for Reynolds number R for which the corresponding eigenvalue ρ lies in the complex half-plane $\mathrm{Re}\rho < 0$. In order to obtain our stated bifurcation result, it is necessary to specify all of the eigenvalues ρ with $\mathrm{Re}\rho \geq 0$ and their corresponding eigenfunctions. However, the general expression of the eigenfunction in (2.1) is too complicated for this purpose. Fortunately, the eigenfunctions can be expressed in a simpler form, as given by the following lemma.

LEMMA 2.1. *Let $a \geq \sqrt{3}/4$ and let ρ be an eigenvalue of* (1.2) *and* (1.5) *such that $\mathrm{Re}\rho > -E\pi^2(a^2 + 1/4)$. Then $a < \sqrt{3}/2$ and the corresponding eigenfunction has the form*

$$\phi(x,y) = e^{\pm ia\pi x} \sum_{n=0}^{\infty} i^n \phi_n \sin(n+1/2)\pi y. \tag{2.2}$$

Proof. To begin with, we show that for given integers m and $k = 0, 1$, the subspace of H^4

$$\left\{ \phi \in H^4;\ \phi = \sum_{n \geq 1-k} e^{ima\pi x} i^n \phi_n \sin(n+k/2)\pi y \right\} \tag{2.3}$$

is invariant with respect to the spectral operator $\Delta^{-2}L(\rho)$, L being defined by (1.5).

For ϕ in the previous subspace, we see that

$$\begin{aligned}
&e^{-ima\pi x} L\phi \\
&= \sum_{n\geq 1-k} \{\pi^2[m^2a^2 + (n+k/2)^2]\rho - ima\pi\} i^n \phi_n \sin(n+k/2)\pi y \\
&\quad + \sum_{n\geq 1-k} E\pi^4[m^2a^2 + (n+k/2)^2]^2 i^n \phi_n \sin(n+k/2)\pi y \\
&\quad - R \sum_{n\geq 1-k} a[a^2 + (n+k/2)^2 - 1] i^{n+1} \phi_n \cos\pi y \sin(n+k/2)\pi y \\
&= \sum_{n\geq 1-k} \{\pi^2[m^2a^2 + (n+k/2)^2]\rho - ima\pi\} i^n \phi_n \sin(n+k/2)\pi y \\
&\quad + \sum_{n\geq 1-k} E\pi^4[m^2a^2 + (n+k/2)^2]^2 i^n \phi_n \sin(n+k/2)\pi y \\
&\quad - \frac{R}{2} \sum_{n\geq 2-k} ma[m^2a^2 + (n-1+k/2)^2 - 1] i^n \phi_{n-1} \sin(n+k/2)\pi y \\
&\quad + \frac{R}{2} \sum_{n\geq 0} ma[m^2a^2 + (n+1+k/2)^2 - 1] i^n \phi_{n+1} \sin(n+k/2)\pi y \\
&\quad + \frac{R}{2} ma(m^2a^2 - 3/4) i\phi_0 \sin(k/2)\pi y.
\end{aligned}$$

We thus have

$$L \sum_{n\geq 1-k} e^{ima\pi x} i^n \phi_n \sin(n+k/2)\pi y = \sum_{n\geq 1-k} e^{ima\pi x} i^n \psi_n \sin(n+k/2)\pi y, \tag{2.4}$$

where ψ_n is a linear combination of ϕ_{n-1}, ϕ_n, and ϕ_{n+1}. More precisely, for $n > 1-k$,

$$\begin{aligned}
\psi_n &= \{\pi^2[m^2a^2 + (n+k/2)^2]\rho - ima\pi + E\pi^4[m^2a^2 + (n+k/2)^2]^2\}\phi_n \\
&\quad - \frac{R}{2} ma[m^2a^2 + (n-1+k/2)^2 - 1]\phi_{n-1} + \frac{R}{2} ma[m^2a^2 + (n+1+k/2)^2 - 1]\phi_{n+1},
\end{aligned}$$

while

$$\begin{aligned}\psi_{1-k} = &\{\pi^2[m^2a^2 + (1-k/2)^2]\rho - ima\pi + E\pi^4[m^2a^2 + (1-k/2)^2]^2\}\phi_{1-k}\\ &+ \frac{R}{2}kma[m^2a^2 + (1+k/2)^2 - 1]\phi_1 + \frac{R}{2}kma(m^2a^2 - 3/4)i\phi_0\\ &+ \frac{R}{2}(1-k)ma(m^2a^2+3)\phi_2.\end{aligned}$$

This gives the invariance of the subspace (2.3), and thus we may assume that any eigenfunction ϕ is in this subspace for some integer m and some $k = 0, 1$.

We can now prove the desired assertion. Indeed, suppose that (ρ, ϕ) solves the spectral problem $L(\rho)\phi = 0$ and thus, by (2.4),

$$\psi_n = 0, \quad n \geq 1 - k. \tag{2.5}$$

If $m = 0$, this implies immediately $\phi_n = 0$ for $n \geq 1 - k$. If $m \neq 0$, the spectral problem yields

$$\sum_{n\geq 1-k} [m^2a^2 + (n + k/2)^2 - 1]\psi_n\bar{\phi}_n = 0,$$

where $\bar{\phi}_n$ is the complex conjugate of ϕ_n, and so its real part satisfies

$$\sum_{n\geq 1-k} [m^2a^2+(n+k/2)^2]\{\mathrm{Re}\rho+E\pi^2[m^2a^2+(n+k/2)^2]\}[m^2a^2+(n+k/2)^2-1]|\phi_n|^2 = 0. \tag{2.6}$$

This result, together with the condition $\mathrm{Re}\rho + E\pi^2(a^2 + 1/4) > 0$ and (2.5), implies

$$\begin{cases} \phi_n \equiv 0 & \text{when } k = 0, a > 0, -\infty < m < \infty,\\ \phi_n \equiv 0 & \text{when } k = 1, a > \sqrt{3}/2, -\infty < m < \infty,\\ \phi_n \equiv 0 & \text{when } k = 1, \sqrt{3}/4 \leq a \leq \sqrt{3}/2, |m| \geq 2,\end{cases}$$

since $m^2a^2 + (n + k/2)^2 - 1 \geq 0$.

The proof is thus complete. □

This lemma immediately gives the stability result contained in assertion (i) of Theorem 1.1.

As we shall see, it is important to realize that $k = 0$ corresponds to a function $\phi(x, y)$ that is antisymmetric about the axis $y = 1$ of the channel, while $\phi(x, y)$ is symmetric about this axis for $k = 1$. It follows from Lemma 2.1 that no bifurcation can arise in the problem governed by (1.1) and (1.2) from an antisymmetric instability.

To address the bifurcation problem, we thus consider the spectral equation (1.5)

$$\mathrm{Re}\ \rho > -E\pi^2(a^2 + 1/4), \quad \phi = \sum_{n=0}^{\infty} e^{\pm ia\pi x} i^n \phi_n \sin(n + 1/2)\pi y \in H^4$$

in the case $\sqrt{3}/4 \leq a < \sqrt{3}/2$. For simplicity, let

$$d_n = \frac{2\pi^2[a^2 + (n+1/2)^2]\rho - 2a\pi i + 2E\pi^4[a^2 + (n+1/2)^2]^2}{Ra[a^2 + (n+1/2)^2 - 1]}; \tag{2.7}$$

it follows, in particular, that $\mathrm{Re}\ d_0 < 0$ and $\mathrm{Re}\ d_n > 0$ with $n \geq 1$.

For the eigenfunction

$$\phi = \sum_{n=0}^{\infty} e^{ia\pi x} i^n \phi_n \sin(n+1/2)\pi y \in H^4, \tag{2.8}$$

let

$$\xi_n = a[a^2 + (n+k/2)^2 - 1]\phi_n. \tag{2.9}$$

The spectral problem $L(\rho)\phi = 0$ is, by (2.4), equivalent to $\psi_n = 0$; see (2.5). With definitions (2.7) and (2.9), the latter formulation becomes

$$\begin{cases} d_0\xi_0 + i\xi_0 + \xi_1 = 0, & n = 0, \\ d_n\xi_n - \xi_{n-1} + \xi_{n+1} = 0, & n \geq 1, \end{cases} \tag{2.10}$$

and (2.6) can be rewritten as

$$\sum_{n\geq 0} \operatorname{Re} d_n |\xi_n|^2 = 0. \tag{2.11}$$

It is readily seen that (2.10) implies that $\xi_n = 0$ for all $n \geq 0$ whenever there exists an $n_0 \geq 0$ for which $\xi_{n_0} = 0$.

Thus we may assume that $\xi_n \neq 0$ for $n \geq 0$. Hence, we obtain from (2.10) that

$$\frac{\xi_1}{\xi_0} = -d_0 - i, \quad \frac{\xi_n}{\xi_{n-1}} = \frac{1}{d_n + \dfrac{\xi_{n+1}}{\xi_n}}, \quad n \geq 1.$$

It follows therewith from (2.8) that $\rho = \rho(R)$ solves the following continued-fraction equation:

$$-d_0 - i = \cfrac{1}{d_1 + \cfrac{1}{d_2 + \cfrac{1}{\ddots}}}. \tag{2.12}$$

Moreover, we see that the spectral problem $L(\rho)\phi = 0$ is equivalent to the complex conjugate spectral problem $L(\bar{\rho})\bar{\phi} = 0$, with

$$\bar{\phi} = \sum_{n=0}^{\infty} e^{-ia\pi x} i^n (-1)^n \bar{\phi}_n \sin(n+1/2)\pi y \in H^4,$$

and thus $(\bar{\rho}, \bar{\phi})$ satisfies the complex conjugate of the three-term recursions (2.10), namely,

$$\begin{cases} \bar{d}_0\bar{\xi}_0 - i\bar{\xi}_0 + \bar{\xi}_1 = 0, & n = 0, \\ \bar{d}_n\bar{\xi}_n - \bar{\xi}_{n-1} + \bar{\xi}_{n+1} = 0, & n \geq 1. \end{cases} \tag{2.13}$$

Note that (2.10) and (2.13) are two distinct difference equations.

The above argument implies the following fundamental lemma.

LEMMA 2.2. *For* $\sqrt{3}/4 \le a < \sqrt{3}/2$, *the spectral problem described by* (1.5) *and* (1.2) *with unknown spectral solution* (ρ, ϕ) *such that*

$$\mathrm{Re}\rho > -E\pi^2(a^2+1/4), \quad \phi(x,y) = \sum_{n=0}^{\infty} e^{ia\pi x} i^n \phi_n \sin(n+1/2)\pi y \in H^4$$

is equivalent to the difference equation (2.10) *with* (d_n, ξ_n) *satisfying the constraints* (2.7), (2.9), (2.11), *and* (2.12). *Moreover, for a nontrivial eigenfunction* ϕ, *the corresponding sequence* $\{\xi_n\}$ *may be represented in the product form*

$$\begin{aligned} \xi_n &= c\gamma_1 \cdots \gamma_n, \quad n \ge 1, \\ \xi_0 &= c, \end{aligned}$$

where c *is an arbitrary complex constant and the factors* γ_n *are given by the infinite continued fractions*

$$\gamma_n = \cfrac{1}{d_n + \cfrac{1}{d_{n+1} + \cfrac{1}{\ddots}}}.$$

Furthermore, the spectral problem described by (1.5) *and* (1.2) *with another unknown solution* $(\bar{\rho}, \bar{\phi})$, *the complex conjugate of* (ρ, ϕ), *is equivalent to the difference equation* (2.13).

3. Existence of a critical Reynolds number. In this section, we show the existence of a critical Reynolds number R_0 and the existence of an eigenvalue $\rho(R)$ such that $\mathrm{Re}\,\rho(R_0) = 0$ and $\mathrm{Im}\,\rho(R_0) > 0$. We begin with the existence of the eigenvalue ρ, which may reach, and eventually cross, the imaginary axis of the complex plane.

LEMMA 3.1. *The spectral problem expressed by* (1.5) *and* (1.2) *admits a unique pair of complex conjugate eigenvalues* $\rho = \rho(R)$ *and* $\bar{\rho} = \bar{\rho}(R)$ *for any* $R > 0$ *and* $\sqrt{3}/4 \le a \le \alpha_0$, *such that* $\mathrm{Re}\,\rho > -E\pi^2(a^2+1/4)$.

Proof. From Lemma 2.2, we see readily that it suffices to show the existence and uniqueness of a function $\rho = \rho(R)$ that satisfies (2.7)–(2.12). Combining (2.7) and (2.12), we may write $\rho(R)$ as

$$\rho = -E\pi^2(a^2+1/4) + \frac{i2a\pi + iRa(3/4-a^2)}{2\pi^2(a^2+1/4)} + \cfrac{\dfrac{Ra(3/4-a^2)}{2\pi^2(a^2+1/4)}}{d_1 + \cfrac{1}{d_2 + \cfrac{1}{\ddots}}}. \tag{3.1}$$

To derive the existence of the unique pair of eigenvalues that satisfies the required strong inequality, we denote by $\Phi_R(\rho)$ the right-hand side of (3.1). It suffices then to show the existence of a fixed point of Φ_R in the complex plane $\mathbb{C}$. Indeed, since $\mathrm{Re}\rho \ge -E\pi^2(a^2+1/4)$, $R > 0$, and $a^2 < 3/4$, we have

$$\mathrm{Re}\Phi_R(\rho) > -E\pi^2(a^2+1/4)$$

and

$$|\Phi_R(\rho)| \le E\pi^2(a^2+1/4) + \frac{2a\pi + Ra(3/4-a^2)}{2\pi^2(a^2+1/4)} + \frac{Ra(3/4-a^2)}{2\pi^2(a^2+1/4)}\frac{1}{\mathrm{Re}\, d_1}$$
$$\le E\pi^2(a^2+1/4) + \frac{2a\pi + Ra(3/4-a^2)}{2\pi^2(a^2+1/4)} + \frac{R^2a^2(3/4-a^2)(a^2+5/4)}{8\pi^4(a^2+1/4)(a^2+9/4)E}. \tag{3.2}$$

Denoting this bound by K_R, we see that Φ_R maps the closed convex set

$$C_R = \{z \in \mathbb{C}; -E\pi^2(a^2+1/4) \le \mathrm{Re} z,\ |z| \le K_R\} \tag{3.3}$$

into itself. It follows from Brouwer's fixed point theorem that there exists a value $\rho(R)$ in this compact set such that $\Phi_R(\rho(R)) = \rho(R)$. This gives the existence of the desired pair of eigenvalues.

By Lemmas 2.1 and 2.2, we see that there is a one-to-one correspondence between each pair of complex conjugate eigenvalues of the spectral problem $L(\rho)\phi = 0$ and a solution ρ to the difference equation (2.10). In order to show that the pair of eigenvalues is unique, we suppose the existence of two pairs of complex conjugate solutions $\rho_j = \rho_j(R)$ and $\bar\rho_j = \bar\rho_j(R)$, with $\mathrm{Re}\rho_j > -E\pi^2(a^2+1/4)$ for $j = 1, 2$ and $R > 0$. This is equivalent to (2.10) admitting two solutions ρ_j for $j = 1, 2$ such that

$$\gamma_n(\rho_j) = \frac{1}{d_n + \gamma_{n+1}(\rho_j)} = \lim_{m\to\infty} \cfrac{1}{d_n + \cfrac{1}{d_{n+1} + \cfrac{1}{\ddots \cfrac{1}{d_{n+m}}}}}$$

for $d_n = d_n(\rho_j)$ defined by (2.7) and $\mathrm{Re}\rho_j > -E\pi^2(a^2+1/4)$. We have

$$\gamma_n(\rho_1) - \gamma_n(\rho_2) = -\gamma_n(\rho_1)\gamma_n(\rho_2)[d_n(\rho_1) - d_n(\rho_2) + \gamma_{n+1}(\rho_1) - \gamma_{n+1}(\rho_2)],$$

and so, by induction,

$$\gamma_1(\rho_1) - \gamma_1(\rho_2) = \sum_{n\ge1}(-1)^n[d_n(\rho_1) - d_n(\rho_2)]\eta_n(\rho_1)\eta_n(\rho_2)$$
$$= \sum_{n\ge1}(-1)^n \frac{2\pi[a^2 + (n+1/2)^2]}{Ra[a^2 + (n+1/2)^2 - 1]}\eta_n(\rho_1)\eta_n(\rho_2)(\rho_1 - \rho_2); \tag{3.4}$$

here $\{\eta_n(\rho_j)\}$ is now the solution specified by Lemma 2.2 such that $\eta_n = \xi_n$ and $\eta_0 = c = 1$. This yields, for $\rho_1 \ne \rho_2$,

$$|\rho_1 - \rho_2|$$
$$= \frac{Ra(3/4-a^2)}{2\pi^2(a^2+1/4)}|\gamma_1(\rho_1) - \gamma_1(\rho_2)|$$
$$\le \frac{3/4-a^2}{2(a^2+1/4)}\sum_{n\ge1}\frac{a^2+(n+1/2)^2}{[a^2+(n+1/2)^2-1]}(|\eta_n(\rho_1)|^2 + |\eta_n(\rho_2)|^2)|\rho_1 - \rho_2|$$
$$< \frac{3/4-a^2}{2a^2+1/2}\sum_{i=1}^{2}\sum_{n\ge1}\frac{[a^2+(n+1/2)^2]\{\mathrm{Re}\rho_i + E\pi^2[a^2+(n+1/2)^2]\}}{[\mathrm{Re}\rho_i + E\pi^2(a^2+1/4)][a^2+(n+1/2)^2-1]}|\eta_n(\rho_i)|^2|\rho_1 - \rho_2|$$
$$= |\rho_1 - \rho_2|,$$

where we have used (2.11). The strong inequality implies $\rho_1 = \rho_2$, which completes the proof of Lemma 3.1. ∎

LEMMA 3.2. *Let $\rho = \rho(R)$ with $R > 0$ be one of the two complex conjugate eigenvalues obtained in Lemma* 3.1. *Then we have* $\mathrm{Im}\rho(R) \neq 0$.

Proof. Assuming otherwise, i.e., $\mathrm{Im}\,\rho = 0$, we derive a contradiction. To this end, notice that

$$\begin{aligned}\left|\frac{\mathrm{Im}\,d_n}{\mathrm{Re}\,d_n}\right| &= \frac{\left|(a^2+(n+1/2)^2)\mathrm{Im}\,\rho\pi^2 - a\pi\right|}{\pi^2(a^2+(n+1/2)^2)\left\{E[a^2+(n+1/2)^2]\pi^2+\mathrm{Re}\,\rho\right\}} \\ &> \frac{\left|(a^2+(n+1+1/2)^2)\pi^2\mathrm{Im}\,\rho - a\pi\right|}{\pi^2(a^2+(n+1+1/2)^2)\left\{E[a^2+(n+1+1/2)^2]\pi^2+\mathrm{Re}\,\rho\right\}} \\ &= \left|\frac{\mathrm{Im}\,d_{n+1}}{\mathrm{Re}\,d_{n+1}}\right|.\end{aligned}$$

This, together with induction on n, implies that

$$\frac{|\mathrm{Im}\,\gamma_1|}{\mathrm{Re}\,\gamma_1} < \left|\frac{\mathrm{Im}\,d_1}{\mathrm{Re}\,d_1}\right|.$$

Thus we have

$$\left|\frac{\mathrm{Im}\,(-d_0 - i)}{-\mathrm{Re}\,d_0}\right| < \left|\frac{\mathrm{Im}\,d_1}{\mathrm{Re}\,d_1}\right|,$$

and so

$$\frac{|2(a^2+1/4)\pi^2\mathrm{Im}\,\rho - 2a\pi - Ra(3/4-a^2)|}{(a^2+1/4)\mathrm{Re}\,\rho + E(a^2+1/4)^2\pi^2} < \frac{2|(a^2+9/4)\pi^2\mathrm{Im}\,\rho - a\pi|}{(a^2+9/4)\mathrm{Re}\,\rho + E(a^2+9/4)^2\pi^2}.$$

Therefore,

$$\begin{aligned}\frac{|2(a^2+1/4)\pi^2\mathrm{Im}\,\rho - 2a\pi - Ra(3/4-a^2)|}{2|(a^2+9/4)\pi^2\mathrm{Im}\,\rho - a\pi|} &< \frac{(a^2+1/4)[\mathrm{Re}\,\rho + E(a^2+1/4)\pi^2]}{(a^2+9/4)[\mathrm{Re}\,\rho + E(a^2+9/4)\pi^2]} \\ &\leq \frac{a^2+1/4}{a^2+9/4},\end{aligned}$$

where we have used the condition $\mathrm{Re}\,\rho + E(a^2+1/4)\pi^2 > 0$. This implies

$$1 + \frac{Ra(3/4-a^2)}{2|(a^2+9/4)\pi^2\mathrm{Im}\,\rho - a\pi|} < \frac{a^2+1/4}{a^2+9/4} + \frac{2\pi^2|\mathrm{Im}\,\rho|}{|(a^2+9/4)\pi^2\mathrm{Im}\,\rho - a\pi|} < 1,$$

which leads to a contradiction and hence $\mathrm{Im}\,\rho > 0$. The proof of Lemma 3.2 is thus complete. ∎

LEMMA 3.3. *For one of the two eigenvalues obtained in Lemma* 3.1, *we have*

$$-\frac{1}{10} < \frac{2\pi^2}{a}\limsup_{R\to\infty}\frac{\mathrm{Im}\rho(R)}{R} < 2. \tag{3.5}$$

Proof. By (2.11), (2.12), and Lemma 2.2, we have

$$-\mathrm{Re}\,d_0|\xi_0|^2 > \mathrm{Re}\,d_1|\xi_1|^2,$$

or

$$-\mathrm{Re}\,d_0 > \mathrm{Re}\,d_1|d_0+i|^2.$$

Together with (2.7), this implies

$$\frac{(a^2+1/4)^2}{3/4-a^2} > \frac{(a^2+9/4)^2}{a^2+5/4}\left[\frac{2\pi^2(a^2+1/4)\mathrm{Im}\,\rho - 2a\pi}{Ra(3/4-a^2)} - 1\right]^2,$$

and so

$$\frac{(a^2+1/4)^2(a^2+5/4)}{(3/4-a^2)(a^2+9/4)^2} > \left[\frac{2\pi^2(a^2+1/4)}{a(3/4-a^2)}\limsup_{R\to\infty}\frac{\mathrm{Im}\,\rho}{R} - 1\right]^2.$$

Therefore, the double inequality (3.5) holds for $\sqrt{3}/4 \le a < \sqrt{3}/2$ and the proof of Lemma 3.3 is thus complete. □

Now we prove the main result of this section.

THEOREM 3.4. *Let $\rho(R)$ be either one of the two complex conjugate eigenvalues obtained in Lemma* 3.1. *Then there exists a critical Reynolds number R_0 such that* $\mathrm{Re}\rho(R_0) = 0$ *and* $\mathrm{Im}\rho(R_0) \neq 0$.

Proof. First, we prove the smoothness of $\rho(R)$ by using the implicit function theorem. Define the function

$$F(\rho, R) = i + d_0 + \cfrac{1}{d_1 + \cfrac{1}{d_2 + \cfrac{1}{\ddots}}} \tag{3.6}$$

with $d_n = d_n(\rho, R)$ given by (2.7). We see immediately that $F(\rho(R), R) = 0$, due to (2.12) and Lemma 3.1. It follows from (3.4) that

$$\frac{\partial F(\rho,R)}{\partial\rho} = \sum_{n\ge 0}(-1)^n\frac{\partial d_n}{\partial\rho}\eta_n^2 = \sum_{n\neq 0}(-1)^n\frac{2\pi[a^2+(n+1/2)^2]}{Ra[(a^2+(n+1/2)^2-1]}\eta_n^2, \tag{3.7}$$

where $\{\eta_n\}$ is the solution specified by Lemma 2.2 such that $\eta_0 = 1$. Hence we have, by (2.11), that

$$\begin{aligned}\left|\frac{\partial F(\rho,R)}{\partial\rho}\right| &\ge \frac{2\pi^2(a^2+1/4)}{Ra(3/4-a^2)}|\eta_0|^2 - \sum_{n\ge 1}\frac{2\pi^2[a^2+(n+1/2)^2]}{Ra[a^2+(n+1/2)^2-1]}|\eta_n|^2 \\ &= \frac{2\pi^2}{Ra}\sum_{n\ge 1}\frac{[a^2+(n+1/2)^2][\mathrm{Re}\rho + E\pi^2(a^2+(n+1/2)^2)]}{[\mathrm{Re}\rho + E\pi^2(a^2+1/4)][a^2+(n+1/2)^2-1]}|\eta_n|^2 \\ &\quad - \sum_{n\ge 1}\frac{2\pi^2[a^2+(n+1/2)^2]}{Ra[a^2+(n+1/2)^2-1]}|\eta_n|^2 \\ &= \frac{2\pi^4 E}{\mathrm{Re}\rho + E\pi^2(a^2+1/4)}\sum_{n\ge 1}\frac{[a^2+(n+1/2)^2](n+1)n}{Ra[a^2+(n+1/2)^2-1]}|\eta_n|^2 > 0.\end{aligned} \tag{3.8}$$

Thus, by the implicit function theorem, $\rho = \rho(R)$ is continuously differentiable.

Next, letting $R \to 0$ in (2.7) and (2.12), we see that

$$\lim_{R\to 0}\mathrm{Re}\,\rho(R) = -E\pi^2(a^2+1/4).$$

Finally, by Lemma 3.2 and the smoothness of $\rho(R)$, it suffices to show that

$$\limsup_{R\to\infty}\mathrm{Re}\,\rho(R) > 0.$$

To do so, we suppose that $\limsup_{R\to\infty} \mathrm{Re}\,\rho(R) \le 0$, which will lead to a contradiction.

Without loss of generality, by Lemma 3.3, we may suppose that

$$\lim_{R\to\infty} \mathrm{Re}\,\rho(R) = \mu, \quad \lim_{R\to\infty} \frac{2\pi^2 \mathrm{Im}\,\rho(R)}{Ra} = \nu \tag{3.9}$$

for some constants μ and ν. Otherwise, we may consider instead a subsequence $\{R_n\}$ that converges to these values.

Applying (2.12) and Lemma 2.2 yields

$$-d_0 - i = \frac{1}{d_1 + \dfrac{1}{d_2 + \gamma_3}}. \tag{3.10}$$

Hence

$$\begin{aligned} \lim_{R\to\infty} \gamma_3 &= \frac{1}{-\dfrac{1}{\lim\limits_{R\to\infty} i\mathrm{Im}\,d_0 + i} - \lim\limits_{R\to\infty} i\mathrm{Im}\,d_1} - \lim_{R\to\infty} i\mathrm{Im}\,d_2 \\ &= i\frac{\lim\limits_{R\to\infty} \mathrm{Im}\,d_0 + 1}{\lim\limits_{R\to\infty} \mathrm{Im}\,d_1(\lim\limits_{R\to\infty} \mathrm{Im}\,d_0 + 1) - 1} - \lim_{R\to\infty} i\mathrm{Im}\,d_2, \end{aligned} \tag{3.11}$$

or

$$\lim_{R\to\infty} (\mathrm{Im}\,\gamma_3 + \mathrm{Im}\,d_2) = \frac{\lim\limits_{R\to\infty} \mathrm{Im}\,d_0 + 1}{\lim\limits_{R\to\infty} \mathrm{Im}\,d_1(\mathrm{Im}\,d_0 + 1) - 1}. \tag{3.12}$$

Furthermore, it follows from (3.10) that

$$-d_0 - i = \frac{d_2 + \gamma_3}{d_1(d_2 + \gamma_3) + 1} = \frac{(d_2 + \gamma_3)(\overline{d_1(d_2 + \gamma_3)} + 1)}{|d_1(d_2 + \gamma_3) + 1|^2},$$

and thus

$$\begin{aligned} &-\mathrm{Re}\,d_0 |d_1(d_2 + \gamma_3) + 1|^2 \\ &= \mathrm{Re}\,(d_2 + \gamma_3)\mathrm{Re}\,(d_1(d_2 + \gamma_3) + 1) - \mathrm{Im}\,(d_2 + \gamma_3)\mathrm{Im}\,\overline{d_1(d_2 + \gamma_3)} \\ &= \mathrm{Re}\,(d_2 + \gamma_3)(\mathrm{Re}\,d_1 \mathrm{Re}\,(d_2 + \gamma_3) + 1) + \mathrm{Re}\,d_1(\mathrm{Im}\,d_2 + \mathrm{Im}\,\gamma_3)^2. \end{aligned} \tag{3.13}$$

Multiplying the last equation by R and passing to the limit $R \to \infty$, we obtain, after using (3.9), (3.11), (3.12), and the positivity of $\mathrm{Re}\,\gamma_3$,

$$\begin{aligned} \lim_{R\to\infty} R|\mathrm{Re}\,d_0| &= \frac{\lim\limits_{R\to\infty} R\mathrm{Re}\,(d_2 + \gamma_3) + \lim\limits_{R\to\infty} R\mathrm{Re}\,d_1 \lim\limits_{R\to\infty} (\mathrm{Im}\,d_2 + \mathrm{Im}\,\gamma_3)^2}{|\lim\limits_{R\to\infty} \mathrm{Im}\,d_1 \lim\limits_{R\to\infty} (\mathrm{Im}\,d_2 + \mathrm{Im}\,\gamma_3) - 1|^2} \\ &\ge \frac{\lim\limits_{R\to\infty} R\mathrm{Re}\,d_2 + \lim\limits_{R\to\infty} R\mathrm{Re}\,d_1 \lim\limits_{R\to\infty} (\mathrm{Im}\,d_2 + \mathrm{Im}\,\gamma_3)^2}{|\lim\limits_{R\to\infty} \mathrm{Im}\,d_1 \lim\limits_{R\to\infty} (\mathrm{Im}\,d_2 + \mathrm{Im}\,\gamma_3) - 1|^2} \\ &= \lim_{R\to\infty} R\mathrm{Re}\,d_1(\lim_{R\to\infty} \mathrm{Im}\,d_0 + 1)^2 \\ &\quad + \lim_{R\to\infty} R\mathrm{Re}\,d_2(1 - \lim_{R\to\infty} \mathrm{Im}\,d_1(\mathrm{Im}\,d_0 + 1))^2. \end{aligned}$$

That is,

$$\begin{aligned}
&\frac{a^2+1/4}{3/4-a^2}\left[\mu+E(a^2+1/4)\pi^2\right]\\
&\geq \frac{a^2+9/4}{a^2+5/4}\left[\mu+E(a^2+9/4)\pi^2\right]\left(1-\frac{a^2+1/4}{3/4-a^2}\nu\right)^2\\
&\quad+\frac{a^2+25/4}{a^2+21/4}\left[\mu+E(a^2+25/4)\pi^2\right]\left[\frac{a^2+9/4}{a^2+5/4}\left(1-\frac{a^2+1/4}{3/4-a^2}\nu\right)\nu-1\right]^2.
\end{aligned}$$

This, together with the condition $-E(a^2+1/4)\pi^2 \leq \mu \leq 0$, implies

$$\begin{aligned}
\frac{(a^2+1/4)^2}{3/4-a^2} &\geq \frac{(a^2+9/4)^2}{a^2+5/4}\left(1-\frac{a^2+1/4}{3/4-a^2}\nu\right)^2\\
&\quad+\frac{(a^2+25/4)^2}{a^2+21/4}\left[1-\frac{a^2+9/4}{a^2+5/4}\left(1-\frac{a^2+1/4}{3/4-a^2}\nu\right)\nu\right]^2\\
&> \frac{(a^2+1/4)^2}{3/4-a^2}
\end{aligned} \tag{3.14}$$

for $\sqrt{3}/4 \leq a \leq \alpha_0$, with $\alpha_0 \simeq 0.80$. This leads to a contradiction and hence $\lim_{R\to\infty} \operatorname{Re}\rho > 0$. The proof of Theorem 3.4 is thus complete. ∎

4. Spectral simplicity condition. This section is devoted to the simplicity of each of the two complex conjugate eigenvalues that cross the imaginary axis. We prove the following theorem.

THEOREM 4.1. *Let the critical Reynolds number R_0 and the eigenvalue $\rho = \rho(R_0)$ with $\operatorname{Im}\rho > 0$ be as shown to exist in Theorem* 3.4. *Then this eigenvalue is simple, i.e.,*

$$\dim \bigcup_{n\geq 1} \left\{\phi \in H^4;\ L^n\phi = 0\right\} = 1, \tag{4.1}$$

where $L = L(\rho)$ is the linear operator defined in (1.5).

Proof. We introduce the invariant subspaces of the spectral problem (1.5) for any integer m:

$$E_{m,k} = \left\{ \phi \in H^4;\ \psi(x,y) = \sum_{n\geq 1-k} i^n \phi_n e^{ima\pi x} \sin(n+k/2)\pi y \right\},\ \ k = 0, 1.$$

Hence the simplicity condition holds true provided that the following assertions are valid:

$$\dim \bigcup_{n\geq 1} \{\phi \in E_{m,k};\ L^n\phi = 0\} = 0 \tag{4.2}$$

for $(m,k) \neq (1,1)$, and

$$\dim \bigcup_{n\geq 1} \{\phi \in E_{1,1};\ L^n\phi = 0\} = 1. \tag{4.3}$$

Equation (4.2) follows immediately from Lemmas 2.1 and 2.2. Thus it remains to prove (4.3). We first note from Lemma 2.2 and Theorem 3.4 that

$$\dim\{\phi \in E_{1,1};\ L\phi = 0\} = 1.$$

By induction, it suffices to show

$$\dim\left\{\phi \in E_{1,1}\,;\ L^2\phi = 0\right\} = 1. \tag{4.4}$$

Indeed, rewrite the equation $L^2\phi = 0$ as

$$L\chi = 0, \tag{4.5}$$

after setting

$$L\phi = \chi \tag{4.6}$$

for some $\phi \in E_{1,1}$. It remains to show that $\phi \equiv 0$. Following the derivation of (2.10), we obtain the equivalent formulation of (4.5) and (4.6) in terms of two coupled difference equations:

$$\begin{cases} d_n\xi_n' - \xi_{n-1}' + \xi_{n+1}' = \xi_n, & n \geq 1, \\ d_n\xi_n - \xi_{n-1} + \xi_{n+1} = 0, & n \geq 1, \\ d_0\xi_0' + i\xi_0' + \xi_1' = \xi_0, & n = 0, \\ d_0\xi_0 + i\xi_0 + \xi_1 = 0, & n = 0; \end{cases} \tag{4.7}$$

here d_n is defined by (2.7) and $\{\xi_n/d_n\}$, $\{\xi_n'\} \in l_2^2$, while l_2^2 denotes the Hilbert space

$$l_2^2 = \left\{ \{\xi_n\};\ \|\{\xi_n\}\|_{l_2^2}^2 = \sum_{n\geq 0} n^2|\xi_n|^2 < \infty \right\}.$$

Define an operator $M : l_2^2 \mapsto l_2^2$ such that $M\{\xi_n\} = \{\eta_n\}$ with

$$\eta_n = \frac{1}{d_n}(\xi_{n+1} - \xi_{n-1}),\ \ \eta_0 = \frac{1}{d_0}(i\xi_0 + \xi_1),\ \ n \geq 1.$$

Equation (4.7) becomes

$$\begin{cases} (1+M)\{\xi_n'\} = \left\{\dfrac{\xi_n}{d_n}\right\}, \\ (1+M)\{\xi_n\} = 0. \end{cases} \tag{4.8}$$

We see that M is compact in l_2^2. It follows from Riesz–Schauder theory (also called the Fredholm alternative principle; see Theorem 5.3 in [15]) that (4.8) is solvable if and only if

$$\sum_{n\geq 0} \frac{\xi_n\bar{\zeta}_n}{d_n} = 0 \tag{4.9}$$

whenever $\{\zeta_n\} \in l_2^2$ is a nontrivial solution of the dual equation

$$(1 + M^*)\{\zeta_n\} = 0, \tag{4.10}$$

where M^* is the dual operator of M.

Such a nontrivial solution of (4.10) is given by

$$\zeta_n + \frac{\zeta_{n-1}}{\bar{d}_{n-1}} - \frac{\zeta_{n+1}}{\bar{d}_{n+1}} = 0, \quad n \geq 1,$$
$$\zeta_0 - \frac{i\zeta_0}{\bar{d}_0} - \frac{\zeta_1}{\bar{d}_1} = 0, \quad n = 0.$$

This becomes, by setting $\hat{\zeta}_n = (-1)^n \bar{\zeta}_n / d_n$,

$$d_n\hat{\zeta}_n - \hat{\zeta}_{n-1} + \hat{\zeta}_{n+1} = 0, \quad n \geq 1,$$
$$d_0\hat{\zeta}_0 + i\hat{\zeta}_0 + \hat{\zeta}_1 = 0, \quad n = 0.$$

By Lemma 2.2 and Theorem 3.4, we have

$$\hat{\zeta}_n = (-1)^n \frac{\bar{\zeta}_n}{d_n} = c\xi_n, \quad n \geq 0,$$

for some constant $c \neq 0$. Thus (4.9) becomes

$$\sum_{n\geq 0} \frac{\xi_n \bar{\zeta}_n}{d_n} = c \sum_{n\geq 0} (-1)^n \xi_n^2 = 0.$$

Hence, it follows from (2.11) that

$$\begin{aligned} 0 &= \left| \sum_{n\geq 0} (-1)^n \xi_n^2 \right| \\ &\geq |\xi_0|^2 - \sum_{n\geq 1} |\xi_n|^2 \\ &= \frac{1}{|\operatorname{Re} d_0|} \sum_{n\geq 1} (\operatorname{Re} d_n - |\operatorname{Re} d_0|)|\xi_n|^2. \end{aligned}$$

Since $\operatorname{Re} d_n - |\operatorname{Re} d_0| > 0$ for $\sqrt{3}/4 \leq a < \sqrt{3}/2$, we thus have $\{\xi_n\} = 0$ and hence (4.4). The proof of Theorem 4.1 is complete. □

5. Transversal crossing condition. The objective of this section is to show that the pair of eigenvalues $\{\rho(R), \bar{\rho}(R)\}$ crosses the imaginary axis transversally and away from the origin. This result reads as follows.

THEOREM 5.1. *Let $\rho = \rho(R_0)$ and R_0 be as characterized in Lemma* 2.2 *and Theorems* 3.4 *and* 4.1. *Then the transversal crossing condition* (1.7) *holds, provided that $E > c_0$ for some constant $c_0 > 0$.*

Proof. Recalling the definition of the function $F = F(\rho, R)$ in (3.6), we see that the eigenvalue $\rho = \rho(R)$ satisfies the equation $F(\rho(R), R) = 0$, and so

$$\frac{\partial F}{\partial \rho} \frac{\mathrm{d}\rho(R)}{dR} + \frac{\partial F}{\partial R} = 0.$$

It follows from the derivation of (3.7) that

$$\frac{\partial F}{\partial R} = -\frac{1}{R} \sum_{n\geq 0} (-1)^n d_n \eta_n^2.$$

Thus, by (2.7) with $R = R_0$, (3.7), and (3.8), we have

$$\pi R_0 \mathrm{Re}\,\frac{\mathrm{d}\rho(R_0)}{\mathrm{d}R} = \pi \mathrm{Re}\,\frac{\displaystyle\sum_{n\geq 0}(-1)^n d_n \eta_n^2}{\displaystyle\sum_{n\geq 0}(-1)^n \frac{\partial d_n}{\partial \rho}\eta_n^2}$$
$$= \mathrm{Re}\,\frac{\displaystyle\sum_{n\geq 0}(-1)^n \frac{\pi^3 E[a^2+(n+1/2)^2]^2 - ia}{a^2+(n+1/2)^2-1}\eta_n^2}{\displaystyle\sum_{n\geq 0}(-1)^n \frac{a^2+(n+1/2)^2}{a^2+(n+1/2)^2-1}\eta_n^2}.$$

This equals, after setting $a^2 + (n+1/2)^2 = \beta_n$,

$$\mathrm{Re}\,\frac{\displaystyle\sum_{n\geq 0}(-1)^n \frac{\pi^3 E\beta_n^2 - ia}{\beta_n - 1}\eta_n^2}{\displaystyle\sum_{n\geq 0}(-1)^n \frac{\beta_n}{\beta_n - 1}\eta_n^2}$$
$$= \pi^3 E\beta_0 + \sum_{n\geq 1}(-1)^n \frac{\pi^3 E\beta_n(\beta_n-\beta_0)}{\beta_n - 1}\mathrm{Re}\,(\xi_n^2) - \frac{a}{\beta_0}\sum_{n\geq 1}(-1)^n \frac{\beta_n-\beta_0}{\beta_n-1}\mathrm{Im}\,(\xi_n^2),$$

where the solution $\{\xi_n\}$ is chosen such that

$$\xi_0^2 = \frac{1}{\displaystyle\sum_{n\geq 0}(-1)^n \frac{\beta_n}{\beta_n-1}\eta_n^2}. \tag{5.1}$$

We thus have

$$\pi R_0 \mathrm{Re}\,\frac{\mathrm{d}\rho(R_0)}{\mathrm{d}R}$$
$$= \pi^3 E\beta_0 \left[1 - \frac{1}{\beta_0}\sum_{n\geq 1}\frac{\beta_n(\beta_n-\beta_0)}{\beta_n-1}|\xi_n|^2\right] - \frac{a}{\beta_0}\sum_{n\geq 1}(-1)^n\frac{\beta_n-\beta_0}{\beta_n-1}\mathrm{Im}\,(\xi_n^2)$$
$$+ \pi^3 E\sum_{n\geq 1}\frac{\beta_n(\beta_n-\beta_0)}{\beta_n-1}[|\xi_n|^2 + (-1)^n\mathrm{Re}\,(\xi_n^2)].$$

Furthermore, by (2.11) and (5.1), we have

$$1 - \frac{1}{\beta_0}\sum_{n\geq 1}\frac{\beta_n(\beta_n-\beta_0)}{\beta_n-1}|\xi_n|^2$$
$$= 1 - \frac{1}{\beta_0}\sum_{n\geq 1}\frac{\beta_n^2}{\beta_n-1}|\xi_n|^2 + \sum_{n\geq 1}\frac{\beta_n}{\beta_n-1}|\xi_n|^2$$

$$
\begin{aligned}
&= |\xi_0|^2 \left[\left| \sum_{n\geq 0} (-1)^n \frac{\beta_n}{\beta_n - 1} \eta_n^2 \right| + \sum_{n\geq 1} \frac{\beta_n}{\beta_n - 1} |\eta_n|^2 + \frac{\beta_0}{\beta_0 - 1} \right] \\
&\geq |\xi_0|^2 \left[\left| \sum_{n\geq 0} (-1)^n \frac{\beta_n}{\beta_n - 1} \mathrm{Re}\,(\eta_n^2) \right| + \sum_{n\geq 1} \frac{\beta_n}{\beta_n - 1} |\eta_n|^2 + \frac{\beta_0}{\beta_0 - 1} \right] \\
&\geq |\xi_0|^2 \sum_{n\geq 1} \frac{\beta_n}{\beta_n - 1} |\eta_n|^2 - |\xi_0|^2 \left| \sum_{n\geq 1} (-1)^n \frac{\beta_n}{\beta_n - 1} \mathrm{Re}\,(\eta_n^2) \right| \\
&\geq |\xi_0|^2 \frac{\beta_1}{\beta_1 - 1} (|\eta_1|^2 - |\mathrm{Re}\,(\eta_1^2)|)
\end{aligned}
$$

and

$$
\begin{aligned}
\left| \sum_{n\geq 1} (-1)^n \frac{\beta_n - \beta_0}{\beta_n - 1} \mathrm{Im}\,(\xi_n^2) \right| &\leq \sum_{n\geq 1} \frac{\beta_n - \beta_0}{\beta_n - 1} |\mathrm{Im}\,(\xi_n^2)| \\
&\leq \frac{1}{\beta_1} \sum_{n\geq 1} \frac{\beta_n^2}{\beta_n - 1} |\xi_n|^2 = \frac{\beta_0^2}{\beta_1(1-\beta_0)} |\xi_0|^2.
\end{aligned}
$$

Collecting terms while using (2.12) and Lemma 2.2, we have

$$
\begin{aligned}
&\frac{R_0}{\pi^2 |\xi_0|^2} \mathrm{Re}\,\frac{\mathrm{d}\rho(R_0)}{\mathrm{d}R} \\
&\geq E\beta_0 (|\eta_1|^2 - |\mathrm{Re}\,\eta_1^2|) + E(\beta_1 - \beta_0) \left(|\eta_1|^2 - \frac{\mathrm{Re}\,\xi_1^2}{|\xi_0|^2} \right) - \frac{a\beta_0}{\pi^3 \beta_1 (1-\beta_0)} \\
&= E\beta_0 (|d_0 + i|^2 - |\mathrm{Re}\,(d_0 + i)^2|) \\
&\quad + 2E \left(|d + i|^2 - \frac{\mathrm{Re}\,[(d_0 + i)^2 \xi_0^2]}{|\xi_0|^2} \right) - \frac{a\beta_0}{\pi^3 \beta_1 (1-\beta_0)}.
\end{aligned}
$$

Let us now apply the following lemma, whose proof will be given at the end of this section.

LEMMA 5.2. *One of the following two estimates,*

$$
\liminf_{E\to\infty} \frac{E}{R_0} > 0 \tag{5.2}
$$

or

$$
\liminf_{E\to\infty} |\mathrm{Re}\ d_0| > 0, \tag{5.3}
$$

holds true.

If $|1 + \mathrm{Im}\, d_0| > |\mathrm{Re}\, d_0|$, we have

$$
\begin{aligned}
\frac{R_0}{\pi^2 |\xi_0|^2} \mathrm{Re}\,\frac{\mathrm{d}\rho(R_0)}{\mathrm{d}R} &\geq E\beta_0 (|d_0 + i|^2 - |\mathrm{Re}\,(d_0 + i)^2|) - \frac{a\beta_0}{\pi^3 \beta_1 (1-\beta_0)} \\
&= 2E\beta_0 (\mathrm{Re}\, d_0)^2 - \frac{a\beta_0}{\pi^3 \beta_1 (1-\beta_0)};
\end{aligned}
$$

the latter is positive due to Lemma 5.2, after letting $E > c_0$ for some constant $c_0 > 0$.

If $|\mathrm{Re}\, d_0| \geq |1 + \mathrm{Im}\, d_0| > |\mathrm{Re}\, d_0|/4$, we see that

$$\begin{aligned}\frac{R_0}{\pi^2|\xi_0|^2}\mathrm{Re}\,\frac{\mathrm{d}\rho(R_0)}{\mathrm{d}R} &\geq E\beta_0(|d_0+i|^2 - |\mathrm{Re}\,(d_0+i)^2|) - \frac{a\beta_0}{\pi^3\beta_1(1-\beta_0)}\\ &= 2E\beta_0(1+\mathrm{Im}\, d_0)^2 - \frac{a\beta_0}{\pi^3\beta_1(1-\beta_0)}\\ &\geq \frac{1}{2}E\beta_0(\mathrm{Re}\, d_0)^2 - \frac{a\beta_0}{\pi^3\beta_1(1-\beta_0)} > 0,\end{aligned}$$

which is positive as well for E large enough.

For the remaining case $|\mathrm{Re}\, d_0|/4 \geq |1 + \mathrm{Im}\, d_0| \geq 0$, we see that $-\mathrm{Re}\,(\xi_0^2) > 0$ due to (2.11) and (5.1). Hence we have

$$\begin{aligned}&\frac{R_0}{\pi^2|\xi_0|^2}\mathrm{Re}\,\frac{\mathrm{d}\rho(R_0)}{\mathrm{d}R}\\ &\geq 2E\left\{|d+i|^2 - \frac{\mathrm{Re}\,[(d_0+i)^2\xi_0^2]}{|\xi_0|^2}\right\} - \frac{a\beta_0}{\pi^3\beta_1(1-\beta_0)}\\ &\geq 2E\left[|d+i|^2 + \frac{2(\mathrm{Re}\, d_0)(1+\mathrm{Im}\, d_0)\mathrm{Im}\,(\xi_0^2)}{|\xi_0|^2}\right] - \frac{a\beta_0}{\pi^3\beta_1(1-\beta_0)}\\ &\geq E(\mathrm{Re}\, d_0)^2 - \frac{a\beta_0}{\pi^3\beta_1(1-\beta_0)}.\end{aligned}$$

This also implies the desired assertion by letting E be large enough. The proof of Theorem 5.1 is complete, subject to proving Lemma 5.2. □

Proof of Lemma 5.2. Let us first note, from (2.11) and Lemma 2.2, that

$$\mathrm{Re}\, d_1|\eta_1|^2 < -\mathrm{Re}\, d_0,$$

or, by (2.12),

$$\frac{\beta_0^2(\beta_1-1)}{\beta_1^2(1-\beta_0)} > |d_0+i|^2 = \left[\frac{2\pi^2\beta_0\mathrm{Im}\,\rho - a\pi}{R_0 a(1-\beta_0)} - 1\right]^2 + \left[\frac{2\pi^4 E\beta_0^2}{R_0 a(1-\beta_0)}\right]^2. \tag{5.4}$$

This implies

$$|d_1| + |d_2| < c_1, \tag{5.5}$$

and $R_0 > c_2 E$ for some positive constants c_1 and c_2, independent of E and R_0. Thus $E \to \infty$ implies $R_0 \to \infty$.

On the contrary, we suppose that

$$\liminf_{E\to\infty} \frac{E}{R_0} = 0, \tag{5.6}$$

which will lead to a contradiction. Indeed, by following the final step in the proof of Theorem 3.4, we use (2.12) or (3.10) to obtain

$$\begin{aligned}\lim_{E\to\infty}\gamma_3 &= \frac{1}{-\dfrac{1}{\lim\limits_{E\to\infty} i\mathrm{Im}\, d_0 + i} - \lim\limits_{E\to\infty} i\mathrm{Im}\, d_1} - \lim_{E\to\infty} i\mathrm{Im}\, d_2\\ &= i\frac{\lim\limits_{E\to\infty}\mathrm{Im}\, d_0 + 1}{\lim\limits_{E\to\infty}\mathrm{Im}\, d_1(\lim\limits_{E\to\infty}\mathrm{Im}\, d_0 + 1) - 1} - \lim_{E\to\infty} i\mathrm{Im}\, d_2\end{aligned} \tag{5.7}$$

and

$$\lim_{E\to\infty}(\operatorname{Im}\gamma_3+\operatorname{Im}d_2)=\frac{\lim\limits_{E\to\infty}\operatorname{Im}d_0+1}{\lim\limits_{E\to\infty}\operatorname{Im}d_1(\operatorname{Im}d_0+1)-1}, \tag{5.8}$$

where, for simplicity of notation, we have supposed the existence of the limit.

Furthermore, multiplying (3.13) by R_0/E and passing to the limit $E\to\infty$, we obtain, after applying (5.4), (5.5), (5.6), (5.7), (5.8), and the positivity of $\operatorname{Re}\gamma_3$,

$$\begin{aligned}\frac{R_0}{E}|\operatorname{Re}d_0| &= \frac{\lim\limits_{E\to\infty}\dfrac{R_0}{E}\operatorname{Re}(d_2+\gamma_3)+\dfrac{R_0}{E}\operatorname{Re}d_1\lim\limits_{E\to\infty}(\operatorname{Im}d_2+\operatorname{Im}\gamma_3)^2}{|\lim\limits_{E\to\infty}\operatorname{Im}d_1\lim\limits_{E\to\infty}(\operatorname{Im}d_2+\operatorname{Im}\gamma_3)-1|^2}\\ &\geq \frac{\dfrac{R_0}{E}\operatorname{Re}d_2+\dfrac{R_0}{E}\operatorname{Re}d_1\lim\limits_{E\to\infty}(\operatorname{Im}d_2+\operatorname{Im}\gamma_3)^2}{|\lim\limits_{E\to\infty}\operatorname{Im}d_1\lim\limits_{E\to\infty}(\operatorname{Im}d_2+\operatorname{Im}\gamma_3)-1|^2};\end{aligned}$$

that is,

$$\begin{aligned}\frac{\beta_0^2}{1-\beta_0} &\geq \frac{\dfrac{\beta_2^2}{\beta_2-1}+\dfrac{\beta_1^2}{\beta_1-1}\lim\limits_{E\to\infty}(\operatorname{Im}d_2+\operatorname{Im}\gamma_3)^2}{|\lim\limits_{E\to\infty}\operatorname{Im}d_1\lim\limits_{E\to\infty}(\operatorname{Im}d_2+\operatorname{Im}\gamma_3)-1|^2}\\ &\geq \frac{\beta_2^2}{\beta_2-1}[1-\lim_{E\to\infty}\operatorname{Im}d_1(\operatorname{Im}d_0+1)]^2\\ &\geq \frac{\beta_2^2}{\beta_2-1}\left[1-\frac{\beta_1(1-\beta_0)}{4\beta_0(\beta_1-1)}\right]^2.\end{aligned}$$

We thus have

$$\frac{(a^2+1/4)^2}{3/4-a^2}\geq\frac{(a^2+25/4)^2}{a^2+21/4}\left[1-\frac{(a^2+9/4)(3/4-a^2)}{4(a^2+1/4)(a^2+5/4)}\right]^2>\frac{(a^2+1/4)^2}{3/4-a^2}.$$

This leads to a contradiction for $\sqrt{3}/4\leq a\leq\alpha_0$, and hence (5.2) is valid. The proof of Lemma 5.2, and hence that of Theorem 5.1, is complete. □

As stated in section 1, assertion (i) of Theorem 1.1 was proven in section 2, while assertion (ii) follows by combining the results of Theorems 3.4, 4.1, and 5.1.

6. Numerical experiments. In this section, we compute numerically the Hopf bifurcation of the zonally periodic problem (1.1) with boundary conditions given by (1.2) and for various values of the aspect ratio a. We discretize a steady-state version of (1.1), as well as the spectral problem (1.5), both using the free-slip boundary conditions (1.2). The finite-difference discretization uses the Arakawa scheme, which conserves energy and enstrophy for 2-D incompressible flows [1]. The spatial resolution is $\Delta x=\Delta y=0.0249$; i.e., there are $N=80$ points in the y direction, $N\Delta y=2$; this resolution is kept the same for all values of a tested ($0.2\leq a\leq 0.86$).

The basic solution $\psi_0=\psi_0(x_i,y_i)$ is found by a pseudoarclength continuation algorithm [9, 23, 39, 42] that solves the discretized steady-state version of (1.1) instead of using (1.3). In order to solve the spectral problem (1.5) we compute the first 10 leading eigenvalues of its discretized version, i.e., those that are closest to the imaginary axis, since it is prohibitive to compute the whole spectrum for our resolution

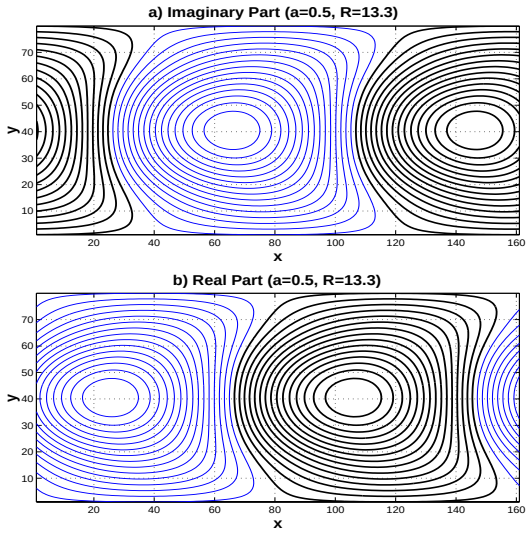


FIG. 6.1. *Unstable oscillatory mode for* $a = 0.5$ *and* $R = 13.3$*:* (a) *Imaginary part* ($t = -T/4$)*; and* (b) *real part* ($t = 0$)*. Light contour lines correspond to negative values and heavy contours to positive ones.*

of $N \times N/a = O(10^4)$ variables. This is achieved using a spectral transformation of the problem combined with a simultaneous iteration technique algorithm (see [39, 40, 42]). An oscillatory eigenfunction $\mathbf{\Phi_r} + i\mathbf{\Phi_i}$ and the corresponding pair of complex conjugate eigenvalues $\kappa_r \pm i\kappa_i$ provide the time-periodic disturbance structure $\mathbf{\Phi}(t)$ with angular frequency κ_i and growth rate κ_r, i.e.,

$$\mathbf{\Phi}(t) = \mathrm{e}^{\kappa_r t} \left[\mathbf{\Phi_r} \cos \kappa_i t - \mathbf{\Phi_i} \sin \kappa_i t \right]. \tag{6.1}$$

Figure 6.1 shows the spatial patterns of the leading eigenvector that loses its stability as R is increased. According to (6.1) this instability propagates westward and it does so for all aspect ratios a we used. The dipolar east-west structure of the destabilizing perturbation is also independent of a. The time-periodic solution is thus characterized by a westward propagation of alternating positive and negative vortices. When a nonlinearly saturated, finite-amplitude version of this oscillatory mode is added to the basic zonal flow Ψ_0, it results in a meandering of the eastward jet (not shown).

The numerically obtained spatial patterns also confirm the theoretical result that this instability is symmetric with respect to the midaxis of the channel. Moreover, numerical results tend to show that there is no other 2-D instability but this one: for fairly large values of $R >> R_0$, there is no other eigenvalue that crosses the imaginary axis.

We now investigate the value of the critical Reynolds number R_0 at the Hopf bifurcation, as a function of the aspect ratio a. We also compute the period $T = 2\pi/\kappa_i$ of the instability at R_0. Both the curves $R_0 = R_0(a)$ and $T = T(a)$ are shown in Figures 6.2(a) and 6.2(b), respectively. The vertical asymptote in both panels strongly suggests that for $a \geq \sqrt{3}/2$ the flow is linearly stable around Ψ_0, in excellent agreement with Theorem 1.1.

Both curves are continuous and monotonic across the entire interval $\sqrt{3}/4 \leq a < \sqrt{3}/2$. This indicates that the difficulties in proving Hopf bifurcation for $\alpha_0 \leq a <$

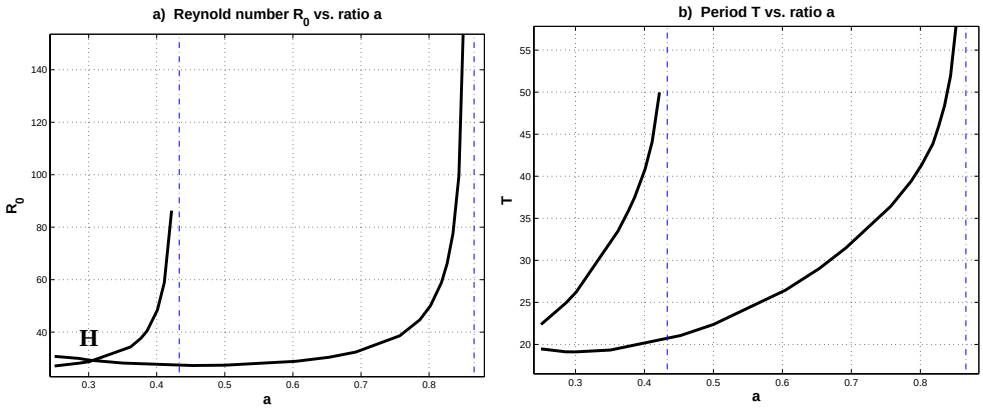


FIG. 6.2. *Dependence of the model's oscillatory instabilities on the aspect ratio a:* (a) *Critical Reynolds number* R_0*; and* (b) *period* T *of the limit cycle at Hopf bifurcation.* T *has been nondimensionalized by* L/U *with* $L = 4000$ km *and* $U = 15$ ms^{-1}*. The vertical dash-dotted straight lines correspond to* $a = \sqrt{3}/4$ *and* $a = \sqrt{3}/2$*, respectively.* H *refers to the simultaneous crossing of the imaginary axis by two distinct eigenmodes, with wavenumbers* $m = 1$ *and* $m = 2$*, respectively.*

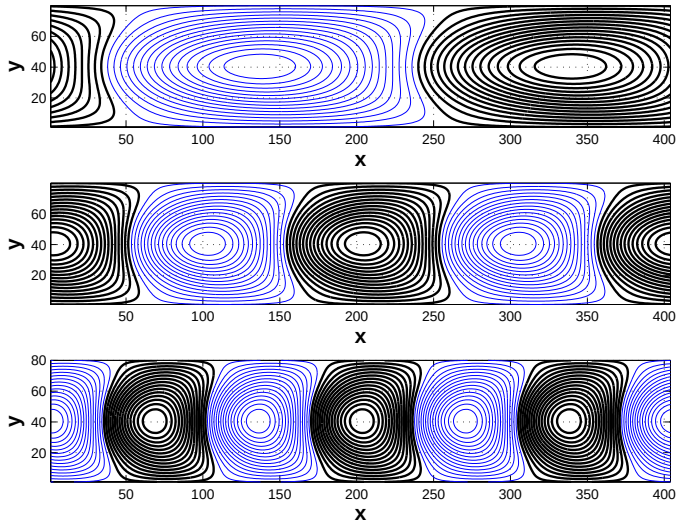


FIG. 6.3. *Real parts of the three unstable modes for* $a = 0.2$ *and* $R = 26.6$.

$\sqrt{3}/2$ are purely technical. In this interval, the flow becomes unstable as well, with a spatial pattern that resembles Figure 6.1 (not shown). The period of the instability increases as the zonal length of the channel decreases and it tends to infinity as a tends to the critical value $\sqrt{3}/2$. Numerical experiments also confirm that only a single instability with wavenumber one is found in the interval $\sqrt{3}/4 \le a < \sqrt{3}/2$.

For $a < \sqrt{3}/4$, interesting phenomena occur that are not captured by our Theorem 1.1. Numerical experiments indicate the existence of additional oscillatory instabilities, characterized by higher spatial harmonics, as shown in Figures 6.2 and 6.3. The competition between these instabilities is expected to generate chaotic dynamics as a is decreased. Codimension-2 Hopf–Hopf bifurcations, with complex global

dynamics nearby, are likely to play a key role in this process. The first of these bifurcations is denoted in Figure 6.2(a) by H and it occurs at $a \simeq 0.31$. At $a = 0.2$ and $R = 26.6$ (not shown in Figure 6.2), three unstable modes, with $m = 1, 2, 3$, coexist. These modes are shown in Figure 6.3 and are all symmetric about the channel's midaxis.

To gain some geophysical insight into the nature of the fundamental instability, we select characteristic dimensional magnitudes for the width $2L$ of the channel and the maximum $\tau_0 U/\pi^3 E$ of the zonal velocity in the midlatitude atmospheric jet. The value of $L = 4000$ km thus yields a meridional extent of the channel of 8000 km. The choice of $U = 15$ ms^{-1} and E between 50 and 100 corresponds to maximum zonal jet velocities between 25 and 50 ms^{-1} that agree well with those observed. Taking the midaxis of the β-channel at 45° N, as is often done in theoretical studies of large-scale atmospheric flow, and $a \simeq 0.26$ yields a channel length of $2L/a$ that corresponds to about 360° in longitude.

With these choices of L, U, and a, the numerical results shown in Figures 6.2(b) and 6.3 yield periods of 20–25 days for the instabilities obtained. Of these, the second one, with zonal wavenumber $m = 2$, has about the correct dimensional wavelength to match the westward-traveling Branstator–Kushnir [3, 22] wave. This wave has been shown to have, indeed, an equivalent-barotropic, i.e., essentially 2-D, structure and a period of about 25 days [11, 12, 13].

7. Symmetry considerations. As discussed in section 1, the symmetry properties of the domain and the forcing, on the one hand, and of the perturbation that gives rise to the bifurcation, on the other, have a decisive effect on whether the bifurcation leads to a branch of stationary or oscillatory solutions. Legras and Ghil [23] and Jin and Ghil [19] pointed out that, in the atmospheric channel problem with bottom topography, back-to-back saddle-node bifurcations resulted when the zonal forcing jet had a flat or unimodal velocity profile; see also Charney and DeVore [6] and Pedlosky [35]. To the contrary, when a higher-order component, which exhibited an inflection point, was present in the forcing jet, an oscillatory instability with so-called intraseasonal periods of 30–60 days could set in by resonance with this higher-order component and lead to a stable limit cycle.

Traveling Rossby waves are the unique type of solutions of (1.1) in the same periodic-channel geometry and in the absence of forcing, topography, and dissipation. The longest known period of this type of so-called free Rossby wave is about 16 days (Madden [29]). The periods of the oscillatory instabilities obtained previously, with bottom topography [6, 19, 23, 35], and here without it, are considerably longer: 25–50 days. Our rigorous result in Theorem 1.1 shows that Hopf bifurcation occurs even for a flat bottom of the channel, in agreement with the numerical results of section 6. The separate effects of forcing and dissipation, on the one hand, and topography, on the other, on the period of oscillatory solutions will be studied further elsewhere.

In the oceanic rectangular-basin problem, on the other hand, given the antisymmetric wind-stress forcing profile of (1.1) here, a steady double-gyre circulation resulted that is antisymmetric with respect to the basin's zonal symmetry axis [4, 18, 42]. This circulation is first destabilized by a pitchfork bifurcation—perfect in QG models [4, 5, 14] and perturbed in shallow-water models [18, 40, 41]—that arises from a purely exponential instability, which is symmetric with respect to the symmetry axis defined here as $y = 1$. Oscillatory instabilities documented numerically in either type of model, QG or shallow-water, also included some that have an asymmetric spatial pattern. The periods of these asymmetric instabilities are also much longer

than those of the oceanic problem's free modes (i.e., the Rossby basin modes arising in the absence of forcing and dissipation); see Simonnet and Dijkstra [39] and Simonnet et al. [41].

The symmetry properties of the Hopf bifurcation here are thus in agreement with those obtained numerically for the atmospheric channel flow. The situation for the oceanic double-gyre problem is more complex and requires further investigation. It might be possible, using some of these symmetry ideas, to adapt the approach and methods used here to this oceanic problem, in spite of the fact that analytic stationary solutions are harder to find for it [18, 44].

Acknowledgments. The first draft of this paper was prepared for presentation at the mathematical congress honoring the memory of Jacques-Louis Lions held in July 2002 at the Collège de France, Paris. It is a pleasure to thank the organizers and the participants of this congress for the impetus to complete the paper in its present form. Constructive comments by Gershon Wolansky and an anonymous reviewer helped improve the presentation and stimulated us to add section 6.

REFERENCES

[1] A. Arakawa, *Computational design for long-term numerical integrations of the equations of atmospheric motion*, J. Comput. Phys., 1 (1966), pp. 119–143.

[2] P. S. Berloff and S. P. Meacham, *On the stability of the wind-driven circulation*, J. Marine Res., 56 (1998), pp. 937–993.

[3] G. W. Branstator, *A striking example of the atmosphere's leading traveling pattern*, J. Atmospheric Sci., 44 (1987), pp. 2310–2323.

[4] P. Cessi and G. R. Ierley, *Symmetry-breaking multiple equilibria in quasi-geostrophic, wind-driven flows*, J. Phys. Oceanogr., 25 (1995), pp. 1196–1202.

[5] K.-I. Chang, M. Ghil, K. Ide, and C.-C. A. Lai, *Transition to aperiodic variability in a wind-driven double-gyre circulation model*, J. Phys. Oceanogr., 31 (2001), pp. 1260–1286.

[6] J. Charney and J. DeVore, *Multiple flow equilibria in the atmosphere and blocking*, J. Atmospheric Sci., 36 (1979), pp. 1205–1216.

[7] Z. M. Chen and W. G. Price, *Remarks on time-dependent periodic Navier-Stokes flow in a two-dimensional torus*, Comm. Math. Phys., 207 (1999), pp. 81–106.

[8] Z. M. Chen and S. Wang, *Steady-state bifurcations of the three-dimensional Kolmogorov problem*, Electron. J. Differential Equations, 58 (2000), pp. 1–32.

[9] H. A. Dijkstra, *Nonlinear Physical Oceanography: A Dynamical Systems Approach to the Large-Scale Ocean Circulation and El Niño*, Kluwer Acad. Publishers, Dordrecht, Norwell, MA, 2000.

[10] M. Ghil and S. Childress, *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics*, Springer-Verlag, New York, 1987.

[11] M. Ghil and K.-C. Mo, *Intraseasonal oscillations in the global atmosphere. Part* I: *Northern Hemisphere and tropics*, J. Atmospheric Sci., 48 (1991), pp. 752–779.

[12] M. Ghil and K.-C. Mo, *Intraseasonal oscillations in the global atmosphere. Part* II: *Southern Hemisphere*, J. Atmospheric Sci., 48 (1991), pp. 780–790.

[13] M. Ghil and A. W. Robertson, *"Waves" vs. "particles" in the atmosphere's phase space: A pathway ro long-range forecasting?*, Proc. Natl. Acad. Sci., 99 (2002), pp. 2493–2500.

[14] M. Ghil, Y. Feliks, and L. Sushama, *Baroclinic and barotropic aspects of the wind-driven ocean circulation*, Phys. D, 167 (2002), pp. 1–35.

[15] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1983.

[16] G. Ierley and V. A. Sheremet, *Multiple solutions and advection-dominated flows in the wind-driven circulation* I: *Slip*, J. Marine Res., 53 (1995), pp. 703–737.

[17] V. I. Iudovich, *Example of the generation of a secondary stationary or periodic flow when there is loss of stability of the laminar flow of a viscous incompressible fluid*, J. Math. Mech., 29 (1965), pp. 587–603.

[18] S. Jiang, F.-F. Jin, and M. Ghil, *Multiple equilibria, periodic, and aperiodic solutions in a wind-driven, double-gyre, shallow-water model*, J. Phys. Oceanogr., 25 (1995), pp. 764–786.

[19] F.-F. Jin and M. Ghil, *Intraseasonal oscillations in the extratropics: Hopf bifurcation and topographic instabilities*, J. Atmospheric Sci., 47 (1990), pp. 3007–3022.

[20] D. D. Joseph and D. Sattinger, *Bifurcating time periodic solutions and their stability*, Arch. Ration. Mech. Anal., 45 (1972), pp. 75–109.
[21] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, UK, 2003.
[22] Y. Kushnir, *Retrograding wintertime low-frequency disturbances over the North Pacific Ocean*, J. Atmospheric Sci., 44 (1987), pp. 2727–2742.
[23] B. Legras and M. Ghil, *Persistent anomalies, blocking and variations in atmospheric predictability*, J. Atmospheric Sci., 42 (1985), pp. 433–471.
[24] J. L. Lions, R. Temam, and S. Wang, *New formulations of the primitive equations of atmosphere and applications*, Nonlinearity, 5 (1992), pp. 237–288.
[25] J. L. Lions, R. Temam, and S. Wang, *On the equations of large-scale ocean*, Nonlinearity, 5 (1992), pp. 1007–1053.
[26] E. N. Lorenz, *Deterministic nonperiodic flow*, J. Atmospheric Sci., 20 (1963), pp. 130–141.
[27] E. N. Lorenz, *The mechanics of vacillation*, J. Atmospheric Sci., 20 (1963), pp. 448–464.
[28] E. N. Lorenz, *The Nature and Theory of the General Circulation of the Atmosphere*, World Meteorological Organization, Geneva, Switzerland, 1967.
[29] R. A. Madden, *Observations of large-scale traveling Rossby waves*, Revs. Geophys., 17 (1979), pp. 1935–1949.
[30] J. E. Marsden and M. McCracken, *The Hopf Bifurcation and Its Applications*, Springer-Verlag, New York, 1976.
[31] S. P. Meacham, *Low-frequency variability in the wind-driven circulation*, J. Phys. Oceanogr., 30 (2000), pp. 269–293.
[32] S. P. Meacham and P. S. Berloff, *Instability of a steady, barotropic, wind-driven circulation*, J. Marine Res., 55 (1997), pp. 885–913.
[33] L. D. Meshalkin and Ya. G. Sinai, *Investigation of the stability of a stationary solution of a system of equations for the plane movement of an incompressible viscous fluid*, J. Math. Mech., 19 (1961), pp. 1700–1705.
[34] B. Nadiga and B. Luce, *Global bifurcation of Shilnikov type in a double gyre ocean model*, J. Phys. Oceanogr., 31 (2001), pp. 2669–2690.
[35] J. Pedlosky, *Resonant topographic waves in barotropic and baroclinic flows*, J. Atmospheric Sci., 38 (1981), pp. 2626–2641.
[36] J. Pedlosky, *Geophysical Fluid Dynamics*, 2nd ed., Springer-Verlag, New York, 1987.
[37] J. Shen, T. Tachim-Medjo, and S. Wang, *On a wind-driven, double-gyre, quasi-geostrophic ocean model: Numerical simulations and structural analysis*, J. Comp. Phys., 155 (1999), pp. 387–409.
[38] V. A. Sheremet, G. Ierley, and V. Kamenkovich, *Eigenanalysis of the two-dimensional wind-driven ocean circulation problem*, J. Marine Res., 55 (1997), pp. 57–92.
[39] E. Simonnet and H. Dijkstra, *Spontaneous generation of low-frequency modes of variability in the wind-driven ocean circulation*, J. Phys. Oceanogr., 32 (2002), pp. 1747–1762.
[40] E. Simonnet, M. Ghil, K. Ide, R. Temam, and S. Wang, *Low-frequency variability in shallow-water models of the wind-driven ocean circulation. Part* I: *Steady-state solutions*, J. Phys. Oceanogr., 33 (2003), pp. 712–728.
[41] E. Simonnet, M. Ghil, K. Ide, R. Temam, and S. Wang, *Low-frequency variability in shallow-water models of the wind-driven ocean circulation. Part* II: *Time-dependent solutions*, J. Phys. Oceanogr., 33 (2003), pp. 729–752.
[42] S. Speich, H. Dijkstra, and M. Ghil, *Successive bifurcations in a shallow-water model, applied to the wind-driven ocean circulation*, Nonlin. Proc. Geophys., 2 (1995), pp. 241–268.
[43] H. Stommel, *Thermohaline convection with two stable regimes of flow*, Tellus, 13 (1961), pp. 224–230.
[44] G. Veronis, *An analysis of wind-driven ocean circulation with a limited number of Fourier components*, J. Atmospheric Sci., 20 (1963), pp. 577–593.
[45] G. Veronis, *Wind-driven ocean circulation. Part* II: *Numerical solution of the nonlinear problem*, Deep-Sea Res., 13 (1966), pp. 30–55.
[46] S. Wang, *Attractors for the* 3*D baroclinic quasi-geostrophic equations of large-scale atmosphere*, J. Math. Anal. Appl., 165 (1992), pp. 266–283.
[47] G. Wolansky, *Existence, uniqueness, and stability of stationary barotropic flow with forcing and dissipation*, Comm. Pure Appl. Math., 41 (1988), pp. 19–46.
[48] G. Wolansky, *The barotropic vorticity equation under forcing and dissipation: Bifurcations of nonsymmetric responses and multiplicity of solutions*, SIAM J. Appl. Math., 49 (1989), pp. 1585–1607.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 369–400

STOCHASTIC MODE REDUCTION FOR THE IMMERSED BOUNDARY METHOD*

PETER R. KRAMER† AND ANDREW J. MAJDA‡

Abstract. We apply the formulation of a stochastic mode reduction method developed in a recent paper of Majda, Timofeyev, and Vanden-Eijnden [*Comm. Pure Appl. Math.*, 54 (2001), pp. 891–974] (MTV) to obtain simplified equations for the dynamics of structures immersed in a thermally fluctuating fluid at low Reynolds (or Kubo) number, as simulated by a recent extension of the immersed boundary (IB) method by Kramer and Peskin [*Proceedings of the Second MIT Conference on Computational Fluid and Solid Mechanics*, Elsevier Science, Oxford, UK, 2003, pp. 1755–1758]. The effective dynamics of the immersed structures are not obvious in the primitive equations, which involve both fluid and structure dynamics, but the procedure of MTV allows the rigorous derivation of a reduced stochastic system for the immersed structures alone. We find, in the limit of small Reynolds (or Kubo) number, that the Lagrangian particle constituents of the immersed structures undergo a drift-diffusive motion with several physically correct features, including the coupling between dynamics of different particles. The MTV procedure is also applied to the spatially discretized form of the IB equations with thermal fluctuations to assist in the design and assessment of numerical algorithms.

Key words. stochastic mode reduction, immersed boundary method, Brownian motion

AMS subject classifications. 60H10, 60H30, 60J60, 60J65, 60J70, 76R50, 82C31, 82C70, 82C80

DOI. 10.1137/S0036139903422139

1. Introduction.

In several applications of modern interest, the governing equations can be written as a complex system of stochastic differential equations, with the variables (modes) evolving over a wide range of characteristic time scales. Sometimes, the variables can be grouped into a "fast" class of modes and a "slow" class of modes, with a wide separation between the time scales of the two classes. In such a situation, one can exploit singular perturbation techniques using the ratio of the fast to slow time scales as a small parameter to reduce the system by averaging the effects of the fast modes on the system. A rigorous procedure for averaging over fast fluctuations in a stochastic system was first provided by Khas'minskii [24, 23] and then later developed into more widely applicable theorems by Kurtz [30] (see also [9]), Ellis and Pinsky [7], and Papanicolaou [39]. (See the textbook [17] for an applied exposition.) Recently, Majda, Timofeyev, and Vanden-Eijnden [34, 35, 36] ([36] hereafter referred to as MTV) have developed these mathematical techniques into a methodological framework for climate modeling, where the governing equations are often essentially quadratically nonlinear and contain both slowly varying climate and "mean flow" modes and more rapidly fluctuating modes. In this work and the companion paper [26], we demonstrate how the MTV framework can be applied productively to a quite different class of applications, namely the simulation of microscale fluid systems with immersed structures and thermal fluctuations, such as microphysiological

*Received by the editors January 30, 2002; accepted for publication (in revised form) June 11, 2003; published electronically December 31, 2003.
http://www.siam.org/journals/siap/64-2/42213.html
†Department of Mathematical Sciences, Rensselaer Polytechnic Institute, 301 Amos Eaton Hall, 110 8th Street, Troy, NY 12180 (kramep@rpi.edu).
‡Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (jonjon@cims.nyu.edu). This author's research was supported in part by ARO grant DAAD19-01-10810, NSF grant DMS9972865, and ONR grant N00014-96-1-0043.

systems, colloid suspensions, and polymer suspensions. In the present paper, we focus on the immersed boundary (IB) method [42] for simulating biological systems, as extended recently by Kramer and Peskin [28] to include thermal fluctuations and thereby extend its applicability to small scales (microns).

The IB method emphasizes the dynamics of the fluid environment in which the biological structures, such as polymers, membranes, and particles, are immersed. The forces generated by the biological structures as they are deformed or interact with external fields are communicated locally as forces on the fluid. The fluid then responds dynamically to these forces in a way represented by the Navier–Stokes equations. The structures are then advected (and strained) by their local fluid velocity. Thermal fluctuations are introduced through forces on the fluid. The immersed structures are not directly thermally forced, but rather undergo thermal fluctuations through advection by the thermally fluctuating fluid [28].

The IB method, then, is described by a coupled system of differential equations with the fluid modes stochastically forced. The details of this dynamical system will be presented in section 2. In the microscopic systems for which the method is designed, viscosity plays a strong role. More precisely, one can define a thermal Reynolds number (product of particle size and thermal velocity divided by kinematic viscosity) which will often be small in typical systems of interest. In such instances, the MTV stochastic mode reduction framework can be applied based on this small parameter to deduce a simplified system governing the immersed particles and structures, with the fluid variables eliminated (see section 3). To unify the discussion with the other simulation methods in [26], it is useful to note that the thermal Reynolds number in the IB method can be identified with a "thermal Kubo number," defined as the ratio of the rate of decorrelation of a particle's (Lagrangian) thermal velocity due to its advection into different fluid regions relative to its rate of decorrelation due to viscous damping.

The main motivation for this work is to obtain a rigorous characterization of the effective dynamics of immersed structures in the IB method at low thermal Reynolds (or Kubo) number. These results can then be used to assess the physical fidelity of the simulation method, to point out possible areas for improvement in the simulation scheme, and to calibrate the numerical parameters in applications. As discussed in [27] and section 2 below, thermal fluctuations are incorporated into the IB simulation equations in a rational manner based on statistical mechanics, but because of a basic approximation in the IB method, it is not a priori clear that the simulated Brownian motion of particles displays physically appropriate behavior. In [27], an approximate semianalytical calculation shows that in fact the IB equations do generate correct physical scaling behavior for various statistical features of multiparticle Brownian motion. Several of these results are corroborated and extended in the present work by the rigorous application of the MTV stochastic mode reduction procedure. In particular, a reduced set of stochastic differential equations describing the dynamics of the immersed structures is derived, with the role of the weak advective nonlinearity operating over long time scales and the nonlinear interaction between the immersed structures rigorously assessed. Such nonlinear effects are neglected in [27].

The immersed structures in the IB method are shown, through the stochastic mode reduction procedure, to obey effective drift-diffusive dynamics at long times, with both the drift coefficient and diffusion coefficient explicitly presented in section 4. Through a study of how the drift of the structures is related to the forces they feel, the structure of the self-diffusion coefficient, and the correlations between the diffusive motion of different particles, we explore to what extent the IB method, in ideal form

without numerical discretization issues, can capture various statistical physical aspects of thermally fluctuating systems. We find that many statistical physical features are properly described by the IB method, but identify a discrepancy between how the correlation in the diffusion of two closely separated particles is simulated by the IB method and its physically proper form for rigid particles. The source of the difference appears to be the lack of a sense of rigidity of particles in the IB method, rather than an artifact arising from the discretization of the fluid.

A cruder physical derivation of the same results is presented in section 5 as an aid to intuitive interpretation. Next, in section 6, we show how the effective dynamics are modified under spatial discretization. We do not consider temporal discretization, since the MTV procedure is not formulated for discrete-time systems, though it would seem that the conclusions concerning the effective dynamics would not be altered substantially [14]. Here we give explicit formulas that indicate how the discrete and continuous long-time dynamics are related. This allows us to calibrate parameters in the discrete numerical simulation as a design principle. A detailed analysis and comparison with numerical simulations will be presented in [27].

We close the introduction by describing some connections of the present work with some other stochastic analytical techniques. Our main technical tool is a singular perturbation analysis of the Kolmogorov backward equation (adjoint to the Fokker–Planck equation), a deterministic second-order parabolic partial differential equation associated with the stochastic dynamics of the full system [9, 30, 39]. The solution of this equation, in the limit of small thermal Reynolds (or Kubo) number, can be shown through a theorem of Kurtz [9, 30] to approach the solution of another second-order parabolic partial differential equation in which the variables corresponding to the fluid modes have been eliminated. The effective stochastic dynamics of the immersed structures, including their drift and diffusion coefficients, can be read off from this limiting equation.

This approach allows the treatment of a system with slow and fast modes, both of which are influenced by each other, as is the case for the IB method in general when the immersed structures do exert force. If, however, the immersed structures are simply force-free particles, then the evolution of the fluid is independent of the particle dynamics. The effective particle dynamics under the IB method can then be analyzed by a variety of other techniques. For example, the dynamics of the particles can be viewed as a random evolution problem [20, 21, 40], where the fluid variables play the role of the auxiliary Markov process parametrizing the advection operator. The long-time limiting effective dynamics, averaging out the influence of the fluid variables, can then be calculated through other fast-averaging formulas [2, 20, 21, 40]. Another approach is to view the motion of the particles as tracers in a turbulent diffusion problem, with the fluid velocity field treated as a prescribed Markovian random, time-dependent field and zero "bare" molecular diffusivity [10, 33]. The long-time behavior of the immersed particles can then be treated through homogenization techniques [1, 11]. If the nonlinear advection term can be safely ignored (due to the low Reynolds number), then the fluid velocity field is nothing more than a superposition of Ornstein–Uhlenbeck processes, for which a simpler analysis is possible [4, 11, 27]. We remark though that the standard turbulent diffusion assumption of point particles must be revised to account for the finite effective size of particles in the IB method.

The derivation of effective dynamics for interacting particle systems on large scales and long times can also be approached through "hydrodynamic limit" techniques [15, 18, 38, 50]. Here, one seeks to pass rigorously from a detailed description of the individual particle dynamics to continuum field equations describing the evolution

of the density and momentum of the collective medium formed by the particles. To our knowledge, such work is primarily focused on Hamiltonian systems with small or zero noise, or on random lattice dynamics with conservation laws. We are not aware of any such applications to systems of particles with strong damping and stochastic driving from the environment, which is the mathematical context of our present study. Moreover, hydrodynamic limit techniques seem most suited for systems in which the constituent particles interact with all other particles according to a universal law governed by their separation distance. It is not clear how to adapt these methods to polymer systems with a variety of bonded interactions. And if hydrodynamic limit techniques could be applied in certain circumstances to the IB method at low thermal Reynolds (or Kubo) number, the results would be complementary. Rather than preserving the Lagrangian framework which accounts for arbitrary N-particle interactions, the hydrodynamic limit would generally be expressed in an Eulerian framework in terms of number densities and correlation functions of the immersed particles.

2. Variables and equations for the IB method. In the IB method [42], the entire system of the fluid with immersed structures is treated as a constant density fluid. We moreover assume that the fluid domain Ω is a cube of side length L with periodic boundary conditions. This is typical for applications of the IB method, because it permits the use of a fast Fourier transform [42]. For the moment, though, we will still consider the space-time domain as continuous.

The evolution of the fluid is given by the incompressible Navier–Stokes equations

$$\rho\left(\frac{\partial \mathbf{u}(\mathbf{x},t)}{\partial t}+\mathbf{u}(\mathbf{x},t)\cdot\nabla\mathbf{u}(\mathbf{x},t)\right)=\mu\nabla^2\mathbf{u}(\mathbf{x},t)-\nabla p(\mathbf{x},t)+\mathbf{f}(\mathbf{x},t),\tag{2.1}$$

$$\nabla\cdot\mathbf{u}(\mathbf{x},t)=0,$$

where $\mathbf{u}(\mathbf{x},t)$ is the fluid velocity, ρ is the density, μ is the dynamic viscosity, p is the pressure, and $\mathbf{f}(\mathbf{x},t)$ is a force density. We decompose the force density into a sum,

$$\mathbf{f}(\mathbf{x},t)=\mathbf{f}_{\mathrm{IS}}(\mathbf{x},t)+\mathbf{f}_T(\mathbf{x},t),$$

with $\mathbf{f}_{\mathrm{IS}}(\mathbf{x},t)$ representing the contribution arising from the immersed structures and $\mathbf{f}_T(\mathbf{x},t)$ representing thermally fluctuating forces from microscopic processes. We next describe how these contributions to the force density are expressed concretely.

The collection of immersed structures will be modelled as a finite collection of Lagrangian particles, located at positions $\mathbf{X}=\{\mathbf{X}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$, where $\boldsymbol{\alpha}$ is a Lagrangian labelling index taking values from some finite set $\mathcal{A}$. The various stresses exerted by the immersed structures in response to deformations will be modelled in general through gradients of some interparticle potential $\Phi(\mathbf{X})$. Note that stresses such as those arising from bending resistance can be modelled by n-body interactions with $n>2$, and still fall within our scope. We assume that there are no external forces, so that the total momentum of the system is conserved ($\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\nabla_{\boldsymbol{\alpha}}\Phi(\mathbf{X})=0$, where $\nabla_{\boldsymbol{\alpha}}$ denotes a gradient with respect to the position of the Lagrangian particle $\mathbf{X}_{\boldsymbol{\alpha}}$). By choosing an appropriate inertial frame, we can then assume that the total system momentum is always zero. (We briefly discuss in section 4.3 how the results would be modified if external forces were allowed to be present.)

With the potential prescribed, we can then describe the force density exerted on the fluid by applying the force exerted on each Lagrangian particle at its current position $\mathbf{X}_{\boldsymbol{\alpha}}(t)$, spread out via a smoothed delta function $\delta_a(\mathbf{x})$ with length scale a.

The choice of smoothing is dictated by numerical considerations, such as compact support and the minimization of oscillations in the fluid-particle interaction as the particles move with respect to the fluid grid [42]. We therefore write

$$\mathbf{f}_{\mathrm{IS}}(\mathbf{x},t) = -\sum_{\boldsymbol{\alpha}\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}}\Phi(\mathbf{X}(t))\delta_a(\mathbf{x}-\mathbf{X}_{\boldsymbol{\alpha}}(t)). \tag{2.2}$$

The parameter a acts as an effective particle size.

The thermal force density is obtained using a fluctuation-dissipation theorem from statistical hydrodynamics [27, 31] and is most clearly expressed in terms of its Fourier series expansion:

$$\begin{aligned} \mathbf{f}_T(\mathbf{x},t) &= \sum_{\mathbf{k}\in S} \mathbf{f}_{T,\mathbf{k}}(t) e^{2\pi i \mathbf{k}\cdot\mathbf{x}/L}, \\ \mathbf{f}_{T,\mathbf{k}}(t) &= \sqrt{\frac{8\pi^2 k^2 \mu k_B T}{L^5}}\, \frac{d\tilde{\mathbf{W}}_{\mathbf{k}}(t)}{dt}, \end{aligned} \tag{2.3}$$

where k_B is Boltzmann's constant, T is the absolute temperature, and $\{\tilde{\mathbf{W}}_{\mathbf{k}}(t)\}_{\mathbf{k}\in S}$ are a collection of standard complex Brownian processes, which are mutually independent except for the complex conjugacy relation

$$\tilde{\mathbf{W}}_{-\mathbf{k}}(t) = \overline{\tilde{\mathbf{W}}_{\mathbf{k}}(t)}, \tag{2.4}$$

which arises from the need to keep $\mathbf{f}_T(\mathbf{x},t)$ real-valued. By a "standard complex Brownian motion," we refer to a mean-zero Gaussian process with stationary increments satisfying

$$\langle d\tilde{\mathbf{W}}(t) \otimes d\tilde{\mathbf{W}}(t')\rangle = 0, \qquad \langle d\tilde{\mathbf{W}}(t) \otimes \overline{d\tilde{\mathbf{W}}(t')}\rangle = \mathcal{I}\delta(t-t')\,dt\,dt', \tag{2.5}$$

where $\mathcal{I}$ is the identity matrix. Please note that the definition of complex Brownian motion processes used in MTV [36] differs by a factor of two in normalization of the variance. In (2.3) and elsewhere in the paper, for simplicity in exposition, we occasionally use the formal notation $d\tilde{\mathbf{W}}(t)/dt$ for the white noise derivative of Brownian motion. Of course, the equations can be given a rigorous interpretation through use of the more proper Itô stochastic differential notation [37]. The set of wavenumbers is just the lattice of integers in three dimensions, with the zero mode excluded since it will always vanish: $S = \mathbb{Z}^3 \setminus \{\mathbf{0}\}$.

We note that the continuum formulation in which all these modes are retained requires care in a serious interpretation, because the velocity field induced by the thermal forcing in (2.3) exhibits an ultraviolet catastrophe due to singular small scale structure. We will not concern ourselves with such subtleties here, because in any numerical implementation the number of modes simulated is finite. Therefore, we will proceed just as if S were a finite collection of modes corresponding to some symmetric Galerkin truncation. However, in our actual discretized implementation, the IB equations are not simply crudely cut off in this way, and we show in section 6 how the results and arguments should be modified to apply to the actual numerical discretization.

With the Navier–Stokes equations (2.1) and the equations (2.2) and (2.3) for the force density, we have defined how the fluid evolves. The particle positions are updated by simple advection by the fluid at a locally interpolated fluid velocity:

$$\frac{d\mathbf{X}_{\boldsymbol{\alpha}}(t)}{dt} = \mathbf{u}_a(\mathbf{X}_{\boldsymbol{\alpha}}(t),t).$$

The same smoothed delta function that was used to spread force is used to interpolate velocity:

$$\mathbf{u}_a(\mathbf{x},t) = \int_\Omega \mathbf{u}(\mathbf{x}',t)\delta_a(\mathbf{x}-\mathbf{x}')\,d\mathbf{x}'; \tag{2.6}$$

this choice (along with the enforced reflection symmetry $\delta_a(\mathbf{x}) = \delta_a(-\mathbf{x})$) conserves energy in the particle-fluid interactions [42]. Note that, particularly in this integration, the delta function δ_a should be viewed as periodic (with its spikes centered at every point of the form (n_1L, n_2L, n_3L) with n_1, n_2, and n_3 integers). Equivalently, the convolution in (2.6) should be viewed as convolution on a torus [25].

2.1. Summary of IB equations in dimensional form. Summarizing, we have the following system of equations for the IB method:

$$\begin{aligned}
\rho\left(\frac{\partial \mathbf{u}(\mathbf{x},t)}{\partial t} + \mathbf{u}(\mathbf{x},t)\cdot\nabla\mathbf{u}(\mathbf{x},t)\right) &= \mu\nabla^2\mathbf{u}(\mathbf{x},t) - \nabla p(\mathbf{x},t) + \mathbf{f}_T(\mathbf{x},t) \\
&\quad - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}}\Phi(\mathbf{X}(t))\delta_a(\mathbf{x}-\mathbf{X}_{\boldsymbol{\alpha}}(t)), \\
\nabla\cdot\mathbf{u}(\mathbf{x},t) &= 0, \\
\frac{d\mathbf{X}_{\boldsymbol{\alpha}}(t)}{dt} &= \mathbf{u}_a(\mathbf{X}_{\boldsymbol{\alpha}}(t),t), \\
\mathbf{u}_a(\mathbf{x},t) &= \int_\Omega \mathbf{u}(\mathbf{x}',t)\delta_a(\mathbf{x}-\mathbf{x}')\,d\mathbf{x}',
\end{aligned} \tag{2.7a}$$

with the thermal forcing given by the following random process:

$$\mathbf{f}_T(\mathbf{x},t) = \sum_{\mathbf{k}\in S} \sqrt{\frac{8\pi^2k^2\mu k_BT}{L^5}}\, e^{2\pi i\mathbf{k}\cdot\mathbf{x}/L}\frac{d\tilde{\mathbf{W}}_{\mathbf{k}}(t)}{dt}. \tag{2.7b}$$

These equations are supplemented with initial conditions

$$\mathbf{X}_{\boldsymbol{\alpha}}(t=0) = \mathbf{X}_{0,\boldsymbol{\alpha}}, \qquad \mathbf{u}(\mathbf{x},t=0) = \mathbf{u}^0(\mathbf{x}).$$

2.2. Nondimensionalization. To prepare for the asymptotic results and calculations in subsequent sections, we nondimensionalize the IB equations. This will make manifest the role of the relevant Kubo number as a small parameter. We choose to nondimensionalize with respect to a length and time scale associated with the thermal forcing.

2.2.1. Parameters of externally specified functions. To facilitate the parametrization of the contributions from the initial data and the prescribed force law of the immersed structures, we express each in terms of dimensionless functions.

The initial velocity field will be described by a magnitude U_0 and length scale ℓ_v, and we write

$$\mathbf{u}^0(\mathbf{x}) = U_0\tilde{\mathbf{u}}^0(\mathbf{x}/\ell_v),$$

where $\tilde{\mathbf{u}}^0$ is a dimensionless function.

We identify ψ as a *force density* induced by the immersed structures, and ℓ_f as a length scale on which the immersed structure forces vary. More explicitly, we assume

$$\Phi(\mathbf{X}) = \psi\ell_f a^3\tilde{\Phi}(\mathbf{X}/\ell_f)$$

for some nondimensional function $\tilde{\Phi}$ with order unity amplitude and order unity gradients of its argument. Then an elementary constituent particle (with volume a^3) will experience a force of magnitude ψa^3.

Envisioning that the immersed structures will be modelled as a collection of elementary constituent particles with effective size a and spacing on the order of a, we will nondimensionalize $\mathbf{X}_{0,\boldsymbol{\alpha}}$ with respect to a:

$$\mathbf{X}_{0,\boldsymbol{\alpha}} = a\tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}}.$$

Of course, the functions described above could have various amplitudes and length scales, depending on the model, but such complications do not bear on the central point of this work.

2.2.2. Reference units. We choose the following units to normalize the equations with respect to the thermal dynamics of the particles:

(i) length scale $\ell_T = a$,
(ii) time scale $\tau_T = \sqrt{\rho a^5/k_B T}$,
(iii) mass $m_T = \rho a^3$.

For example, the mass reference unit is just the mass associated with an elementary particle in the IB formulation (a single delta function), and length and time units are chosen so that the reference velocity

$$V_T \equiv \frac{\ell_T}{\tau_T} = \sqrt{\frac{k_B T}{\rho a^3}}$$

has the order of magnitude of the thermalized velocity of an elementary particle (since the IB system with thermal forcing respects the equipartition law [27]). Note that, after nondimensionalization, the fluid density, the width of the delta function (as well as the grid spacing), and the root-mean-square of the fluid velocity averaged over an elementary particle region are all order unity.

2.2.3. Nondimensional groups. With the above nondimensionalization, the IB dynamics are governed by the following nondimensional groups:

(i) Kubo number based on thermal forcing

$$\mathrm{Ku_T} = \frac{\ell_T^2}{\nu \tau_T} = \sqrt{\frac{k_B T}{\rho \nu^2 a}}, \tag{2.8}$$

where $\nu = \mu/\rho$ is the kinematic viscosity of the fluid;

(ii) nondimensionalized measures of the effects of structural forces and initial velocity

$$\phi = \frac{\psi a^4}{k_B T}, \qquad \Upsilon = \sqrt{\frac{U_0^2 \rho a^3}{k_B T}} = \frac{U_0}{V_T};$$

(iii) length scale ratios

$$\tilde{K} = \frac{L}{a}, \qquad \tilde{\ell}_f = \frac{\ell_f}{a}, \qquad \tilde{\ell}_v = \frac{\ell_v}{a}.$$

We pause to clarify why the nondimensional group $\mathrm{Ku_T}$ is identified as a Kubo number and can also be viewed here as a thermal particle Reynolds number. A

Kubo number is generally defined as the ratio of the time scales of decorrelation of a particle's motion due to intrinsic (Eulerian) processes to that due to advection across variable spatial structure [29, 52]. In the present context, $\tau_T = \ell_T/V_T$ is the time scale on which one of the Lagrangian particles would change its velocity due to its motion (at typical speed V_T) because of its sampling of a new environment after it moves a distance equal to its size $\ell_T = a$. The time scale $\ell_T^2/\nu = a^2/\nu$, on the other hand, describes the intrinsic (Eulerian) decorrelation rate of the fluid velocity averaged over the region occupied by a particle of size a.

Alternatively, the group (2.8) can be viewed as a thermal particle Reynolds number $\mathrm{Re_T}$, since it is the product of a characteristic particle length scale ($\ell_T = a$) and the thermal velocity of the fluid $V_T = \ell_T/\tau_T$ divided by the kinematic viscosity [51]. (Note that the notion of the thermal velocity of a fluid makes sense only if it is discretized or smoothed over some finite region (here taken to be of width a).) The fact that the thermal particle Reynolds number $\mathrm{Re_T} = \mathrm{Ku_T}$ decreases with the length scale a may be surprising; the reason is that the root-mean-square velocity sampled over a region a scales as $a^{-3/2}$ because of its short-range correlations. Therefore, while $\mathrm{Re_T}$ is small, it is not as minuscule as one might expect based on macroscopic intuition. For typical parameter values $k_B = 1.4 \times 10^{-16}$erg/K, $T = 300$ K, $\nu = 0.01$ cm/s, we have

$$\mathrm{Re_T} \approx \sqrt{\frac{5 \times 10^{-4}\mu\mathrm{m}}{a}}.$$

The elementary constituent size in a numerical simulation will be of the order $a \sim 0.01 - 0.1\mu$m, and thus we see that the Reynolds number based on the thermal forcing can be expected to be on the order of 10^{-2}–10^{-1}. We emphasize the thermal Kubo number interpretation because it allows the most parallel handling of the three simulation methods discussed in the companion paper [26].

2.2.4. Nondimensionalized IB equations. We now nondimensionalize the independent and dependent variables with respect to the reference units described in section 2.2.2, but denote their nondimensional versions by the same symbols. We retain special symbols for the nondimensionalized externally prescribed functions, as defined in section 2.2.3,

$$\begin{aligned}
\left(\frac{\partial \mathbf{u}(\mathbf{x},t)}{\partial t} + \mathbf{u}(\mathbf{x},t)\cdot\nabla\mathbf{u}(\mathbf{x},t)\right) &= \mathrm{Ku_T^{-1}}\nabla^2\mathbf{u}(\mathbf{x},t) - \nabla p(\mathbf{x},t) + \mathbf{f}_T(\mathbf{x},t) \\
&\quad - \phi\sum_{\boldsymbol{\alpha}\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}}\tilde{\Phi}\left(\frac{\mathbf{X}(t)}{\tilde{\ell}_f}\right)\delta_1(\mathbf{x}-\mathbf{X}_{\boldsymbol{\alpha}}(t)), \\
\nabla\cdot\mathbf{u}(\mathbf{x},t) &= 0, \\
\frac{d\mathbf{X}_{\boldsymbol{\alpha}}(t)}{dt} &= \mathbf{u}_1(\mathbf{X}_{\boldsymbol{\alpha}}(t),t), \\
\mathbf{u}_1(\mathbf{x},t) &= \int_\Omega \mathbf{u}(\mathbf{x}',t)\delta_1(\mathbf{x}-\mathbf{x}')\,d\mathbf{x}',
\end{aligned} \tag{2.9a}$$

with the thermal forcing given by the following random process

$$\mathbf{f}_T(\mathbf{x},t) = \mathrm{Ku_T^{-1/2}}\tilde{K}^{-3/2}\sum_{\mathbf{k}\in S}\sqrt{2\left(4\pi^2\left(\frac{k}{\tilde{K}}\right)^2\right)}e^{2\pi i\mathbf{k}\cdot\mathbf{x}/\tilde{K}}\frac{d\tilde{\mathbf{W}}_{\mathbf{k}}(t)}{dt} \tag{2.9b}$$

and initial data

$$\mathbf{X}_{\boldsymbol{\alpha}}(t=0) = \tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}}, \qquad \mathbf{u}(\mathbf{x}, t=0) = \Upsilon \tilde{\mathbf{u}}^0 \left(\frac{\mathbf{x}}{\tilde{\ell}_v} \right). \tag{2.9c}$$

(We note that $\nabla_{\boldsymbol{\alpha}} \tilde{\Phi}(\mathbf{X}/\tilde{\ell}_f)$ is to be interpreted as the gradient with respect to the $\boldsymbol{\alpha}$ coordinate of $\tilde{\Phi}$, evaluated at $\mathbf{X}/\tilde{\ell}_f$.)

3. Stochastic mode reduction procedure. For the microscopic applications for which the IB method with thermal fluctuations has been designed, the systems are at low thermal Kubo number, so the fluid motion is strongly damped by viscosity. The positions of the immersed structures, however, have no such damping terms in their equations of motion, and should therefore evolve on a slower time scale than the fluid variables. To make these notions quantitative, we consider the Kubo number based on the thermal forcing, $\mathrm{Ku_T}$, as a small parameter. Then we see from (2.9) that, at least formally, the (nondimensionalized) fluid variables evolve on the fast time scale $\mathrm{Ku_T^{-1}}$, while the immersed structure positions evolve on a slower time scale of $O(1)$ or longer. The IB equations are therefore well suited for the stochastic mode reduction framework developed in MTV. That is, we can systematically eliminate the fluid variables from consideration for small $\mathrm{Ku_T}$ and obtain a closed stochastic equation for the immersed structure positions $\{\mathbf{X}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ alone. We now sketch this stochastic mode reduction procedure for the IB system. The detailed calculations can be found in the appendix. The result will be presented at the beginning of section 4.

3.1. Fourier expansion of IB equations. We prepare by expanding the velocity field (which is assumed periodic) in a Fourier series:

$$\mathbf{u}(\mathbf{x}, t) = \sum_{\mathbf{k}\in S} e^{2\pi i \mathbf{k}\cdot\mathbf{x}/\tilde{K}} \hat{\mathbf{u}}_{\mathbf{k}}(t). \tag{3.1}$$

The Navier–Stokes equations now become a coupled collection of stochastic ordinary differential equations. The nondimensionalized IB system (2.9), expressed in terms of the Fourier coefficients of the velocity field, reads

$$\begin{aligned} d\hat{\mathbf{u}}_{\mathbf{k}}(t) &= -\mathcal{B}_{\mathbf{k}}(\mathbf{U}(t), \mathbf{U}(t))\, dt - 4\pi^2 (k/\tilde{K})^2 \mathrm{Ku_T^{-1}} \hat{\mathbf{u}}_{\mathbf{k}}(t)\, dt \\ &\quad - \phi \mathcal{P}_{\mathbf{k}} \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}} \tilde{\Phi}(\mathbf{X}(t)/\tilde{\ell}_f) \hat{\delta}_{1,\mathbf{k}} e^{-2\pi i \mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}(t)/\tilde{K}}\, dt \\ &\quad + \mathrm{Ku_T^{-1/2}} \tilde{K}^{-3/2} \sqrt{2(4\pi^2(k/\tilde{K})^2)}\, \mathcal{P}_{\mathbf{k}} d\tilde{\mathbf{W}}_{\mathbf{k}}(t), \\ d\mathbf{X}_{\boldsymbol{\alpha}}(t) &= \tilde{K}^3 \sum_{\mathbf{k}\in S} e^{2\pi i \mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}(t)/\tilde{K}} \hat{\mathbf{u}}_{\mathbf{k}}(t) \hat{\delta}_{1,\mathbf{k}}\, dt, \end{aligned} \tag{3.2a}$$

with initial data

$$\mathbf{X}_{\boldsymbol{\alpha}}(t=0) = \tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}}, \qquad \hat{\mathbf{u}}_{\mathbf{k}}(\mathbf{x}, t=0) = \Upsilon \hat{\tilde{\mathbf{u}}}^0_{\mathbf{k},\tilde{\ell}_v}. \tag{3.2b}$$

The Fourier expansion coefficients of the nonlinear advection term are

$$\mathcal{B}_{\mathbf{k}}(\mathbf{U}, \mathbf{U}) = 2\pi i \tilde{K}^{-1} \mathcal{P}_{\mathbf{k}} \sum_{\mathbf{k}'\in S} (\mathbf{u}_{\mathbf{k}'} \cdot \mathbf{k}) \mathbf{u}_{\mathbf{k}-\mathbf{k}'}. \tag{3.2c}$$

The effect of the pressure has been replaced in the standard way [32] through the introduction of a projection tensor which enforces incompressibility of each Fourier velocity mode:

$$\mathcal{P}_{\mathbf{k}} = \mathcal{I} - \frac{\mathbf{k} \otimes \mathbf{k}}{k^2}.$$

In our notation, variables with $\mathbf{k}$ subscripts indicate Fourier coefficients, as defined through the nondimensionalized Fourier transform (3.1). Note that $\{\hat{\tilde{\mathbf{u}}}^0_{\mathbf{k},\tilde{\ell}_v}\}_{\mathbf{k}\in S}$ are the Fourier coefficients of $\tilde{\mathbf{u}}^0(\mathbf{x}/\tilde{\ell}_v)$ and not of $\tilde{\mathbf{u}}^0(\mathbf{x})$. Finally, $\mathbf{U}$ is a shorthand for the collection of all Fourier velocity modes $\{\hat{\mathbf{u}}_{\mathbf{k}}\}_{\mathbf{k}\in S}$, though we recall that the zero mode $\mathbf{u}_0 \equiv \mathbf{0}$ because of our assumption that the global system momentum is conserved. To avoid excessive proliferation of $\mathcal{P}_{\mathbf{k}}$ symbols, we will consider the variables $\hat{\mathbf{u}}_{\mathbf{k}}$ to always be constrained to satisfy the incompressibility condition $\mathbf{k} \cdot \hat{\mathbf{u}}_{\mathbf{k}} = 0$. We will not concern ourselves unduly with the infinite number of modes in S, an ideal continuum fluid system. Indeed, all results converge, and our main interest is really in a discretized finite version of these equations (section 6).

For the sake of integration and differentiation, it will be convenient to follow a convention in complex analysis of treating $\hat{\mathbf{u}}_{\mathbf{k}}$ and $\hat{\mathbf{u}}^*_{\mathbf{k}}$ as independent variables, each with two degrees of freedom (due to the transversality condition $\mathbf{k} \cdot \hat{\mathbf{u}}_{\mathbf{k}} = \mathbf{k} \cdot \hat{\mathbf{u}}^*_{\mathbf{k}} = \mathbf{0}$). Note, however, that because $\mathbf{u}(\mathbf{x}, t)$ is a real-valued vector field, its Fourier coefficients must satisfy the complex conjugacy relations

$$\hat{\mathbf{u}}_{-\mathbf{k}}(t) = \hat{\mathbf{u}}^*_{\mathbf{k}}(t).$$

Therefore, we can consider $\{\hat{\mathbf{u}}_{\mathbf{k}}\}_{\mathbf{k}\in S}$ as a complete set of independent fluid coordinates.

3.2. Kolmogorov backward equation formulation. The calculation is performed on the Kolmogorov backward equation associated with the nondimensionalized IB equations (3.2):

$$\begin{aligned}
-\frac{\partial \rho(s, \mathbf{X}, \mathbf{U}|t)}{\partial s} = \sum_{\mathbf{k}\in S} \Bigg[& -\mathcal{B}_{\mathbf{k}}(\mathbf{U}, \mathbf{U}) - \phi \mathcal{P}_{\mathbf{k}} \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}} \tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right) \hat{\delta}_{1,\mathbf{k}} e^{-2\pi i \mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}} \\
& - \mathrm{Ku}_{\mathrm{T}}^{-1} \left(4\pi^2 \left(\frac{k}{\tilde{K}}\right)^2\right) \hat{\mathbf{u}}_{\mathbf{k}} \Bigg] \cdot \frac{\partial \rho}{\partial \hat{\mathbf{u}}_{\mathbf{k}}} \\
& + \mathrm{Ku}_{\mathrm{T}}^{-1} \sum_{\mathbf{k}\in S} \tilde{K}^{-3} \left(4\pi^2 \left(\frac{k}{\tilde{K}}\right)^2\right) \frac{\partial}{\partial \hat{\mathbf{u}}^*_{\mathbf{k}}} \cdot \frac{\partial \rho}{\partial \hat{\mathbf{u}}_{\mathbf{k}}} \\
& + \tilde{K}^3 \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \sum_{\mathbf{k}\in S} e^{2\pi i \mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}} \hat{\delta}_{1,\mathbf{k}} \hat{\mathbf{u}}_{\mathbf{k}} \cdot \frac{\partial \rho}{\partial \mathbf{X}_{\boldsymbol{\alpha}}},
\end{aligned} \tag{3.3}$$

$$\rho(s = t, \mathbf{X}, \mathbf{U}|t) = f(\mathbf{X}, \mathbf{U}).$$

The solution $\rho(s, \mathbf{X}, \mathbf{U}|t)$ to this Kolmogorov backward equation has the mathematical interpretation as the following conditional expectation:

$$\rho(s, \mathbf{X}, \mathbf{U}|t) = \mathbb{E}\left[f(\mathbf{X}(t), \mathbf{U}(t))|\mathbf{X}(s) = \mathbf{X}, \mathbf{U}(s) = \mathbf{U}\right], \tag{3.4}$$

where $\mathbf{X}(t)$ and $\mathbf{U}(t)$ evolve according to the IB equations (3.2) forward in time, conditioned on their starting at time $s < t$ from values $\mathbf{X}(s) = \mathbf{X}$ and $\mathbf{U}(s) = \mathbf{U}$,

and $\mathbb{E}$ denotes an average over the stochastic noise terms in the evolution equations. Because of the slight complication of using complex coordinates, we provide a formal derivation of the Kolmogorov backward equation in the appendix.

Some notational remarks are in order. We hope the reader will not be confused by our previous use of ρ as a symbol for fluid density; with the nondimensionalization in section 2.2, the fluid density has been removed from the problem and we have reallocated its symbol. To avoid straining the reader's eyes with numerous subscripts upon subscripts, we have used partial derivative notation to represent gradients with respect to vectorial modes when they apply to functions of both $\mathbf{U}$ and $\mathbf{X}$:

$$\frac{\partial}{\partial \hat{\mathbf{u}}_{\mathbf{k}}} = \nabla_{\hat{\mathbf{u}}_{\mathbf{k}}}, \qquad \frac{\partial}{\partial \mathbf{X}_{\boldsymbol{\alpha}}} = \nabla_{\mathbf{X}_{\boldsymbol{\alpha}}}.$$

We still use the abbreviation $\nabla_{\boldsymbol{\alpha}}$ for $\nabla_{\mathbf{X}_{\boldsymbol{\alpha}}}$ when applied to a function only of $\mathbf{X}$. Following the usual practice in complex analysis, $\hat{\mathbf{u}}_{\mathbf{k}}$ and $\hat{\mathbf{u}}^*_{\mathbf{k}}$ are considered to be independent variables in differentiation, so that, for example, $\partial g(\hat{\mathbf{u}}^*_{\mathbf{k}})/\partial \hat{\mathbf{u}}_{\mathbf{k}} = 0$. Finally, we have suppressed the time arguments of most terms in the Kolmogorov backward equation; they are all understood to be evaluated at the running time argument s.

The Kolmogorov backward equation is not being used here to actually solve for the evolution of some expectation of some function of the system variables, but merely as a tool to cast the stochastic dynamics in terms of a deterministic PDE. Perhaps the Kolmogorov forward (or Fokker–Planck) equation, which describes the evolution of the probability density of the system variables, may be a more intuitive formulation, but rigorous theorems are generally easier to prove for the backward equation (see MTV section 4.4 and references therein).

3.3. Identification of small parameter and rescaling of time. We identify $\varepsilon = \mathrm{Ku_T}$ as the small parameter and rescale to a longer time $t \to t/\varepsilon$. This temporal rescaling is necessary to see nontrivial dynamics in the $\varepsilon \downarrow 0$ limit, as we shall discuss in section 5. The Kolmogorov backward equation for the rescaled function

$$\rho^{\varepsilon}(s, \mathbf{X}, \mathbf{U}|t) = \rho(s/\varepsilon, \mathbf{X}, \mathbf{U}|t/\varepsilon)$$

may then be written as

$$\begin{aligned} -\frac{\partial \rho^{\varepsilon}(s, \mathbf{X}, \mathbf{U}|t)}{\partial s} &= \varepsilon^{-2}\mathcal{L}_1\rho^{\varepsilon} + \varepsilon^{-1}\mathcal{L}_2\rho^{\varepsilon}, \\ \rho^{\varepsilon}(s = t, \mathbf{X}, \mathbf{U}|t) &= f(\mathbf{X}, \mathbf{U}), \end{aligned} \tag{3.5a}$$

with differential operators

$$\begin{aligned} \mathcal{L}_1 &= \sum_{\mathbf{k}\in S}\left[-\left(4\pi^2\left(\frac{k}{\tilde{K}}\right)^2\right)\hat{\mathbf{u}}_{\mathbf{k}}\right]\cdot\frac{\partial}{\partial \hat{\mathbf{u}}_{\mathbf{k}}} + \left(4\pi^2\left(\frac{k}{\tilde{K}}\right)^2\right)\tilde{K}^{-3}\frac{\partial}{\partial \hat{\mathbf{u}}^*_{\mathbf{k}}}\cdot\frac{\partial}{\partial \hat{\mathbf{u}}_{\mathbf{k}}}, \\ \mathcal{L}_2 &= \sum_{\mathbf{k}\in S}\left[-\mathcal{B}_{\mathbf{k}}(\mathbf{U},\mathbf{U}) - \phi\mathcal{P}_{\mathbf{k}}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\nabla_{\boldsymbol{\alpha}}\tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right)\hat{\delta}_{1,\mathbf{k}}e^{-2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}}\right]\cdot\frac{\partial}{\partial \hat{\mathbf{u}}_{\mathbf{k}}} \\ &\quad + \tilde{K}^3\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\sum_{\mathbf{k}\in S}e^{2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}}\hat{\delta}_{1,\mathbf{k}}\hat{\mathbf{u}}_{\mathbf{k}}\cdot\frac{\partial}{\partial \mathbf{X}_{\boldsymbol{\alpha}}}. \end{aligned} \tag{3.5b}$$

3.4. Computation of limiting equation. We now apply singular perturbation techniques to this problem to find the equation satisfied by $\rho_0 \equiv \lim_{\varepsilon \downarrow 0} \rho^\varepsilon$. This calculation is presented in the appendix. We find that

$$\begin{aligned} -\frac{\partial \rho_0}{\partial s} &= \bar{\mathcal{L}} \rho_0, \\ \rho_0(s = t, \mathbf{X}|t) &= f(\mathbf{X}), \end{aligned} \tag{3.6a}$$

where the limiting differential operator is given by

$$\begin{aligned} \bar{\mathcal{L}} g(\mathbf{X}) = &-\phi \tilde{K}^3 \sum_{\mathbf{k} \in S} \frac{\hat{\delta}_{1,\mathbf{k}}^2}{4\pi^2 (k/\tilde{K})^2} \mathcal{P}_{\mathbf{k}} \sum_{\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathcal{A}} \nabla_{\boldsymbol{\alpha}'} \tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right) e^{2\pi i \mathbf{k} \cdot (\mathbf{X}_{\boldsymbol{\alpha}} - \mathbf{X}_{\boldsymbol{\alpha}'})/\tilde{K}} \cdot \frac{\partial g(\mathbf{X})}{\partial \mathbf{X}_{\boldsymbol{\alpha}}} \\ &+ \tilde{K}^3 \sum_{\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathcal{A}} \sum_{\mathbf{k} \in S} \frac{|\hat{\delta}_{1,\mathbf{k}}|^2}{4\pi^2 (k/\tilde{K})^2} e^{2\pi i \mathbf{k} \cdot (\mathbf{X}_{\boldsymbol{\alpha}} - \mathbf{X}_{\boldsymbol{\alpha}'})} \frac{\partial}{\partial \mathbf{X}_{\boldsymbol{\alpha}}} \cdot \mathcal{P}_{\mathbf{k}} \cdot \frac{\partial g(\mathbf{X})}{\partial \mathbf{X}_{\boldsymbol{\alpha}'}}. \end{aligned} \tag{3.6b}$$

3.4.1. Passage to reduced stochastic representation. We realize now that the PDE for ρ_0 is again a Kolmogorov backward equation for a Markov process, which we present in the next proposition. This relation can be checked through a stochastic Taylor expansion [22], as in the appendix. We use the fact that $\hat{\delta}_{1,\mathbf{k}} = \hat{\delta}_{1,\mathbf{k}}^*$ due to the assumed even symmetry $\delta_1(\mathbf{x}) = \delta_1(-\mathbf{x})$.

4. Effective dynamics for immersed structures at low thermal Kubo number. The outcome of the stochastic mode reduction procedure is summarized in the following proposition.

PROPOSITION 4.1 (IB dynamics at small Kubo number). *Suppose the IB system* (2.9) *conserves total momentum* ($\sum_{\boldsymbol{\alpha} \in \mathcal{A}} \nabla_{\boldsymbol{\alpha}} \Phi(\mathbf{X}) = 0$). *Then, in the limit* $\mathrm{Ku_T} \to 0$, *with all other nondimensional quantities held fixed, the solution for the immersed structure dynamics* $\{\mathbf{X}_{\boldsymbol{\alpha}}(t)\}_{\boldsymbol{\alpha} \in \mathcal{A}}$, *obtained from the complete coupled fluid-structure system and rescaled in time as*

$$\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t) = \lim_{\mathrm{Ku_T} \to 0} \mathbf{X}_{\boldsymbol{\alpha}}(t/\mathrm{Ku_T}),$$

converges in law to the solution of the following simplified stochastic differential system involving only the structure variables $\{\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t)\}$:

$$\begin{aligned} d\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t) &= \bar{\mathbf{V}}_{\boldsymbol{\alpha}}(\bar{\mathbf{X}}(t))\, dt + \sum_{\mathbf{k} \in S} \mathcal{S}_{\mathbf{k}}(\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t))\, d\tilde{\mathbf{W}}_{\mathbf{k}}(t), \\ \bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t = 0) &= \tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}}, \end{aligned} \tag{4.1}$$

where the stochastic complex white noise terms $d\tilde{\mathbf{W}}_{\mathbf{k}}(t)$ *are defined below* (2.3) *and are given the Itô interpretation. The explicit expression for the drift term is*

$$\bar{\mathbf{V}}_{\boldsymbol{\alpha}}(\mathbf{X}) = -\phi \sum_{\boldsymbol{\alpha}'} \bar{\mathcal{M}}(\mathbf{X}_{\boldsymbol{\alpha}} - \mathbf{X}_{\boldsymbol{\alpha}'}) \nabla_{\boldsymbol{\alpha}'} \tilde{\Phi}(\mathbf{X}/\tilde{\ell}_f), \tag{4.2}$$

and the matrix coefficients of the stochastic terms are

$$\mathcal{S}_{\mathbf{k}}(\mathbf{x}) = \sqrt{2\hat{\bar{\mathcal{M}}}_{\mathbf{k}}}\, e^{2\pi i \mathbf{k} \cdot \mathbf{x}/\tilde{K}}. \tag{4.3}$$

We have defined the mobility matrix function

$$\bar{\mathcal{M}}(\mathbf{r}) = \sum_{\mathbf{k}\in S} \hat{\bar{\mathcal{M}}}_{\mathbf{k}} e^{2\pi i \mathbf{k}\cdot\mathbf{r}/\tilde{K}} \tag{4.4}$$

and its Fourier coefficients

$$\hat{\bar{\mathcal{M}}}_{\mathbf{k}} \equiv \tilde{K}^3 \frac{\mathcal{P}_{\mathbf{k}} |\hat{\delta}_{1,\mathbf{k}}|^2}{4\pi^2 (k/\tilde{K})^2}.$$

The $\{\hat{\delta}_{1,\mathbf{k}}\}_{\mathbf{k}\in S}$ *are the Fourier coefficients of the delta function* $\delta_1(\mathbf{x})$ *(which we recall is to be viewed as a periodic smoothed delta function on the lattice generated by the fundamental cubic fluid domain),*

$$\hat{\delta}_{1,\mathbf{k}} = \frac{1}{\tilde{K}^3} \int_\Omega e^{-2\pi i \mathbf{k}\cdot\mathbf{x}/\tilde{K}} \delta_1(\mathbf{x})\, d\mathbf{x},$$

and we have introduced the projection tensor which enforces incompressibility of each Fourier velocity mode:

$$\mathcal{P}_{\mathbf{k}} = \mathcal{I} - \frac{\mathbf{k}\otimes\mathbf{k}}{k^2}.$$

The asymptotic statement in the proposition can be justified rigorously (see MTV section 4.4), provided that only a finite number of Fourier modes are retained. Later, in section 6, we modify these results to describe finite truncations which are better suited for numerical simulations. The asymptotics reported in the proposition are uniformly valid if the parameters ϕ and Υ are order unity or become small. Large values of ϕ might be of interest in structural models with some vibrational modes that may have time scales comparable to those of the fluid (or at least much faster than other translational and rotational modes of the structures). In this case, it may be desirable to apply the stochastic mode reduction procedure to eliminate some of the fast vibrational modes as well as the fluid modes. This falls outside the scope of our present results, and we leave its study for future work.

We will defer a discussion of the physical origin of the dynamics for $\bar{\mathbf{X}}(t)$ until section 5. There we will also develop the somewhat complicated formulas for the effective drift and diffusion coefficients into some more transparent consequents for one- and two-particle motion.

Here, we make only some brief remarks about the mathematical structure (section 4.1) and physical fidelity (section 4.2) of the effective dynamics, and comment on how the situation would be changed in a system which did not conserve momentum (section 4.3).

4.1. Mathematical remarks.

1. Perhaps somewhat surprisingly, in the low thermal Kubo number limit, the nonlinear advection term in the Navier–Stokes equation has no influence on the effective equation for $\bar{\mathbf{X}}(t)$. This fact is a consequence of the nonlinear advection term's having zero mean when averaged against the invariant measure (see (A.5)) for the velocity modes:

$$\langle \mathbf{u}\cdot\nabla\mathbf{u}\rangle = \langle \nabla\cdot(\mathbf{u}\otimes\mathbf{u})\rangle = \mathbf{0}.$$

A fundamental reason for the vanishing of the nonlinear advection term is the symmetry of the velocity field statistics under parity inversion:

$$\mathbf{x}\to -\mathbf{x}, \qquad \mathbf{u}\to -\mathbf{u}. \tag{4.5}$$

2. The factors of $\tilde{K}^3$ and $\tilde{K}^{3/2}$ in the drift and noise coefficients do not indicate a divergence with the size of the period cell, since $\hat{\delta}_{1,\mathbf{k}}$ scales as $O(\tilde{K}^{-3})$ and decays rapidly for $|k| \geq \tilde{K}$. Therefore, the $\tilde{K} \to \infty$ limits of the drift and noise contributions are in fact finite.

4.2. Physical fidelity of effective dynamics.

1. The drift term $\bar{\mathbf{V}}_{\boldsymbol{\alpha}}(\mathbf{X})$ is in accord with what one would expect on physical grounds, based on a low Reynolds number approximation in which the Navier–Stokes approximations are governed by a quasi-steady balance between the viscous and forcing terms (see section 5.1 below). The IB dynamics, however, do not produce a divergence-drift term, which, strictly speaking, should be present on physical grounds. We discuss this term in more detail in [26], where we can see its explicit form in the coarse-graining of particle-based dynamics. This term is small except when the elementary particles are close together (compared to their sizes); see [12, pp. 232–233].

2. The mobility matrix $\bar{\mathcal{M}}(\mathbf{r})$ describing the particle velocities in response to forces is a symmetric, nonnegative definite matrix function by Khinchin's theorem [53]. This implies that for any configuration of N particles the $3N \times 3N$ matrix relating the effective hydrodynamic velocity of each particle to the force on each particle is a symmetric, nonnegative definite matrix in the ordinary sense. This is also in accord with what one expects from physical arguments [19].

3. The random component of the IB dynamics can be shown to generate a physically appropriate absolute diffusion and relative diffusion of Lagrangian particles, provided that the particles are not too close together (relative to their sizes) [28]. In the IB dynamics, however, the motion of one particle is completely unaffected by the presence of other particles which do not generate force. In physical reality, though, the diffusion of a particle is hindered over the long run by the presence of other particles because they affect the fluid motion through a change of boundary conditions induced by rigidity of the particle, even if the particle does not induce any net force or torque [44, 45]. This correction to the motion of the particles is naturally proportional to the volume density occupied by the immersed particles.

4. The reduced system for the effective dynamics of the immersed structures obeys the Einstein relation [6, 16, 46]. The mobility matrix $\bar{\mathcal{M}}(\mathbf{r})$ in (4.4) is exactly the same as the matrix of relative diffusivities between different particles, as we show explicitly in section 5.2. The IB system should, on first principles, satisfy this Einstein relation because it is founded on statistical mechanical principles, but this fact is not at all transparent in the primitive formulation (2.9).

In summary, the IB method appears to generate physically correct dynamics of immersed structures in the presence of thermal fluctuations, provided that the particles constituting the structures are not too closely situated relative to their effective sizes. On the other hand, our analysis in the present work indicates that there is some quantitative difference between the statistical behavior of particles in the IB method and the physical behavior of rigid particles when the separation distance is comparable to the particle sizes. In practice, the structures (polymers, membranes, etc.) in the IB method are generally constructed with the elementary particles separated by a distance on the order of their effective size, so this regime is worth some scrutiny. One reason for the difference in behavior is surely that the Lagrangian particles in the IB method do not act on the fluid as rigid particles with a definite surface do. In particular, the fluid does not respond to the presence of an IB elementary particle unless that particle experiences a force, whereas a force-free rigid particle does exert

stress on the fluid to move it out of the way and to satisfy the no-slip condition on the surface. For physiological applications for which the IB method is primarily designed [42], the immersed structures are often elastic, so it may well be desirable to simulate it numerically using elementary particles that are not fully rigid. Moreover, the closely spaced IB particles will, often in applications, be modelled with some direct force interaction whose effects may dominate those of the rigidity of the particles. If some partial rigidity (or solidity) effects are still desired, however, they could perhaps be incorporated through a modification [41] of the IB approach described in [28]. A detailed exploration of theses issues is beyond the scope of this work and will be explored elsewhere.

We remark that the possible need for special handling of closely spaced particles in a fluid would not be unique to the IB method. Straightforward implementations of the particle-based method of simulations (to be described in [26]) based on Oseen (see [8]) or Rotne and Prager [47] hydrodynamic interaction approximations, which cause the divergence-drift term to cancel out, simulate rigid particle motion accurately only when their separation distance is large compared to their sizes. Only through a more elaborate introduction of lubrication forces as in Stokesian dynamics [3, 48] could the hydrodynamic interaction between closely spaced rigid particles be simulated accurately.

4.3. Changes in presence of nontrivial global system momentum. The simplified stochastic equations would require changes if the global system momentum were not a conserved quantity, such as if the system were subject to some fixed external potential. First of all, we would need to include the zero Fourier mode of the velocity,

$$\mathbf{u}_0(t) = \tilde{K}^{-3} \int_\Omega \mathbf{u}(\mathbf{x}, t)\, d\mathbf{x},$$

as a slow mode along with the immersed structure positions. The extent to which the effective dynamics are changed by the inclusion of this slow mode depends on its amplitude.

4.3.1. Weak zero velocity mode. If the slow mode $\mathbf{u}_0(t)$ has a small amplitude ($O(\mathrm{Ku_T})$ or less), then the stochastic mode reduction procedure can be carried through with simple changes, and the resulting effective equations would be changed in the following way:

1. The stochastic differential equations (4.1) for $d\mathbf{X}_{\boldsymbol{\alpha}}(t)$ would include a drift term $\mathbf{u}_0^\sharp(t)\, dt$, where $\mathbf{u}_0^\sharp = \mathrm{Ku_T^{-1}}\mathbf{u}_0(t)$.

2. The following evolution equation for $\mathbf{u}_0^\sharp(t)$ would be included in the effective dynamics:

$$d\mathbf{u}_0^\sharp(t) = -\phi^\sharp \sum_{\boldsymbol{\alpha}'} \nabla_{\boldsymbol{\alpha}'} \tilde{\Phi}(\mathbf{X}/\tilde{\ell}_f)\, dt,$$

where $\phi^\sharp = \phi \mathrm{Ku_T^{-1}}$.

Two situations in which the global system momentum can be expected to be weak are when

1. the forcing by the immersed structures is weak ($\phi \sim O(\mathrm{Ku_T})$), or

2. the external force couples to the system in such a way that the total force experienced self-averages to a small quantity when there are many immersed structures.

4.3.2. Strong zero velocity mode. If, on the other hand, the zero velocity mode has stronger amplitude, then more significant changes need to be made. For example, if the zero velocity mode has amplitude comparable with the (order unity) amplitudes of the other Fourier modes, then the hierarchy of significant dynamical variation is altered in the following way:

1. As before, the fluid velocity modes (other than the zero mode) vary on a fast time scale $O(\mathrm{Ku_T})$ (in the original nondimensional time coordinate).

2. The presence of a global system momentum would induce immersed structure motion on an $O(1)$ time scale and would itself vary on an $O(1)$ time scale as the structures were moved through the external potential.

3. The drift and diffusive motion due to internal forces and thermal fluctuations evolve on a slow time scale $O(\mathrm{Ku_T^{-1}})$.

Now, the fast velocity mode dynamics are unchanged by the presence of the global momentum. The evolution of the zero velocity Fourier mode and immersed structure positions on the $O(1)$ time scale is, to leading order, independent of the fast velocity modes. It is hard, however, to provide a general closed-form description for these $O(1)$ time scale dynamics for general nonlinear external potentials. Of course, the equations are easily solved for linear external potentials (such as gravity), but the resulting dynamics will be unphysical at long time scales (global system momentum growing unboundedly) unless the effects of backflow are somehow introduced. Of central interest in this paper is the effective drift-diffusive motion of the immersed structures on time scales $\mathrm{Ku_T^{-1}}$. The stochastic mode reduction procedure can be carried through to do this only if we can find an appropriate change of variables which removes the $O(1)$ time-scale motion of the particles. The formal procedure for doing so is presented in MTV section 5.3, but an explicit result requires a closed-form solution of the $O(1)$ time-scale dynamics, which is not generally available for the IB equations in the presence of a nonlinear external potential.

For simplicity, we hereafter consider only systems which conserve global system momentum, which we can arrange to be zero.

5. Physical discussion of effective dynamics. We wish here to provide some simple physical derivations of the drift and diffusion terms for the immersed structures as reported in Proposition 4.1, to provide an intuitive picture to complement the systematic mathematical derivation of section 3.

5.1. Drift term. At low Kubo number, the viscous dissipation term in the fluid momentum evolution equation in (2.9a) formally dominates the inertial terms. As the thermal forcing has mean zero, we can then suppose that the fluid motion inducing the deterministic part of the evolution of the immersed structures is given by the following simplified balance of viscous and pressure forces against the forces induced by the straining of the immersed structures:

$$\mathrm{Ku_T^{-1}}\nabla^2\mathbf{u}(\mathbf{x},t) - \nabla p(\mathbf{x},t) - \phi\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\nabla_{\boldsymbol{\alpha}}\tilde{\Phi}(\mathbf{X}(t)/\tilde{\ell}_f)\delta_1(\mathbf{x}-\mathbf{X}_{\boldsymbol{\alpha}}(t)) = 0,$$

$$\nabla\cdot\mathbf{u}(\mathbf{x},t) = 0.$$

Solving this linear system by a Fourier transform, we obtain

$$\mathbf{u}(\mathbf{x},t) = \sum_{\mathbf{k}\in S} e^{2\pi i\mathbf{k}\cdot\mathbf{x}/\tilde{K}}\hat{\mathbf{u}}_{\mathbf{k}}(t),$$

$$\hat{\mathbf{u}}_{\mathbf{k}}(t) = -\mathrm{Ku_T}\phi\sum_{\boldsymbol{\alpha}'\in\mathcal{A}}\frac{\mathcal{P}_{\mathbf{k}}\nabla_{\boldsymbol{\alpha}'}\tilde{\Phi}(\mathbf{X}(t)/\tilde{\ell}_f)\hat{\delta}_{1,\mathbf{k}}e^{-2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}'}(t)/\tilde{K}}}{4\pi^2(k/\tilde{K})^2},$$

with the Fourier symbols defined in section 3.1. Substituting this into the evolution equation in (3.2a) for $\mathbf{X}_{\boldsymbol{\alpha}}(t)$, we obtain the following expression for the deterministic component of the motion for the immersed structures at low Kubo number:

$$\left(\frac{d\mathbf{X}_{\boldsymbol{\alpha}}(t)}{dt}\right)_{\text{det}} = -\text{Ku}_{\text{T}}\phi\tilde{K}^3 \sum_{\boldsymbol{\alpha}'\in\mathcal{A}} \frac{\mathcal{P}_{\mathbf{k}}\nabla_{\boldsymbol{\alpha}'}\tilde{\Phi}(\mathbf{X}(t)/\tilde{\ell}_f)\hat{\delta}^2_{1,\mathbf{k}}e^{2\pi i\mathbf{k}\cdot(\mathbf{X}_{\boldsymbol{\alpha}}(t)-\mathbf{X}_{\boldsymbol{\alpha}'}(t))/\tilde{K}}}{4\pi^2(k/\tilde{K})^2}.$$

Note how the velocity induced by the immersed structure is $O(\text{Ku}_{\text{T}})$, and so will only produce a significant displacement over a $O(\text{Ku}_{\text{T}}^{-1})$ time scale. Upon rescaling time in this way, we recover the systematically computed drift coefficient (4.2).

5.2. Diffusion term. The stochastic dynamics of the immersed structures are in response to the stochastic fluctuations of the fluid, which are in turn due to the thermal forcing. We ignore structural forces here since they contribute, to leading order, to the deterministic rather than the random component of the particle motion. The motion of the structures is then purely random (with zero mean drift) and is described at a basic level by the evolution of the second-order moments of the coordinates. One such measure of the random motion, which can be cleanly computed, is the diffusion correlation tensor

$$\mathcal{D}(\mathbf{r}) \equiv \frac{1}{2}\frac{d}{dt}\left\langle (\mathbf{X}_{\boldsymbol{\alpha}}(t)-\mathbf{X}_{\boldsymbol{\alpha}}(t'))\otimes(\mathbf{X}_{\boldsymbol{\alpha}'}(t)-\mathbf{X}_{\boldsymbol{\alpha}'}(t'))\,|\mathbf{X}_{\boldsymbol{\alpha}}(t')=\mathbf{x}+\mathbf{r},\mathbf{X}_{\boldsymbol{\alpha}'}(t')=\mathbf{x}\right\rangle_{t=t'}, \tag{5.1}$$

which describes the correlation in the random motion of two particles during the moment at which they are situated with relative separation $\mathbf{r}$. Note that this diffusion correlation tensor does not depend on $\mathbf{x}$ nor t', provided that the fluid is in thermal equilibrium so that the system is statistically invariant under space and time translations. The single-particle diffusivity is just given by the diagonal entries $\mathcal{D}(\mathbf{0})$ because two coincident particles will move as a single particle due to the common random flow environment.

We now provide a direct but approximate calculation for the diffusion correlation tensor and then show that it agrees with the results of Proposition 4.1. To do this, we will ignore the nonlinear advection term in the fluid equation, on the grounds that the Reynolds number is small. We have the following approximation for the evolution of the random component of the Fourier modes of the fluid velocity field:

$$d\hat{\mathbf{u}}_{\mathbf{k}}(t) = -4\pi^2(k/\tilde{K})^2\text{Ku}_{\text{T}}^{-1}\hat{\mathbf{u}}_{\mathbf{k}}(t)\,dt + \text{Ku}_{\text{T}}^{-1/2}\tilde{K}^{-3/2}\sqrt{2(4\pi^2(k/\tilde{K})^2)}\,\mathcal{P}_{\mathbf{k}}d\tilde{\mathbf{W}}_{\mathbf{k}}(t).$$

These linear stochastic differential equations can be solved explicitly when the fluid is in thermal equilibrium. We find that each Fourier mode of the velocity field is independent of the others and evolves as a mean-zero Gaussian Markov process with correlation function

$$\begin{aligned}\langle\hat{\mathbf{u}}_{\mathbf{k}}(t)\otimes\hat{\mathbf{u}}_{\mathbf{k}}(t')\rangle &= 0,\\ \langle\hat{\mathbf{u}}_{\mathbf{k}}(t)\otimes\hat{\mathbf{u}}^*_{\mathbf{k}}(t')\rangle &= \tilde{K}^{-3}e^{-\text{Ku}_{\text{T}}^{-1}4\pi^2(k/\tilde{K})^2|t-t'|}\mathcal{P}_{\mathbf{k}}.\end{aligned}$$

The fluctuating component of the smoothed version of the velocity field, $\mathbf{u}_1$, which advects the immersed structures, also has independent Fourier coefficients which evolve

according to mean-zero Gaussian random processes with correlation structure

$$\begin{aligned}&\langle \hat{\mathbf{u}}_{1,\mathbf{k}}(t) \otimes \hat{\mathbf{u}}_{1,\mathbf{k}}(t')\rangle = 0,\\ &\langle \hat{\mathbf{u}}_{1,\mathbf{k}}(t) \otimes \hat{\mathbf{u}}^*_{1,\mathbf{k}}(t')\rangle = \tilde{K}^3 e^{-\mathrm{Ku}_\mathrm{T}^{-1} 4\pi^2 (k/\tilde{K})^2 |t-t'|} |\hat{\delta}_{1,\mathbf{k}}|^2 \mathcal{P}_{\mathbf{k}}.\end{aligned} \tag{5.2}$$

It will be convenient in the following development to define the increment in the Lagrangian particle positions:

$$\Delta \mathbf{X}_{\boldsymbol{\alpha}}(t) = \mathbf{X}_{\boldsymbol{\alpha}}(t+\Delta t) - \mathbf{X}_{\boldsymbol{\alpha}}(t).$$

Suppose that at time t we have two elementary particles situated at $\mathbf{X}_{\boldsymbol{\alpha}}(t) = \mathbf{x}+\mathbf{r}$ and $\mathbf{X}_{\boldsymbol{\alpha}'}(t) = \mathbf{x}$. The second-order moments of the changes in position of these particles due to the random thermal fluctuations in the fluid are given by

$$\begin{aligned}&\langle \Delta \mathbf{X}_{\boldsymbol{\alpha}}(t) \otimes \Delta \mathbf{X}_{\boldsymbol{\alpha}'}(t)\rangle\\ &\quad = \int_t^{t+\Delta t} ds \int_t^{t+\Delta t} ds' \, \langle \mathbf{u}_1(\mathbf{X}_{\boldsymbol{\alpha}}(s), s) \otimes \mathbf{u}_1(\mathbf{X}_{\boldsymbol{\alpha}'}(s'), s')\rangle\\ &\quad = \sum_{\mathbf{k}\in S} \sum_{\mathbf{k}'\in S} \int_t^{t+\Delta t} ds \int_t^{t+\Delta t} ds' \, \langle \hat{\mathbf{u}}_{1,\mathbf{k}}(s) \otimes \hat{\mathbf{u}}_{1,\mathbf{k}'}(s') e^{2\pi i (\mathbf{k}\cdot \mathbf{X}_{\boldsymbol{\alpha}}(s) + \mathbf{k}'\cdot \mathbf{X}_{\boldsymbol{\alpha}'}(s'))/\tilde{K}}\rangle.\end{aligned} \tag{5.3}$$

Suppose now that we consider $\mathrm{Ku}_\mathrm{T} \ll \Delta t \ll 1$. The velocity field decorrelates on the short time scale Ku_T (see (5.2)), but the particle positions, which integrate this fluid velocity, will change very little over the time interval Δt. Therefore, if we condition on the position of the particles at time t, as in the definition of the diffusion correlation tensor (5.1), we can reasonably approximate $\mathbf{X}_{\boldsymbol{\alpha}}(s)$ and $\mathbf{X}_{\boldsymbol{\alpha}'}(s')$ to be frozen within the last integrand in (5.3):

$$\begin{aligned}\mathcal{D}(\mathbf{r}) &\approx \frac{1}{2\Delta t} \sum_{\mathbf{k}\in S} \sum_{\mathbf{k}'\in S} \int_t^{t+\Delta t} ds \int_t^{t+\Delta t} ds' \, \langle \hat{\mathbf{u}}_{1,\mathbf{k}}(s) \otimes \hat{\mathbf{u}}_{1,\mathbf{k}'}(s')\rangle e^{2\pi i (\mathbf{k}\cdot(\mathbf{x}+\mathbf{r}) + \mathbf{k}'\cdot\mathbf{x})/\tilde{K}}\\ &= \frac{\tilde{K}^3}{2\Delta t} \sum_{\mathbf{k}\in S} |\hat{\delta}_{1,\mathbf{k}}|^2 \mathcal{P}_{\mathbf{k}} \int_t^{t+\Delta t} ds \int_t^{t+\Delta t} ds' \, e^{-\mathrm{Ku}_\mathrm{T}^{-1} 4\pi^2 (k/\tilde{K})^2 |s-s'|} e^{2\pi i \mathbf{k}\cdot\mathbf{r}/\tilde{K}}\\ &\approx \frac{\tilde{K}^3}{\Delta t} \mathrm{Ku}_\mathrm{T} \sum_{\mathbf{k}\in S} |\hat{\delta}_{1,\mathbf{k}}|^2 \frac{\mathcal{P}_{\mathbf{k}}}{4\pi^2 (k/\tilde{K})^2} e^{2\pi i \mathbf{k}\cdot\mathbf{r}/\tilde{K}} \Delta t \quad \text{for } \mathrm{Ku}_\mathrm{T} \ll \Delta t \ll 1.\end{aligned}$$

Therefore, the diffusion correlation tensor in the low Kubo number limit is given by the above approximate calculation as

$$\mathcal{D}(\mathbf{r}) = \tilde{K}^3 \mathrm{Ku}_\mathrm{T} \sum_{\mathbf{k}\in S} |\hat{\delta}_{1,\mathbf{k}}|^2 \frac{\mathcal{P}_{\mathbf{k}}}{4\pi^2 (k/\tilde{K})^2} e^{2\pi i \mathbf{k}\cdot\mathbf{r}/\tilde{K}}. \tag{5.4}$$

Notice again that this diffusion is $O(\mathrm{Ku}_\mathrm{T})$, so the random motion is significant only on $O(\mathrm{Ku}_\mathrm{T}^{-1})$ time scales.

After this rescaling, this heuristically deduced law of coupled diffusion of the immersed structures agrees with what the noise terms of the effective dynamics in Proposition 4.1 would produce in the absence of drift. Indeed, if we have $\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t) = \mathbf{x}+\mathbf{r}$ and $\bar{\mathbf{X}}_{\boldsymbol{\alpha}'}(t) = \mathbf{x}$, then applying the rules of (Itô) stochastic calculus [37], we have over

a short time interval of duration Δt

$$\begin{aligned}\langle \Delta\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t) \otimes \Delta\bar{\mathbf{X}}_{\boldsymbol{\alpha}'}(t)\rangle &= \left\langle \left(\sum_{\mathbf{k}\in S} \mathcal{S}_{\mathbf{k}}(\mathbf{x}+\mathbf{r})\,\Delta\tilde{\mathbf{W}}_{\mathbf{k}}(t)\right) \otimes \left(\sum_{\mathbf{k}'\in S} \mathcal{S}_{\mathbf{k}'}(\mathbf{x})\,\Delta\tilde{\mathbf{W}}_{\mathbf{k}'}(t)\right)\right\rangle \\ &\quad + o(\Delta t) \\ &= \sum_{\mathbf{k},\mathbf{k}'\in S} \langle \mathcal{S}_{\mathbf{k}}(\mathbf{x}+\mathbf{r}) \langle \Delta\tilde{\mathbf{W}}_{\mathbf{k}}(t) \otimes \Delta\tilde{\mathbf{W}}_{\mathbf{k}'}(t)\rangle \cdot \mathcal{S}^{\dagger}_{\mathbf{k}'}(\mathbf{x})\rangle,\end{aligned}$$

with the noise increment

$$\Delta\tilde{\mathbf{W}}_{\mathbf{k}}(t) \equiv \int_t^{t+\Delta t} d\tilde{\mathbf{W}}_{\mathbf{k}}(t).$$

Now using the statistical properties of the complex white noise processes (see (2.5) and surrounding discussion), we have

$$\begin{aligned}\langle \Delta\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t) \otimes \Delta\bar{\mathbf{X}}_{\boldsymbol{\alpha}'}(t)\rangle &= \sum_{\mathbf{k},\mathbf{k}'\in S} \langle \mathcal{S}_{\mathbf{k}}(\mathbf{x}+\mathbf{r}) \cdot \mathcal{I}\Delta t \delta_{\mathbf{k},-\mathbf{k}'} \cdot \mathcal{S}^{\dagger}_{\mathbf{k}'}(\mathbf{x})\rangle + o(\Delta t) \\ &= \sum_{\mathbf{k}\in S} \langle \mathcal{S}_{\mathbf{k}}(\mathbf{x}+\mathbf{r}) \cdot \mathcal{S}^{\dagger}_{-\mathbf{k}}(\mathbf{x})\rangle\,\Delta t + o(\Delta t).\end{aligned}$$

Substituting the expression (4.3) for $\mathcal{S}_{\mathbf{k}}(\mathbf{x})$, dividing by $2\Delta t$, and taking $\Delta t \to 0$, we obtain agreement with the heuristically deduced expression (5.4). The difference of the factor $\mathrm{Ku_T}$ is due to the fact that $\bar{\mathbf{X}}_{\boldsymbol{\alpha}}$ is defined with respect to the rescaled time $t/\mathrm{Ku_T}$.

We now directly observe the Einstein relation, which in our nondimensionalization reads $\mathcal{D}(\mathbf{r}) = \mathcal{M}(\mathbf{r}) = \mathrm{Ku_T}\bar{\mathcal{M}}(\mathbf{r})$, where $\mathcal{M}(\mathbf{r})$ is the mobility matrix (4.4) expressed in terms of the original time scale.

5.3. Discussion. The physical derivations of the drift and diffusion at low Kubo number are intended to provide some intuition for the results stated in Proposition 4.1. These formal arguments, however, involve assumptions which have varying degrees of plausibility and confidence, so the systematic and rigorous approach developed in section 3 is valuable. In particular, the systematic calculation allows precise assessment of the influence of the nonlinear advection term. Though indeed it is $O(\mathrm{Ku_T})$ weak relative to the viscous diffusion term, we are considering motion on $O(\mathrm{Ku_T^{-1}})$ time scales, so the nonlinearity could in principle have an $O(1)$ integrated influence on the particle motion. We found though, in section 3 that the nonlinear advection term does not in fact have an $O(1)$ effect due to cancellation caused by a parity symmetry (4.5) which it possesses.

6. Coarse-graining of the discretized IB method. In a numerical implementation, the velocity field can be represented by only a finite number of parameters. For a spectral code which retains the Fourier representation of the derivatives applied to the linear terms in the Navier–Stokes equations, the results of the continuum formulation would carry over by a simple Galerkin truncation. But such an abrupt spectral cutoff is usually not desirable in numerical simulations.

The version of the IB method implemented by Kramer and Peskin in [28] defines the fluid on a discrete, periodic, cubic mesh with (dimensional) spacing $h = L/K$, where K is an integer. The traditional IB method [42] has $h = a$, but there is no

difficulty in extending the numerical simulation approach for $h = ma$, where m is a nonnegative integer. An equivalent representation of the velocity field is through a finite set of Fourier coefficients $\{\hat{\mathbf{u}}_{\mathbf{k}}\}$ sufficient to resolve the velocity field on this mesh. The spatial derivatives appearing in the Navier–Stokes equations in (2.9a) must be replaced by operators with domain and range consistent with the finite-dimensional function space supported by the numerical resolution. The usual implementation is through finite-difference operators. We work out now how the effective dynamics of the particles are altered due to the discretization. These results are important for providing a benchmark against which numerical simulations can be compared more precisely. Temporal discretization lends itself less readily to the MTV framework, so we keep time continuous here.

6.1. Spatial discretization. We write the dynamical equations in the spatially discretized IB system (2.9) as

$$\begin{aligned}
&\frac{\partial \mathbf{u}(\mathbf{x},t)}{\partial t} + \mathcal{B}^{(d)}_{\mathbf{k}}(\mathbf{u},\mathbf{u})(\mathbf{x},t) = \mathrm{Ku}_{\mathrm{T}}^{-1}(\Delta^{(d)}_{\tilde{h}}\mathbf{u})(\mathbf{x},t) - (\nabla^{(d)}_{\tilde{h}}p)(\mathbf{x},t) + \mathbf{f}_T(t) \\
&\qquad\qquad - \phi \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}} \tilde{\Phi}\left(\frac{\mathbf{X}(t)}{\tilde{\ell}_f}\right) \delta_1(\mathbf{x} - \mathbf{X}_{\boldsymbol{\alpha}}(t)), \\
&(\tilde{\nabla}^{(d)}_{\tilde{h}} \cdot \mathbf{u})(\mathbf{x},t) = 0, \\
&\frac{d\mathbf{X}_{\boldsymbol{\alpha}}(t)}{dt} = \mathbf{u}_1(\mathbf{X}_{\boldsymbol{\alpha}}(t),t), \\
&\mathbf{u}_1(\mathbf{x},t) = \tilde{h}^3 \sum_{\mathbf{x}'\in\tilde{h}\mathbb{Z}^3_K} \mathbf{u}(\mathbf{x}',t)\delta_1(\mathbf{x}-\mathbf{x}').
\end{aligned} \tag{6.1}$$

In these equations, $\mathbf{x}$ is restricted to taking values on the periodic cubic lattice $\tilde{h}\times\mathbb{Z}^3_K$, where

$$\mathbb{Z}^3_K \equiv [1,2,\ldots,K]^3,$$

and $\tilde{h}$ is a nondimensional length scale ratio:

$$\tilde{h} \equiv \frac{h}{a} = \frac{\tilde{K}}{K}.$$

$\mathcal{B}^{(d)}_{\mathbf{k}}$ is some finite-difference approximation for the nonlinear advection operator, $\Delta^{(d)}_{\tilde{h}}$ is a discrete Laplacian, $\nabla^{(d)}_{\tilde{h}}$ is a discrete gradient, and $\tilde{\nabla}^{(d)}_{\tilde{h}}$ is another discrete gradient used in defining a divergence operator. The current implementation of the IB method [42] takes the usual centered-difference approximations for the linear differential operators, namely,

$$\begin{aligned}
\nabla^{(d)}_{\tilde{h}} &= \sum_{m=1}^{3} \hat{e}_m D^0_{\tilde{h},m}, \\
\Delta^{(d)}_{\tilde{h}} &= \sum_{m=1}^{3} D^+_{\tilde{h},m} D^-_{\tilde{h},m}, \\
\tilde{\nabla}^{(d)}_{\tilde{h}} &= \nabla^{(d)}_{\tilde{h}},
\end{aligned} \tag{6.2}$$

where

$$(D^+_{\tilde{h},m}g)(\mathbf{x}) = \frac{g(\mathbf{x}+\tilde{h}\hat{e}_m) - g(\mathbf{x})}{\tilde{h}},$$

$$(D^-_{\tilde{h},m}g)(\mathbf{x}) = \frac{g(\mathbf{x}) - g(\mathbf{x}-\tilde{h}\hat{e}_m)}{\tilde{h}},$$

$$(D^0_{\tilde{h},m}g)(\mathbf{x}) = \frac{g(\mathbf{x}+\tilde{h}\hat{e}_m) - g(\mathbf{x}-\tilde{h}\hat{e}_m)}{2\tilde{h}},$$

and $\hat{e}_m$ denotes a unit vector in the mth coordinate direction. The nonlinear term in the Navier–Stokes equation is discretized by a skew-symmetric central differencing scheme which conserves energy exactly (see [42]):

$$\mathcal{B}^{(d)}_{\mathbf{k}}(\mathbf{u},\mathbf{u})(\mathbf{x},t) \equiv \frac{1}{2}\left(\mathbf{u}(\mathbf{x},t)\cdot\nabla^{(d)}_{\tilde{h}}\mathbf{u}(\mathbf{x},t) + \nabla^{(d)}_{\tilde{h}}\cdot(\mathbf{u}(\mathbf{x},t)\otimes\mathbf{u}(\mathbf{x},t))\right).$$

Other discretizations can also be contemplated. For example, an upwind differencing scheme for the nonlinear term has often been used for numerical simulations at higher Reynolds number to provide numerical stability [43]. A purely divergence-form discretization would conserve total momentum exactly. Finally, Cowen [5] is investigating other discretizations for the divergence operator, adapted to the choice of the interpolation/spreading delta function, which improve volume conservation properties. For this reason, we will express the formulas for the effective drift and diffusion coefficients in terms of the general abstract differential operators $\nabla^{(d)}_{\tilde{h}}$, $\tilde{\nabla}^{(d)}_{\tilde{h}}$, and $\Delta^{(d)}_{\tilde{h}}$, without assuming that they take the specific form of (6.2). It is important to note, however, that the pressure term will conserve energy only if $\tilde{\nabla}^{(d)}_{\tilde{h}} = \nabla^{(d)}_{\tilde{h}}$.

Because the finite-difference derivative operators remain invariant under translations by a grid spacing $\tilde{h}$, they act diagonally as multiplication operators on Fourier modes. In this way we can define their action on Fourier modes $\mathbf{k}$ through the symbols $\mathcal{F}_{\mathbf{k}}(\cdot)$, defined in general for an operator $\mathcal{O}$ with translation invariance on the basic lattice, by

$$\mathcal{O}\mathbf{g}(\mathbf{x},t) = \sum_{\mathbf{k}\in\mathbb{Z}^3_K} e^{2\pi i\mathbf{k}\cdot\mathbf{x}/\tilde{K}}\mathcal{F}_{\mathbf{k}}(\mathcal{O})\hat{\mathbf{g}}(\mathbf{k},t),$$

where

$$\mathbf{g}(\mathbf{x},t) = \sum_{\mathbf{k}\in\mathbb{Z}^3_K} e^{2\pi i\mathbf{k}\cdot\mathbf{x}/\tilde{K}}\hat{\mathbf{g}}(\mathbf{k},t).$$

For the implemented version described in (6.2), the Fourier representation of the finite-difference operators would be

$$\mathcal{F}_{\mathbf{k}}(\Delta^{(d)}_{\tilde{h}}) = -\frac{4}{\tilde{h}^2}\sum_{m=1}^{3}\sin^2\frac{\pi k_m}{K},$$

$$\mathcal{F}_{\mathbf{k}}(\nabla^{(d)}_{\tilde{h}}) = \mathcal{F}_{\mathbf{k}}(\tilde{\nabla}^{(d)}_{\tilde{h}}) = \frac{i}{\tilde{h}}\sum_{m=1}^{3}\hat{e}_m\sin\frac{2\pi k_m}{K}.$$

Note that these Fourier representations of the discretized operators converge to the Fourier representation of the corresponding continuum operators as $K\to\infty$ (with the period cell length $\tilde{K} = K\tilde{h}$ held fixed).

The thermal forcing must be changed as follows (see [28]) in order to maintain the correct statistical mechanics for the spatially discretized system (6.1):

$$\mathbf{f}_T(\mathbf{x},t) = \mathrm{Ku}_\mathrm{T}^{-1/2}\tilde{K}^{-3/2}\sum_{\mathbf{k}\in S_K}\sqrt{2\mathcal{F}_\mathbf{k}(-\Delta_{\tilde{h}}^{(d)})}e^{2\pi i\mathbf{k}\cdot\mathbf{x}/\tilde{K}}\mathcal{P}_\mathbf{k}^{(dI)}\frac{d\tilde{\mathbf{W}}_\mathbf{k}(t)}{dt}, \tag{6.3}$$

where

$$S_K = \mathbb{Z}_K^3 \setminus (K,K,K) \tag{6.4}$$

and the projection tensor

$$\mathcal{P}_\mathbf{k}^{(dI)} = \mathcal{I} - \frac{\mathcal{F}_\mathbf{k}(\tilde{\nabla}_{\tilde{h}}^{(d)})\otimes\mathcal{F}_\mathbf{k}(\tilde{\nabla}_{\tilde{h}}^{(d)})}{\mathcal{F}_\mathbf{k}(\tilde{\nabla}_{\tilde{h}}^{(d)}\cdot\tilde{\nabla}_{\tilde{h}}^{(d)})} \tag{6.5}$$

must be included when $\tilde{\nabla}_{\tilde{h}}^{(d)} \neq \nabla_{\tilde{h}}^{(d)}$ for proper results. When $\tilde{\nabla}_{\tilde{h}}^{(d)} = \nabla_{\tilde{h}}^{(d)}$, the inclusion or omission of the factor $\mathcal{P}_\mathbf{k}^{(dI)}$ in (6.3) has no effect on the system dynamics.

6.2. Changes in drift and diffusion due to discretization. Whatever the precise forms of the discretized derivatives, the resulting drift and noise terms in the effective dynamics (4.1) have the form

$$\begin{aligned}\bar{\mathbf{V}}_{\boldsymbol{\alpha}}(\mathbf{X}) &= -\phi\sum_{\boldsymbol{\alpha}'\in\mathcal{A}}\nabla_{\boldsymbol{\alpha}'}\tilde{\Phi}\left(\frac{\mathbf{X}}{\bar{\ell}_f}\right)\bar{\mathcal{M}}^{(d)}(\mathbf{X}_{\boldsymbol{\alpha}'},\mathbf{X}_{\boldsymbol{\alpha}}-\mathbf{X}_{\boldsymbol{\alpha}'}),\\ \mathcal{S}_\mathbf{k}(\mathbf{x}) &= \sqrt{2}\tilde{K}^{3/2}\frac{\mathcal{P}_\mathbf{k}^{(dI)}e^{2\pi i\mathbf{k}\cdot\mathbf{x}/\tilde{K}}}{\sqrt{\mathcal{F}_\mathbf{k}(-\Delta_{\tilde{h}}^{(d)})}}\sum_{\mathbf{p}\in\mathbb{Z}^3}\hat{\delta}_{1,K\mathbf{p}+\mathbf{k}}e^{2\pi i\mathbf{p}\cdot\mathbf{x}/\tilde{h}},\end{aligned} \tag{6.6a}$$

with the mobility matrix function given by

$$\begin{aligned}\bar{\mathcal{M}}^{(d)}(\mathbf{x}',\mathbf{r}) &= \sum_{\mathbf{q}\in S}\sum_{\mathbf{p}\in\mathbb{Z}^3}\hat{\bar{\mathcal{M}}}_{\mathbf{p},\mathbf{q}}^{(d)}e^{2\pi i\mathbf{q}\cdot\mathbf{r}/\tilde{K}}e^{2\pi i\mathbf{p}\cdot\mathbf{x}/\tilde{h}},\\ \hat{\bar{\mathcal{M}}}_{\mathbf{p},\mathbf{q}}^{(d)} &= \tilde{K}^3\frac{\mathcal{P}_\mathbf{q}^{(dII)}\hat{\delta}_{1,K\mathbf{p}+\mathbf{q}}\hat{\delta}_{1,\mathbf{q}}^*}{\mathcal{F}_\mathbf{q}(-\Delta_{\tilde{h}}^{(d)})}.\end{aligned} \tag{6.6b}$$

Note that in general we must distinguish two versions of the projection tensor $\mathcal{P}_\mathbf{k}$ in the discretized formalism:

$$\mathcal{P}_\mathbf{k}^{(dII)} = \mathcal{I} - \frac{\mathcal{F}_\mathbf{k}(\nabla_{\tilde{h}}^{(d)})\otimes\mathcal{F}_\mathbf{k}(\tilde{\nabla}_{\tilde{h}}^{(d)})}{\mathcal{F}_\mathbf{k}(\nabla_{\tilde{h}}^{(d)}\cdot\tilde{\nabla}_{\tilde{h}}^{(d)})},$$

and $\mathcal{P}_\mathbf{k}^{(dI)}$ as given in (6.5). In the discretized formalism, the projection tensors $\mathcal{P}_\mathbf{k}$ appearing in the dynamical equation for the velocity modes in (3.2a), except in front of the noise term, are to be replaced by $\mathcal{P}_\mathbf{k}^{(dII)}$. On the other hand, the projection tensor $\mathcal{P}_\mathbf{k}$ appearing in front of the noise term and in (A.11) should be replaced by $\mathcal{P}_\mathbf{k}^{(dI)}$.

The effective drift and random terms for the discretized IB formalism have the same form as those appearing in the continuum IB formalism (Proposition 4.1), with the following key differences:

1. The mobility matrix function depends not only on the relative separation of the location of applied force and the responding particle, but also on the absolute location of the applied force relative to the fluid grid. Indeed, the mobility matrix function is readily seen to be periodic in the force location $\mathbf{x}'$ along the grid axes with period equal to the grid spacing $\tilde{h}$. This implies that the dynamics of the particles will depend somewhat on their position relative to the fluid grid. The grid-induced oscillations of the mobility matrix function are quantified by the $\mathbf{p} \neq \mathbf{0}$ terms in (6.6b) and will be studied quantitatively in [27].

2. The formulas involve a clear aliasing of wavenumbers separated by integer vector multiples of K, the number of grid points in each direction.

3. The mobility matrix function $\bar{\mathcal{M}}^{(d)}(\mathbf{r})$ is a symmetric, nonnegative definite matrix function if and only if the discretized derivatives obey $\nabla_{\tilde{h}}^{(d)} = \tilde{\nabla}_{\tilde{h}}^{(d)}$. In this case, the effective dynamics still obey the Einstein relation between the mobility matrix and the relative diffusivities of particles, though we no longer have a simple relation between the Fourier coefficients of the noise and the mobility matrix as in (4.3). The violation of the Einstein relation when $\nabla_{\tilde{h}}^{(d)} \neq \tilde{\nabla}_{\tilde{h}}^{(d)}$ occurs because the discretized pressure term does not conserve energy in this case.

4. The dissipation factor for each mode is naturally changed from the continuum value $4\pi^2(k/\tilde{K})^2$ to the value $\mathcal{F}_{\mathbf{k}}(-\Delta_{\tilde{h}}^{(d)})$ appropriate to the discretized viscous diffusion operator.

5. The sum is taken over a finite set of modes S_K (6.4), and there is no issue of ultraviolet divergence.

As we emphasize in section A.6, the nonlinear advection term in the continuum formulation makes no contribution to the effective structure dynamics in the low $\mathrm{Ku_T}$ limit due to a parity symmetry (4.5). This symmetry is also preserved under the implemented discretization (6.2) or if the nonlinear advection term is alternatively discretized by upwind differencing. In general, provided the discretization scheme respects the parity symmetry (4.5), the presence of the discretized weak nonlinear advection term does not change the effective dynamics of the particles on nondimensional time scales $O(\mathrm{Ku_T^{-1}})$ to leading order. If, however, the discretization violates this symmetry, there may be spurious contamination from the discretized nonlinearity.

The effective immersed structure dynamics derived for the immersed structures evolving according to the discretized IB method can be used as a design criterion for a numerical procedure. For example, the single-particle diffusivity (which is the same as $\mathcal{D}(\mathbf{0})$ given by a discretized modification of the formula (5.4) following the above discussion) may be used as a means to identify the effective size of the simulated particle. Of course this size will be order unity in our nondimensionalized units, but since the elementary particles are represented as smoothed delta functions rather than objects with rigid boundaries, it is not a priori clear how to associate a definite size value to the particles. The general Stokes–Einstein formula [16, 46] relating the diffusivity of a particle to its size, along with the explicit formulas for the effective diffusion of the IB particles, gives us a quantitative way to associate the simulation parameters with the effective particle sizes which are desired in a simulation. The above formulas for the coarse-grained discretized IB dynamics are used extensively in [28, 27] to explore how well the IB method replicates the correct statistical physics of immersed structure motion and to provide a benchmark for the results of numerical simulations.

7. Conclusions. We have demonstrated how the stochastic mode reduction framework developed by Majda, Timofeyev, and Vanden-Eijnden (MTV) can be ap-

plied to obtain coarse-grained approximations for the equations underlying the IB method for the simulation of microfluid systems with thermal fluctuations. In this way we were able to characterize rigorously the effective drift and diffusion behavior of immersed structures in the IB method. In particular, with full rigor, the equations governing the coarse-grained spatially continuous and discrete IB methods have been compared with each other and also with the desired physical behavior. Provided that the pressure gradient is discretized in an energy-conserving way and the nonlinear advection term is discretized in a manner (such as a central difference approximation) which respects the parity symmetry (4.5), the continuous and spatially discretized IB methods have structurally similar coarse-grained dynamics with explicit formulas identifying the drift and diffusion coefficients. The main difference between the structural form of the coarse-grained dynamics of the elementary particles in the IB method and those of physically proper rigid particles (as well as the particle-based simulation methods discussed in [26]) is the absence of the "divergence-drift" term for the IB method, which is significant only for Lagrangian particles spaced closely compared to their effective sizes. As discussed in section 4.2, one source of this discrepancy is lack of rigidity of the Lagrangian particles in the IB method, which may be desirable for physiological systems with flexible and elastic structures. Incorporation of rigidity into the IB method [41] might bring its effective dynamics into closer agreement with those of rigid particle-based methods [3, 8, 26, 48] at small separation distances. Finally, the explicit formulas for the effective diffusion of the Lagrangian particles under spatial discretization provide a means for choosing the size parameter precisely in applications so as to match desired diffusion coefficients.

Appendix. Details of computations for stochastic mode reduction. The details of the calculation from section 3 are presented here.

A.1. Derivation of Kolmogorov backward equation. We provide here a formal derivation of the Kolmogorov backward equation (3.3) based on a stochastic Taylor expansion [22, 37]. Consider the change of $\rho(s, \mathbf{X}, \mathbf{U}|t)$ over a small time interval Δs, using its definition (3.4) and the law of total expectation (see [49]):

$$\begin{aligned}
&\rho(s+\Delta s, \mathbf{X}, \mathbf{U}|t) - \rho(s, \mathbf{X}, \mathbf{U}|t) \\
&\quad = \rho(s+\Delta s, \mathbf{X}, \mathbf{U}|t) \\
&\qquad\qquad - \mathbb{E}\left[f(\mathbf{X}(t), \mathbf{U}(t))|\mathbf{X}(s) = \mathbf{X}, \mathbf{U}(s) = \mathbf{U}\right] \\
&\quad = \rho(s+\Delta s, \mathbf{X}, \mathbf{U}|t) \\
&\qquad\qquad - \mathbb{E}\left[\mathbb{E}\left(f(\mathbf{X}(t), \mathbf{U}(t))|\mathbf{X}(s+\Delta s), \mathbf{U}(s+\Delta s)\right)\big|\mathbf{X}(s) = \mathbf{X}, \mathbf{U}(s) = \mathbf{U}\right] \\
&\quad = \rho(s+\Delta s, \mathbf{X}, \mathbf{U}|t) \\
&\qquad\qquad - \mathbb{E}\left[\rho(s+\Delta s, \mathbf{X}(s+\Delta s), \mathbf{U}(s+\Delta s)|t)|\mathbf{X}(s) = \mathbf{X}, \mathbf{U}(s) = \mathbf{U}\right] \\
&\quad = \mathbb{E}\left[\rho(s+\Delta s, \mathbf{X}, \mathbf{U}|t) - \rho(s+\Delta s, \mathbf{X}(s+\Delta s), \mathbf{U}(s+\Delta s)|t)\right. \\
&\qquad\qquad\qquad\qquad \left.\big|\mathbf{X}(s) = \mathbf{X}, \mathbf{U}(s) = \mathbf{U}\right].
\end{aligned}$$

Now we will want to perform a stochastic Taylor expansion of the second term, using the following expressions for the increments in the system variables during the small time step Δs:

$$\mathbf{X}_{\boldsymbol{\alpha}}(s+\Delta s) - \mathbf{X}_{\boldsymbol{\alpha}}(s) = \tilde{K}^3 \sum_{\mathbf{k}\in S} e^{2\pi i \mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}(s)} \hat{\mathbf{u}}_{\mathbf{k}}(s)\hat{\delta}_{1,\mathbf{k}}\Delta s + O((\Delta s)^2),$$

$$\begin{aligned}\Delta\hat{\mathbf{u}}_{\mathbf{k}}(s) \equiv \hat{\mathbf{u}}_{\mathbf{k}}(s+\Delta s) - \hat{\mathbf{u}}_{\mathbf{k}}(s) &= -\mathcal{B}_{\mathbf{k}}(\mathbf{U}(s),\mathbf{U}(s))\Delta s - 4\pi^2(k/\tilde{K})^2 \mathrm{Ku}_{\mathrm{T}}^{-1}\hat{\mathbf{u}}_{\mathbf{k}}(s)\Delta s \\ &\quad - \phi\mathcal{P}_{\mathbf{k}} \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}}\tilde{\Phi}(\mathbf{X}_{\boldsymbol{\alpha}}(s)/\tilde{\ell}_f)\hat{\delta}_{1,\mathbf{k}} e^{-2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}(s)}\Delta s \\ &\quad + \mathrm{Ku}_{\mathrm{T}}^{-1/2}\tilde{K}^{-3/2}\sqrt{2(4\pi^2(k/\tilde{K})^2)}\,\mathcal{P}_{\mathbf{k}}\Delta\tilde{\mathbf{W}}_{\mathbf{k}}(s) + o(\Delta s),\end{aligned}$$

which arise upon integrating the equations of motion (3.2) over the short time interval Δs and using the Itô property of the noise [37]. The noise increments $\Delta\tilde{\mathbf{W}}_{\mathbf{k}}(s) = \int_s^{s+\Delta s} d\tilde{\mathbf{W}}_{\mathbf{k}}(s)$ are Gaussian, mean-zero complex random variables, which are independent except for the complex conjugacy property

$$\Delta\tilde{\mathbf{W}}_{-\mathbf{k}} = \overline{\Delta\tilde{\mathbf{W}}_{\mathbf{k}}}. \tag{A.1}$$

Their covariances are given by integration of (2.5):

$$\langle \Delta\tilde{\mathbf{W}}_{\mathbf{k}} \otimes \Delta\tilde{\mathbf{W}}_{\mathbf{k}}\rangle = 0, \qquad \langle \Delta\tilde{\mathbf{W}}_{\mathbf{k}} \otimes \overline{\Delta\tilde{\mathbf{W}}_{\mathbf{k}}}\rangle = \Delta s\mathcal{I}. \tag{A.2}$$

Now we can compute

$$\begin{aligned}&\rho(s+\Delta s,\mathbf{X},\mathbf{U}|t) - \rho(s,\mathbf{X},\mathbf{U}|t) \\ &\quad = \mathbb{E}\left[\rho(s+\Delta s,\mathbf{X},\mathbf{U}|t) - \left(\rho(s+\Delta s,\mathbf{X},\mathbf{U}|t) + \sum_{\boldsymbol{\alpha}\in\mathcal{A}}(\Delta\mathbf{X}_{\boldsymbol{\alpha}})\cdot\frac{\partial\rho(s,\mathbf{X},\mathbf{U}|t)}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\right.\right. \\ &\qquad\quad + \sum_{\mathbf{k}\in S}(\Delta\hat{\mathbf{u}}_{\mathbf{k}})\cdot\frac{\partial\rho(s,\mathbf{X},\mathbf{U}|t)}{\partial\hat{\mathbf{u}}_{\mathbf{k}}} \\ &\qquad\quad \left.\left. + \frac{1}{2}\sum_{\mathbf{k},\mathbf{k}'\in S}(\Delta\hat{\mathbf{u}}_{\mathbf{k}})\cdot\frac{\partial^2\rho(s,\mathbf{X},\mathbf{U}|t)}{\partial\hat{\mathbf{u}}_{\mathbf{k}}\partial\hat{\mathbf{u}}_{\mathbf{k}'}}\cdot(\Delta\hat{\mathbf{u}}_{\mathbf{k}'})\right)\,\middle|\,\mathbf{X}(s)=\mathbf{X},\mathbf{U}(s)=\mathbf{U}\right] \\ &\qquad + o(\Delta s) \\ &\quad = -\tilde{K}^3\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\sum_{\mathbf{k}\in S} e^{2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}(s)}\hat{\delta}_{1,\mathbf{k}}\hat{\mathbf{u}}_{\mathbf{k}}(s)\cdot\frac{\partial\rho(s,\mathbf{X},\mathbf{U}|t)}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\Delta s \\ &\qquad + \sum_{\mathbf{k}\in S}\left[\mathcal{B}_{\mathbf{k}}(\mathbf{U}(s),\mathbf{U}(s)) + 4\pi^2\left(\frac{k}{\tilde{K}}\right)^2\mathrm{Ku}_{\mathrm{T}}^{-1}\hat{\mathbf{u}}_{\mathbf{k}}(s)\right. \\ &\qquad\qquad \left. + \phi\mathcal{P}_{\mathbf{k}}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\nabla_{\boldsymbol{\alpha}}\tilde{\Phi}\left(\frac{\mathbf{X}_{\boldsymbol{\alpha}}(s)}{\tilde{\ell}_f}\right)\hat{\delta}_{1,\mathbf{k}}e^{-2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}(s)}\right]\cdot\frac{\partial\rho(s,\mathbf{X},\mathbf{U}|t)}{\partial\hat{\mathbf{u}}_{\mathbf{k}}}\Delta s \\ &\qquad - \sum_{\mathbf{k}\in S}\mathrm{Ku}_{\mathrm{T}}^{-1}\tilde{K}^{-3}\left(4\pi^2\left(\frac{k}{\tilde{K}}\right)^2\right)\mathcal{P}_{\mathbf{k}} : \frac{\partial^2\rho(s,\mathbf{X},\mathbf{U}|t)}{\partial\hat{\mathbf{u}}_{\mathbf{k}}\partial\hat{\mathbf{u}}_{\mathbf{k}}^*}\Delta s + o(\Delta s).\end{aligned} \tag{A.3}$$

We used (A.1) and (A.2) to collapse the sum over $\mathbf{k}$ and $\mathbf{k}'$ in the last term to a single sum (with $\mathbf{k}' = -\mathbf{k}$). Dividing the final relation by Δs, then sending $\Delta s \to 0$, leads to the Kolmogorov backward equation (3.3). The $\mathcal{P}_{\mathbf{k}}$ can be dropped from the term involving second derivatives of $\hat{\mathbf{u}}_{\mathbf{k}}$ because $\mathbf{k}\cdot\partial/(\partial\hat{\mathbf{u}}_{\mathbf{k}}) = 0$ follows automatically from the fact that $\hat{\mathbf{u}}_{\mathbf{k}}$ is understood to always be restricted so that $\mathbf{k}\cdot\hat{\mathbf{u}}_{\mathbf{k}} = 0$.

A.2. Asymptotic expansion of solution. We show now how to derive the limiting equation (3.6) from the original Kolmogorov backward equation (3.5) with the small parameter ε. We expand the solution in powers of ε:

$$\rho^\varepsilon = \rho_0 + \varepsilon\rho_1 + \varepsilon^2\rho_2 + \cdots.$$

Then, writing out the first three equations of the asymptotic hierarchy, we have

$$\begin{aligned} &\mathcal{L}_1\rho_0 = 0, \\ &\mathcal{L}_1\rho_1 = -\mathcal{L}_2\rho_0, \\ &\mathcal{L}_1\rho_2 = -\frac{\partial\rho_0}{\partial s} - \mathcal{L}_2\rho_1. \end{aligned} \tag{A.4}$$

We will now solve these in succession; all three must be considered to obtain a full description of the evolution of ρ_0. Our presentation will take the form of a formal calculation, but the results find rigorous support from the theorem of Kurtz [9, 30] for any finite truncation of the set of modes S.

Since we are interested in deriving effective equations for only the (slow) particle coordinates $\mathbf{X}$, we can restrict attention to initial data which depends only on $\mathbf{X}$:

$$f = f(\mathbf{X}).$$

In the solvability conditions that will follow after consideration of the second- and higher-order equations of the hierarchy, it will be helpful to explicitly identify some analytical properties of $\mathcal{L}_1$. Since $\mathcal{L}_1$ is the generator of the Ornstein–Uhlenbeck process, the projection $\mathcal{P}_N$ onto its null space can be identified with the operation of averaging against its invariant measure (MTV):

$$\begin{aligned} (\mathcal{P}_N g)(\mathbf{X}) = (\mathbb{E}_{OU} g)(\mathbf{X}) &\equiv \int_{\mathbb{C}^S} d\mathbf{U}\, \pi_{OU}(\mathbf{U}) g(\mathbf{X}, \mathbf{U}), \\ \pi_{OU}(\mathbf{U}) &= \left(\prod_{\mathbf{k}\in S} \frac{\tilde{K}^3}{4\pi} \right) \exp\left(-\frac{1}{2}\tilde{K}^3 \sum_{\mathbf{k}\in S} |\hat{\mathbf{u}}_{\mathbf{k}}|^2 \right). \end{aligned} \tag{A.5}$$

Also, the null space of $\mathcal{L}_1^*$ is exactly spanned by π_{OU}, so in applying the solvability conditions, it is helpful to note that (see [13])

$$g \in \operatorname{Ran}\mathcal{L}_1 \Leftrightarrow g \in (\operatorname{Ker}\mathcal{L}_1^*)^\perp \Leftrightarrow \mathbb{E}_{OU} g = 0. \tag{A.6}$$

A.3. First equation in asymptotic hierarchy. The leading-order equation in (A.4) implies simply that ρ_0 does not depend on the fast variables $\mathbf{U}$:

$$\rho_0 = \rho_0(s, \mathbf{X}|t).$$

Another way of expressing this is

$$\mathbb{E}_{OU}\rho_0 = \rho_0.$$

A.4. Second equation in asymptotic hierarchy. The solvability condition for the second equation in (A.4) is

$$\mathbb{E}_{OU}\mathcal{L}_2\rho_0 = \mathbb{E}_{OU}\mathcal{L}_2\mathbb{E}_{OU}\rho_0 = 0,$$

which is trivially satisfied because $\mathbb{E}_{OU}\mathcal{L}_2\mathbb{E}_{OU} = 0$.

Therefore, the equation may be solved directly by writing

$$\rho_1 = -\mathcal{L}_1^{-1}\mathcal{L}_2\mathbb{E}_{OU}\rho_0 + \tilde{\rho}_1. \tag{A.7}$$

Since $\mathcal{L}_1$ has a one-dimensional null-space, the inverse operator $\mathcal{L}_1^{-1}$ should be thought of as a particular continuous choice of an inverse image associated with each function in the range of $\mathcal{L}_1$. We make a specific choice in section A.6. The function $\tilde{\rho}_1 = \tilde{\rho}_1(\mathbf{X}, s|t)$ is a function in the null space of $\mathcal{L}_1$; its presence reflects the one-dimensional indeterminacy of the inversion of $\mathcal{L}_1$.

A.5. Third equation in asymptotic hierarchy. Substituting the solution (A.7) into the third equation in (A.4) and applying the solvability condition, we obtain the desired evolution equation for ρ_0 in operator-theoretic form:

$$\begin{aligned} -\frac{\partial \rho_0}{\partial s} &= -\mathbb{E}_{OU}\mathcal{L}_2\mathcal{L}_1^{-1}\mathcal{L}_2\mathbb{E}_{OU}\rho_0, \\ \rho_0(s = t, \mathbf{X}|t) &= f(\mathbf{X}). \end{aligned} \tag{A.8}$$

The arbitrary function $\tilde{\rho}_1 \in \operatorname{Ker}\mathcal{L}_1$ has now disappeared because $\mathbb{E}_{OU}\mathcal{L}_2\tilde{\rho}_1 = 0$. Therefore, we see that the evolution equation for ρ_0 will not depend on the particular way in which we choose to invert $\mathcal{L}_1$.

A.6. Explicit computation of limiting PDE. To express the differential operator

$$\bar{\mathcal{L}} \equiv -\mathbb{E}_{OU}\mathcal{L}_2\mathcal{L}_1^{-1}\mathcal{L}_2\mathbb{E}_{OU},$$

appearing on the right-hand side of (A.8), in a concrete form, we follow the development in Appendix B of MTV. To map the formulas appearing there to the present problem, we write

$$\mathcal{L}_1 = \sum_{\mathbf{k}\in S}\left(-\gamma_{\mathbf{k}}\hat{\mathbf{u}}_{\mathbf{k}}\cdot\frac{\partial}{\partial\hat{\mathbf{u}}_{\mathbf{k}}} + \frac{\sigma_{\mathbf{k}}^2}{2}\frac{\partial}{\partial\hat{\mathbf{u}}_{\mathbf{k}}}\cdot\frac{\partial}{\partial\hat{\mathbf{u}}_{\mathbf{k}}^*}\right),$$

with

$$\gamma_{\mathbf{k}} \equiv 4\pi^2(k/\tilde{K})^2, \qquad \sigma_{\mathbf{k}} \equiv \sqrt{2\tilde{K}^{-3}(4\pi^2(k/\tilde{K})^2)}. \tag{A.9}$$

We pass to the representation of $\mathcal{L}_2$ acting on Fourier transformed functions of the fast variables $\mathbf{U}$. We must be careful in defining this Fourier transform, however, because the variables constituting $\mathbf{U}$ are complex and are constrained by the complex conjugacy relations $\hat{\mathbf{u}}_{-\mathbf{k}} = \hat{\mathbf{u}}_{\mathbf{k}}^*$. Therefore, we define the Fourier transform of functions $g(\mathbf{U})$ by

$$\hat{g}(\mathbf{P}) = \int_{\mathbb{C}^S}\exp\left[\frac{1}{4}i\sum_{\mathbf{k}\in S}\left(\hat{\mathbf{u}}_{\mathbf{k}}\cdot(\mathbf{p}_{\mathbf{k}} + \mathbf{p}_{-\mathbf{k}}^*) + \hat{\mathbf{u}}_{\mathbf{k}}^*\cdot(\mathbf{p}_{-\mathbf{k}} + \mathbf{p}_{\mathbf{k}}^*)\right)\right] g(\mathbf{U})\, d\mathbf{U},$$

where $\mathbf{P} = \{\mathbf{p}_{\mathbf{k}}\}_{\mathbf{k}\in S}$. This artifice first ensures that the exponent is purely imaginary (so that the Fourier integration is at least well defined in each mode), and second allows us to identify $\mathbf{p}_{-\mathbf{k}}$ with $\mathbf{p}_{\mathbf{k}}^*$ just as we have been identifying $\hat{\mathbf{u}}_{-\mathbf{k}}$ with $\hat{\mathbf{u}}_{\mathbf{k}}^*$. Indeed, $\hat{g}$ must perforce depend on $\mathbf{p}_{\mathbf{k}}$ and $\mathbf{p}_{-\mathbf{k}}^*$ in the same way, and thus these

variables can be identified with each other. Also, we can and will restrict the domain of definition of the variables $\mathbf{p_k}$ to the hyperplane $\mathbf{p_k} \cdot \mathbf{k} = 0$; the component of $\mathbf{p_k}$ parallel to $\mathbf{k}$ is irrelevant because of the restriction $\hat{\mathbf{u}}_\mathbf{k} \cdot \mathbf{k} = 0$. Finally, it can be readily checked that the Fourier transform rules for derivatives carry over to our present definition of the Fourier transform in a straightforward way:

$$\hat{\mathbf{u}}_\mathbf{k} \to -i \frac{\partial}{\partial \mathbf{p_k}}, \qquad \frac{\partial}{\partial \hat{\mathbf{u}}_\mathbf{k}} \to -i \mathbf{p_k}.$$

Proceeding, then, we write down the Fourier transform of the operator $\mathcal{L}_2$ as

$$\begin{aligned}\hat{\mathcal{L}}_2 = \sum_{\mathbf{k}\in S} \Bigg[& -2\pi \tilde{K}^{-1} \mathcal{P}_\mathbf{k} \sum_{\mathbf{k}'\in S} \mathbf{k} \cdot \frac{\partial}{\partial \mathbf{p}_{\mathbf{k}'}} \frac{\partial}{\partial \mathbf{p}_{\mathbf{k}-\mathbf{k}'}} \\ & + i\phi \mathcal{P}_\mathbf{k} \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}} \tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right) \hat{\delta}_{1,\mathbf{k}} e^{-2\pi i \mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}} \Bigg] \cdot \mathbf{p_k} \\ & - i\tilde{K}^3 \sum_{\mathbf{k}\in S} \hat{\delta}_{1,\mathbf{k}} \sum_{\boldsymbol{\alpha}\in\mathcal{A}} e^{2\pi i \mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}} \frac{\partial}{\partial \mathbf{p_k}} \cdot \frac{\partial}{\partial \mathbf{X}_{\boldsymbol{\alpha}}}.\end{aligned}$$

We can now adapt relation (B.4) of MTV to our present case with complex-valued variables:

$$\begin{aligned}& -\mathbb{E}_{OU} \mathcal{L}_2 \mathcal{L}_1^{-1} \mathcal{L}_2 \mathbb{E}_{OU} g(\mathbf{X}) \\ & \quad = \int_{\mathbb{C}^S} d\mathbf{P}\, \hat{\bar{P}}_{OU}(\mathbf{P}) \hat{\mathcal{L}}_2 \int_0^\infty dt \\ & \quad \exp\left(\sum_{\mathbf{k}\in S} \gamma_\mathbf{k} t - \frac{1}{4} \sum_{\mathbf{k}\in S} \frac{\sigma_\mathbf{k}^2 |\mathbf{p_k}|^2}{\gamma_\mathbf{k}} (e^{2\gamma_\mathbf{k} t} - 1) \right) \left[\hat{\mathcal{L}}_2 (g(\mathbf{X}) \delta(\mathbf{P}')) \right]_{\mathbf{P}' = \boldsymbol{\beta}(\mathbf{P},t)},\end{aligned}$$

where $\mathbf{P}$ is shorthand notation for the collection $\{\mathbf{p_k}\}_{\mathbf{k}\in S}$ (similarly for $\mathbf{P}'$),

$$\hat{\bar{P}}_{OU}(\mathbf{P}) = \exp\left(-\frac{1}{4} \sum_{\mathbf{k}\in S} \frac{\sigma_\mathbf{k}^2 |\mathbf{p_k}|^2}{\gamma_\mathbf{k}} \right)$$

is the Fourier transform of the invariant measure $\pi_{OU}(\mathbf{U})$ in (A.5), and

$$\boldsymbol{\beta}(\mathbf{P}, t) = \{ e^{\gamma_\mathbf{k} t} \mathbf{p_k} \}_{\mathbf{k}\in S}.$$

We have here chosen to define $\mathcal{L}_1^{-1}$ in terms of its Fourier transform:

$$(\hat{\mathcal{L}}_1^{-1} b)(\mathbf{P}) = -\int_0^\infty \exp\left(-\frac{1}{4} \sum_{\mathbf{k}\in S} \frac{\sigma_\mathbf{k}^2 |\mathbf{p_k}|^2}{\gamma_\mathbf{k}} (e^{2\gamma_\mathbf{k} t} - 1) \right) \exp\left(\sum_{\mathbf{k}\in S} \gamma_\mathbf{k} t \right) \hat{b}(\boldsymbol{\beta}(\mathbf{P}, t))\, dt.$$

Computing now the action of the rightmost $\hat{\mathcal{L}}_2$, we obtain

$$\begin{aligned}& \exp\left(\sum_{\mathbf{k}\in S} \gamma_\mathbf{k} t \right) \left(\left[\hat{\mathcal{L}}_2 (g(\mathbf{X}) \delta(\mathbf{P}')) \right]_{\mathbf{P}' = \boldsymbol{\beta}(\mathbf{P},t)} \right) \\ & \qquad = -i\tilde{K}^3 \sum_{\mathbf{k}\in S} \hat{\delta}_{1,\mathbf{k}} e^{-\gamma_\mathbf{k} t} \sum_{\boldsymbol{\alpha}\in\mathcal{A}} e^{2\pi i \mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}} \frac{\partial g}{\partial \mathbf{X}_{\boldsymbol{\alpha}}} \cdot \frac{\partial \delta(\mathbf{P})}{\partial \mathbf{p_k}}.\end{aligned}$$

Continuing,

$$\begin{aligned}
&\hat{\mathcal{L}}_1^{-1}\hat{\mathcal{L}}_2(g(\mathbf{X})\delta(\mathbf{P}))\\
&\quad= -\int_0^\infty dt\, \exp\left(-\frac{1}{4}\sum_{\mathbf{k}\in S}\frac{\sigma_{\mathbf{k}}^2|\mathbf{p}_{\mathbf{k}}|^2}{\gamma_{\mathbf{k}}}(e^{2\gamma_{\mathbf{k}}t}-1)\right)\\
&\qquad\qquad\times\left[-i\tilde{K}^3\sum_{\mathbf{k}\in S}\hat{\delta}_{1,\mathbf{k}}e^{-\gamma_{\mathbf{k}}t}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}e^{2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}}\frac{\partial g}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\cdot\frac{\partial\delta(\mathbf{P})}{\partial\mathbf{p}_{\mathbf{k}}}\right]\\
&\quad= i\tilde{K}^3\sum_{\mathbf{k}\in S}\hat{\delta}_{1,\mathbf{k}}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}e^{2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}}\frac{\partial g}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\cdot\frac{\partial\delta(\mathbf{P})}{\partial\mathbf{p}_{\mathbf{k}}}\int_0^\infty e^{-\gamma_{\mathbf{k}}t}\,dt\\
&\quad= i\tilde{K}^3\sum_{\mathbf{k}\in S}\frac{\hat{\delta}_{1,\mathbf{k}}}{\gamma_{\mathbf{k}}}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}e^{2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}}\frac{\partial g}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\cdot\frac{\partial\delta(\mathbf{P})}{\partial\mathbf{p}_{\mathbf{k}}}.
\end{aligned}$$

Next, using the distribution identity $\mathbf{p}_{\mathbf{k}}\otimes\frac{\partial\delta(\mathbf{p}_{\mathbf{k}})}{\partial\mathbf{p}_{\mathbf{k}}}=-\mathcal{I}\delta(\mathbf{p}_{\mathbf{k}})$, we have

$$\begin{aligned}
&\hat{\mathcal{L}}_2\hat{\mathcal{L}}_1^{-1}\hat{\mathcal{L}}_2(g(\mathbf{X})\delta(\mathbf{P}))\\
&\quad= -i\tilde{K}^3\sum_{\mathbf{k}\in S}\left[-2\pi\tilde{K}^{-1}\mathcal{P}_{\mathbf{k}}\sum_{\mathbf{k}'\in S}\mathbf{k}\cdot\frac{\partial}{\partial\mathbf{p}_{\mathbf{k}'}}\frac{\partial}{\partial\mathbf{p}_{\mathbf{k}-\mathbf{k}'}}\right.\\
&\qquad\qquad\left.+\,i\phi\mathcal{P}_{\mathbf{k}}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\nabla_{\boldsymbol{\alpha}}\tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right)\hat{\delta}_{1,\mathbf{k}}e^{-2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}}\right]\\
&\qquad\qquad\cdot\left[\frac{\hat{\delta}_{1,\mathbf{k}}}{\gamma_{\mathbf{k}}}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}e^{2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}}\frac{\partial g(\mathbf{X})}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\delta(\mathbf{P})\right]\\
&\qquad+\tilde{K}^6\sum_{\mathbf{k},\mathbf{k}'\in S}\frac{\hat{\delta}_{1,\mathbf{k}}\hat{\delta}_{1,\mathbf{k}'}}{\gamma_{\mathbf{k}}}\sum_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathcal{A}}e^{2\pi i(\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}+\mathbf{k}'\cdot\mathbf{X}_{\boldsymbol{\alpha}'})/\tilde{K}}\frac{\partial^2 g(\mathbf{X})}{\partial\mathbf{X}_{\boldsymbol{\alpha}}\partial\mathbf{X}_{\boldsymbol{\alpha}'}}:\frac{\partial^2\delta(\mathbf{P})}{\partial\mathbf{p}_{\mathbf{k}}\partial\mathbf{p}_{\mathbf{k}'}}\\
&\qquad+\tilde{K}^6\sum_{\mathbf{k},\mathbf{k}'\in S}\frac{\hat{\delta}_{1,\mathbf{k}}\hat{\delta}_{1,\mathbf{k}'}}{\gamma_{\mathbf{k}}}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}e^{2\pi i(\mathbf{k}+\mathbf{k}')\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}}\frac{\partial g(\mathbf{X})}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\cdot\frac{\partial^2\delta(\mathbf{P})}{\partial\mathbf{p}_{\mathbf{k}}\partial\mathbf{p}_{\mathbf{k}'}}\cdot(2\pi i\mathbf{k}).
\end{aligned}\tag{A.10}$$

Finally, taking the leftmost expectation defining $\bar{\mathcal{L}}$, we have

$$\begin{aligned}
\bar{\mathcal{L}}g(\mathbf{X}) &= -\int_{\mathbb{C}^S}\hat{\bar{P}}_{OU}(\mathbf{P})\hat{\mathcal{L}}_2\hat{\mathcal{L}}_1^{-1}\hat{\mathcal{L}}_2(g(\mathbf{X})\delta(\mathbf{P}))\\
&= -\tilde{K}^3\sum_{\mathbf{k}\in S}\phi\mathcal{P}_{\mathbf{k}}\sum_{\boldsymbol{\alpha}'\in\mathcal{A}}\nabla_{\boldsymbol{\alpha}'}\tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right)\hat{\delta}_{1,\mathbf{k}}e^{-2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}'}/\tilde{K}}\hat{\delta}_{1,\mathbf{k}}\gamma_{\mathbf{k}}^{-1}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}e^{2\pi i\mathbf{k}\cdot\mathbf{X}_{\boldsymbol{\alpha}}/\tilde{K}}\cdot\frac{\partial g(\mathbf{X})}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\\
&\quad+\tilde{K}^6\sum_{\mathbf{k}\in S}\frac{|\hat{\delta}_{1,\mathbf{k}}|^2\sigma_{\mathbf{k}}^2}{2\gamma_{\mathbf{k}}^2}\sum_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathcal{A}}e^{2\pi i\mathbf{k}\cdot(\mathbf{X}_{\boldsymbol{\alpha}}-\mathbf{X}_{\boldsymbol{\alpha}'})/\tilde{K}}\frac{\partial}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\cdot\mathcal{P}_{\mathbf{k}}\cdot\frac{\partial g(\mathbf{X})}{\partial\mathbf{X}_{\boldsymbol{\alpha}'}}\\
&\quad+\tilde{K}^6\sum_{\mathbf{k},\mathbf{k}'\in S}\frac{\hat{\delta}_{1,\mathbf{k}}\hat{\delta}_{1,\mathbf{k}'}\sigma_{\mathbf{k}}^2}{2\gamma_{\mathbf{k}}^2}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\frac{\partial g(\mathbf{X})}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\cdot\mathcal{P}_{\mathbf{k}}\cdot(2\pi i\mathbf{k})
\end{aligned}$$

$$= -\phi\tilde{K}^3 \sum_{\mathbf{k}\in S} \frac{\hat{\delta}_{1,\mathbf{k}}^2}{\gamma_{\mathbf{k}}} \mathcal{P}_{\mathbf{k}} \sum_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}'} \tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right) e^{2\pi i\mathbf{k}\cdot(\mathbf{X}_{\boldsymbol{\alpha}}-\mathbf{X}_{\boldsymbol{\alpha}'})/\tilde{K}} \cdot \frac{\partial g(\mathbf{X})}{\partial \mathbf{X}_{\boldsymbol{\alpha}}}$$

$$+ \tilde{K}^6 \sum_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathcal{A}} \sum_{\mathbf{k}\in S} \frac{|\hat{\delta}_{1,\mathbf{k}}|^2\sigma_{\mathbf{k}}^2}{2\gamma_{\mathbf{k}}^2} e^{2\pi i\mathbf{k}\cdot(\mathbf{X}_{\boldsymbol{\alpha}}-\mathbf{X}_{\boldsymbol{\alpha}'})/\tilde{K}} \frac{\partial}{\partial \mathbf{X}_{\boldsymbol{\alpha}}} \cdot \mathcal{P}_{\mathbf{k}} \cdot \frac{\partial g(\mathbf{X})}{\partial \mathbf{X}_{\boldsymbol{\alpha}'}}.$$

We have used the fact that

$$\mathbb{E}_{OU} \frac{\partial^2 \delta(\mathbf{P})}{\partial \mathbf{p}_{\mathbf{k}} \partial \mathbf{p}_{\mathbf{k}'}} = \begin{cases} -\frac{\sigma_{\mathbf{k}}^2}{2\gamma_{\mathbf{k}}} \mathcal{P}_{\mathbf{k}} & \text{if } \mathbf{k}+\mathbf{k}'=0, \\ 0 & \text{otherwise.} \end{cases} \tag{A.11}$$

We show in more detail how the contribution from the first term in brackets in (A.10), arising from the nonlinear advection term in the Navier–Stokes equations, vanishes upon averaging. From (A.11) we have

$$\mathbb{E}_{OU} \sum_{\mathbf{k}'} \frac{\partial}{\partial \mathbf{p}_{\mathbf{k}'}} \frac{\partial \delta(\mathbf{P})}{\partial \mathbf{p}_{\mathbf{k}-\mathbf{k}'}} = 0 \quad \text{unless } \mathbf{k} = \mathbf{0}.$$

However, since this expression appears in an inner product with $\mathbf{k}$ in (A.10), the term arising from nonlinear advection in the Navier–Stokes equations makes no contribution at all upon averaging.

We finally complete the calculation of $\bar{\mathcal{L}}$ by converting back to our original problem parameters, using the expressions (A.9) for $\gamma_{\mathbf{k}}$ and $\sigma_{\mathbf{k}}$. We thereby arrive at (3.6).

Acknowledgments. The authors thank Eric Vanden-Eijnden for helpful discussions and suggestions, and the referees for constructive comments which improved the presentation.

REFERENCES

[1] M. Avellaneda and M. Vergassola, *Stieltjes integral representation of effective diffusivities in time-dependent flows*, Phys. Rev. E, 52 (1995), pp. 3249–3251.

[2] A. N. Borodin, *A limit theorem for solutions of differential equations with random right-hand side*, Theor. Probability Appl., 22 (1977), pp. 482–497.

[3] J. F. Brady and G. Bossis, *Stokesian dynamics*, in Annu. Rev. Fluid Mech. 20, Annual Reviews, Palo Alto, CA, 1988, pp. 111–157.

[4] R. A. Carmona and L. Xu, *Homogenization for time-dependent two-dimensional incompressible Gaussian flows*, Ann. Appl. Probab., 7 (1997), pp. 265–279.

[5] N. Cowen, Ph.D. thesis, Courant Institute of Mathematical Sciences, New York University, New York, 2001.

[6] A. Einstein, *Investigations on the Theory of the Brownian Movement*, R. Fürth, ed., A. D. Cowper, translator, Dover Publications, New York, 1956.

[7] R. S. Ellis and M. A. Pinsky, *The first and second fluid approximations to the linearized Boltzmann equation*, J. Math. Pures Appl., 54 (1975), pp. 125–156.

[8] D. L. Ermak and J. A. McCammon, *Brownian dynamics with hydrodynamic interactions*, J. Chem. Phys., 69 (1978), pp. 1352–1360.

[9] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley & Sons, New York, 1986.

[10] G. Falkovich, K. Gawędzki, and M. Vergassola, *Particles and fields in fluid turbulence*, Rev. Modern Phys., 73 (2001), pp. 913–975.

[11] A. Fannjiang and T. Komorowski, *Diffusive and nondiffusive limits of transport in nonmixing flows*, SIAM J. Appl. Math., 62 (2002), pp. 909–923.

[12] The Faraday Division, *Concentrated Colloidal Dispersions*, Faraday Discussions of the Chemical Society 76, Royal Society of Chemistry, London, 1983.

[13] G. B. Folland, *Introduction to Partial Differential Equations*, 2nd ed., Princeton University Press, Princeton, NJ, 1995.

[14] R. F. Fox and G. E. Uhlenbeck, *Contributions to non-equilibrium thermodynamics.* I. *Theory of hydrodynamical fluctuations*, Phys. Fluids, 13 (1970), pp. 1893–1902.
[15] J. Fritz, *On the diffusive nature of entropy flow in infinite systems: Remarks to a paper "Nonlinear diffusion limit for a system with nearest neighbor interactions"* [*Comm. Math. Phys.* 118 (1988), *no.* 1, 31–59; *MR* 89*m*:60255] *by M. Z. Guo, G. C. Papanicolaou, and S. R. S. Varadhan*, Comm. Math. Phys., 133 (1990), pp. 331–352.
[16] G. Gallavotti, *Statistical Mechanics*, Springer-Verlag, Berlin, 1999.
[17] C. W. Gardiner, *Handbook of Stochastic Methods*, Springer Ser. Synergetics 13, 2nd ed., Springer-Verlag, Berlin, 1985.
[18] M. Z. Guo, G. C. Papanicolaou, and S. R. S. Varadhan, *Nonlinear diffusion limit for a system with nearest neighbor interactions*, Comm. Math. Phys., 118 (1988), pp. 31–59.
[19] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics with Special Applications to Particulate Media*, Prentice–Hall, Englewood Cliffs, NJ, 1965.
[20] R. Hersh, *Random evolutions: A survey of results and problems*, Rocky Mountain J. Math., 4 (1974), pp. 443–477.
[21] R. Hersh and G. Papanicolaou, *Non-commuting random evolutions, and an operator-valued Feynman–Kac formula*, Comm. Pure Appl. Math., 25 (1972), pp. 337–367.
[22] S. Karlin and H. M. Taylor, *A Second Course in Stochastic Processes*, Academic Press, Boston, 1981.
[23] R. Z. Khas'minskii, *Principle of averaging for parabolic and elliptic differential equations and for Markov processes with small diffusion*, Theor. Probability Appl., 8 (1963), pp. 1–21.
[24] R. Z. Khas'minskii, *A limit theorem for the solutions of differential equations with random right-hand sides*, Theor. Probability Appl., 11 (1966), pp. 390–406.
[25] T. W. Körner, *Fourier Analysis*, Cambridge University Press, Cambridge, UK, 1988.
[26] P. R. Kramer and A. J. Majda, *Stochastic mode reduction for particle-based simulation methods for complex microfluid systems*, SIAM J. Appl. Math., 64 (2003), pp. 401–422.
[27] P. R. Kramer and C. S. Peskin, *An Extension of the Immersed Boundary Method Including Thermal Fluctuations*, manuscript, 2003.
[28] P. R. Kramer and C. S. Peskin, *Incorporating thermal fluctuations into the immersed boundary method*, in the Proceedings of the Second MIT Conference on Computational Fluid and Solid Mechanics, Cambridge, MA, 2003, K. J. Bathe, ed., Elsevier Science, Oxford, UK, 2003, pp. 1755–1758.
[29] R. Kubo, *Stochastic Liouville equations*, J. Math. Phys., 4 (1963), pp. 174–183.
[30] T. G. Kurtz, *A limit theorem for perturbed operator semigroups with applications to random evolutions*, J. Funct. Anal., 12 (1973), pp. 55–67.
[31] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics. Vol.* 9: *Statistical Physics*, Pergamon Press, Oxford, UK, 1980.
[32] M. Lesieur, *Turbulence in Fluids*, Fluid Mech. Appl. 1, 2nd ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.
[33] A. J. Majda and P. R. Kramer, *Simplified models for turbulent diffusion: Theory, numerical modelling and physical phenomena*, Phys. Rep., 314 (1999), pp. 237–574.
[34] A. Majda, I. Timofeyev, and E. Vanden-Eijnden, *A priori tests of a stochastic mode reduction strategy*, Phys. D, 170 (2002), pp. 206–252.
[35] A. J. Majda, I. Timofeyev, and E. Vanden-Eijnden, *Models for stochastic climate prediction*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 14687–14691.
[36] A. J. Majda, I. Timofeyev, and E. Vanden-Eijnden, *A mathematical framework for stochastic climate models*, Comm. Pure Appl. Math., 54 (2001), pp. 891–974.
[37] B. Øksendal, *Stochastic Differential Equations*, 5th ed., Springer-Verlag, Berlin, 1998.
[38] S. Olla, S. R. S. Varadhan, and H.-T. Yau, *Hydrodynamical limit for a Hamiltonian system with weak noise*, Comm. Math. Phys., 155 (1993), pp. 523–560.
[39] G. C. Papanicolaou, *Some probabilistic problems and methods in singular perturbations*, Rocky Mountain J. Math., 6 (1976), pp. 653–674.
[40] G. C. Papanicolaou and W. Kohler, *Asymptotic theory of mixing stochastic ordinary differential equations*, Comm. Pure Appl. Math., 27 (1974), pp. 641–668.
[41] C. S. Peskin, *private communication.*
[42] C. S. Peskin, *The immersed boundary method*, Acta Numer., 11 (2002), pp. 479–517.
[43] C. S. Peskin and D. M. McQueen, *A general method for the computer simulation of biological systems interacting with fluids*, in Biological Fluid Dynamics, C. P. Ellington and T. J. Pedley, eds., The Company of Biologists Limited, Cambridge, UK, 1995, pp. 265–276.
[44] P. N. Pusey and R. J. A. Tough, *Hydrodynamic interactions and diffusion in concentrated particle suspensions*, in Concentrated Colloidal Suspensions, Faraday Discussions of the Chemical Society 76, Royal Society of Chemistry, London, 1983, pp. 123–136.
[45] J. M. Rallison and E. J. Hinch, *The effect of particle interactions on dynamic light scattering*

from a dilute suspension, J. Fluid Mech., 167 (1986), pp. 131–168.
[46] F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw–Hill, New York, 1965.
[47] J. Rotne and S. Prager, *Variational treatment of hydrodynamic interaction in polymers*, J. Chem. Phys., 50 (1969), pp. 4831–4837.
[48] A. Sierou and J. F. Brady, *Accelerated Stokesian dynamics simulations*, J. Fluid Mech., 448 (2001), pp. 115–146.
[49] Y. G. Sinai, *Probability Theory*, Springer-Verlag, Berlin, 1992.
[50] H. Spohn, *Large Scale Dynamics of Interacting Particles*, Springer-Verlag, Berlin, 1991.
[51] D. J. Tritton, *Physical Fluid Dynamics*, 2nd ed., Clarendon Press, Oxford, UK, 1988.
[52] M. Vlad, F. Spineanu, J. H. Misguich, and R. Balescu, *Collisional effects on diffusion scaling laws in electrostatic turbulence*, Phys. Rev. E, 61 (2000), pp. 3023–3032.
[53] A. M. Yaglom, *Correlation Theory of Stationary and Related Random Functions. Volume* I: *Basic Results*, Springer-Verlag, Berlin, 1987.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 401–422

STOCHASTIC MODE REDUCTION FOR PARTICLE-BASED SIMULATION METHODS FOR COMPLEX MICROFLUID SYSTEMS*

PETER R. KRAMER† AND ANDREW J. MAJDA‡

Abstract. We illustrate the stochastic mode reduction procedure as formulated recently by Majda, Timofeyev, and Vanden-Eijnden [*Comm. Pure Appl. Math.*, 54 (2001), pp. 891–974] (MTV) on the equations of motion underlying various particle-based simulation approaches (such as Stokesian dynamics and Brownian dynamics) and the conceptually distinct dissipative particle dynamics (DPD) simulation approaches for complex microfluid systems. The resulting coarse-grained dynamics are compared and contrasted with each other. We show that the stochastic mode reduction procedure provides a way to recover the Smoluchowski dynamics for a standard model of multiple interacting particles in a fluid. The DPD, however, has some subtle aspects which obstruct the application of the stochastic mode reduction procedure. We discuss the mathematical and physical properties of the DPD method that underlie this difficulty.

Key words. Brownian dynamics, Brownian motion, dissipative particle dynamics, Smoluchowski reduction

AMS subject classifications. 60H10, 60H30, 60J60, 60J65, 60J70, 76R50, 82C31, 82C70, 82C80

DOI. 10.1137/S0036139903422140

1. Introduction.

In [24], we showed how the stochastic mode reduction framework of Majda, Timofeyev, and Vanden-Eijnden [31, 32, 33] (with MTV hereafter referring to [33]) could be used to assist in the design and analysis of the immersed boundary (IB) method [26, 25, 40] for numerically simulating microphysiological systems consisting of various elastic structures and particles immersed in a fluid with thermal fluctuations. The procedure exploited the smallness of the thermal Reynolds number, which implied a separation of time scales for the dynamics of the fluid and the immersed structures, and derived a simplified stochastic system for the immersed structures with the fluid variables eliminated using rigorous singular perturbation techniques [14, 18, 29, 39]. It is natural to explore how this approach works on other numerical simulation techniques for complex fluid systems consisting of immersed structures and thermal fluctuations. We study here the application of the stochastic mode reduction framework of MTV to the equations underlying particle-based (PB) dynamics schemes such as Brownian dynamics [10], Stokesian dynamics [6, 46], and the conceptually distinct dissipative particle dynamics [1, 11, 13, 23, 35, 38]. These numerical approaches differ fundamentally from the IB method in that they do not treat the fluid dynamically, but rather immediately model its effective influence on the particles. The detailed equations of motion for each of the simulation methods involve some nontrivial approximations or assumptions, even before any numerical discretization is contemplated.

As we shall discuss briefly in the relevant sections, even though the equations

*Received by the editors January 30, 2002; accepted for publication (in revised form) June 11, 2003; published electronically December 31, 2003.
http://www.siam.org/journals/siap/64-2/42214.html
†Department of Mathematical Sciences, Rensselaer Polytechnic Institute, 301 Amos Eaton Hall, 110 8th Street, Troy, NY 12180 (kramep@rpi.edu).
‡Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (jonjon@cims.nyu.edu). This author's research was supported in part by ARO grant DAAD19-01-10810, NSF grant DMS9972865, and ONR grant N00014-96-1-0043.

and variables underlying these simulation approaches are rather different from each other and the IB method, they each can be treated within the MTV framework to deduce a simplified system governing the immersed particles and structures when the viscous friction with the fluid is sufficiently strong. The notion of Reynolds number is a bit obscure for the methods discussed in the present paper, since viscosity does not enter as a governing parameter. Instead, we take the small parameter in the asymptotic analysis as a certain "thermal Kubo number," which is defined as the ratio of the rate of decorrelation of a particle's (Lagrangian) thermal velocity due to its advection into different fluid regions relative to its rate of decorrelation due to viscous damping. The limit of small thermal Kubo number seems appropriate for many microfluid applications and plays a universal role in preparing for a stochastic mode reduction procedure for the methods described in the present paper as well as the IB method discussed in [24].

For the IB method, a small thermal Kubo number implies a separation between the fast time scales of the fluid and the slow time scales of the immersed structures. For the methods treated in the present paper, the fluid is not treated dynamically. Instead, the system variables consist of the positions and momenta of the elementary components of the immersed structures. Stochastic forcing modeling thermal fluctuations is naturally applied to the momenta of the particles. A low thermal Kubo number implies that the time scale for the momentum variables is much faster than that of the position variables. The stochastic mode reduction procedure of MTV then suggests a simplification of the system through elimination of the momentum variables and the development of a new set of effective "coarse-grained" equations for the position variables. By coarse-grained dynamics for the particles in these simulation schemes we mean a simplified description involving only the particle positions and not the rapidly fluctuating velocities. The derivations of the simplified coarse-grained dynamics can be accomplished rigorously for the classical equations underlying PB methods such as Brownian dynamics and Stokesian dynamics, but the dissipative particle dynamics (DPD) scheme has some aspects which can obstruct the deduction of simplified effective dynamics.

The application of the MTV procedure to eliminate the fast momentum variables from the classical PB equations (section 2) yields the standard Smoluchowski (or Brownian dynamics) limit for the positions of the immersed structures, which have been computed previously with several other approaches [10, 37, 47, 49, 50]. The variation of the Chapman–Enskog procedure used in [47] has some foundational similarities to the method of derivation we present, but the present approach is based on a rigorous theorem [29], whereas the validity of the Chapman–Enskog expansion is less certain [45, 47]. Moreover, we calculate the differential operators associated with the simplified dynamics in a different way, involving a slight extension of the framework developed in MTV to a context in which the noise driving the fast modes is not diagonal.

A similar calculation can be made to derive effective equations for the elementary particles in a DPD system (section 3). These resulting equations look formally very similar to those arising from the classical PB equations. Due to the finite-ranged nature of the frictional interactions, however, degeneracies emerge which can spoil the MTV stochastic mode reduction procedure. The formal calculation appears to apply only for some fraction of realizations of the dynamics which increases with increasing density. We cannot yet, however, make this into a rigorous statement. It would be interesting to extend the rigorous aspect of the MTV framework to cover systems with degeneracies such as DPD, but we will not attempt this here.

2. Application of stochastic mode reduction for PB simulation methods. Many scientists and engineers wishing to simulate microscopic fluid-particle systems with thermal fluctuations adopt a PB perspective in contrast to the fluid-based perspective of the IB method. In the PB setup, one appeals to the low Reynolds number of the system to eliminate the fluid degrees of freedom by assuming that the fluid is always in a quasi-steady state determined by the time-independent Stokes equations responding to whatever forces the particles happen to be exerting at that moment [6, 10, 20]. The fluid then ceases to be an independent dynamical quantity. The state of the system can be completely described by the positions $\mathbf{X} = \{\mathbf{X}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ and velocities $\mathbf{V} = \{\mathbf{V}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ of each immersed particle. These variables are updated according to Newton's laws, incorporating three kinds of forces:

(i) interparticle forces governed by some general potential of the particle configuration $\Phi(\mathbf{X})$,

(ii) hydrodynamic drag forces which are determined by the momentary positions and velocities of the particles, and

(iii) thermally fluctuating forces.

We provide details in section 2.1. Note that the approximation that the fluid evolves according to the quasi-steady Stokes equations requires stronger assumptions than simply noting that the Reynolds number is small [8], and these stronger assumptions are not necessarily met in practice [25].

While the PB simulation approaches are founded on these Newton law evolution equations for $\mathbf{X}$ and $\mathbf{V}$, these equations are not suitable for numerical simulation in many situations where the particle velocities decorrelate on a much faster time scale than the time over which the particle position variables change significantly. The equations of motion in this case are very stiff, but the disparity in time scales also permits a simplification of the system to one involving only the slow variables (particle positions). For the present example, this is known as the reduction of the full Fokker–Planck equation (involving position and momentum variables) to the Smoluchowski equation (involving only position variables). This procedure has been undertaken in various ways (formal asymptotic arguments [27], multiple time scale expansions [3], Chapman–Enskog expansion [47], projection operators [37, 49], asymptotic moment equation expansions [50], and asymptotic integration of stochastic differential equations [10, 22]). We show here how this reduction can be performed within the MTV stochastic mode reduction framework. While the ideas behind the calculations are related to those in the earlier works (particularly the adaptation of the Chapman–Enskog expansion approach in [47] and the multiple time scale derivations for the case of a single particle in [3, 49]), the MTV framework has a distinct organizational and computational structure which is systematic, fairly transparent, and founded on a rigorous theorem [29].

2.1. Formulation. We now write the equations on which the PB methods are based, with particle position and momentum resolved. For simplicity in exposition, we assume that all particles have the same mass m, though similar procedures can be applied to the more general case, as we shall discuss in section 2.5.

$$
\begin{aligned}
d\mathbf{X}_{\boldsymbol{\alpha}}(t) &= \mathbf{V}_{\boldsymbol{\alpha}}(t)\,dt,\\
m\,d\mathbf{V}_{\boldsymbol{\alpha}}(t) &= -\sum_{\boldsymbol{\alpha}'\in\mathcal{A}} \mathcal{R}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}(\mathbf{X}(t))\mathbf{V}_{\boldsymbol{\alpha}'}(t)\,dt - \nabla_{\boldsymbol{\alpha}}\Phi(\mathbf{X}(t))\,dt\\
&\quad + \sqrt{2k_BT}\sum_{\boldsymbol{\alpha}'\in\mathcal{A}} \mathcal{S}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}(\mathbf{X}(t))\,d\mathbf{W}_{\boldsymbol{\alpha}'}.
\end{aligned}
\tag{2.1}
$$

These equations, as previously stated, are nothing more than Newton's second law, with the right-hand side of the second equation in (2.1) describing the three kinds of forces that are assumed to act on the particles. The first term is the model for the force felt by the particles due to the fluid, under the governing assumption that the fluid is always in a quasi-steady state responding to the currently applied forces and positions of the particles. The tensor $\mathcal{R}$ is the hydrodynamic resistance matrix [6, 20], which is symmetric, positive definite, and depends on the momentary particle configuration $\mathbf{X}$. Its entries $\mathcal{R}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}$ indicate how the motion of particle $\boldsymbol{\alpha}'$ induces a hydrodynamic drag force on particle $\boldsymbol{\alpha}$. The second term in the second equation in (2.1) is just the force felt by a particle due to its direct interaction with other particles via the generalized potential Φ, just as in the IB method, but it is now permissible for Φ to include external forces which do not conserve total momentum. The third term describes the random thermal fluctuations, with strength specified by the matrix $\mathcal{S}(t)$, which is (any) square root of the symmetric, positive definite matrix $\mathcal{R}(t)$:

$$\mathcal{R}(\mathbf{X}) = \mathcal{S}(\mathbf{X})\mathcal{S}^{\dagger}(\mathbf{X}).$$

The $\{\mathbf{W}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ are a collection of independent standard real-valued Brownian motions, which are mean zero Gaussian processes with stationary increments satisfying

$$\langle d\mathbf{W}(t) \otimes d\mathbf{W}(t')\rangle = \mathcal{I}\delta(t-t')\,dt\,dt'. \tag{2.2}$$

Note that the Brownian motion $\mathbf{W}_{\boldsymbol{\alpha}}$ does not describe the Brownian motion of the particle with label $\boldsymbol{\alpha}$. Indeed, the Brownian motions of the particles are actually coupled, while $\{\mathbf{W}_{\boldsymbol{\alpha}}(t)\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ are independent. One may think of the $\{\mathbf{W}_{\boldsymbol{\alpha}}(t)\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ as a diagonalization of the thermal noise driving the particles. If the particles are well separated, then $\mathbf{W}_{\boldsymbol{\alpha}}(t)$ gives a leading order *approximation* to the thermal Brownian motion of particle $\boldsymbol{\alpha}$.

The equations of motion are supplemented by the initial conditions

$$\mathbf{X}_{\boldsymbol{\alpha}}(t=0) = \mathbf{X}_{0,\boldsymbol{\alpha}}, \qquad \mathbf{V}_{\boldsymbol{\alpha}}(t=0) = \mathbf{V}_{0,\boldsymbol{\alpha}}.$$

In Stokesian dynamics [6, 46], an imposed background spatially linear flow $\mathbf{u}^{\infty}(\mathbf{x})$ is admitted, and can be incorporated into our framework by generalizing the resistance force term in (2.1) to

$$-\sum_{\boldsymbol{\alpha}'\in\mathcal{A}} \mathcal{R}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}(\mathbf{X}(t)) \cdot (\mathbf{V}_{\boldsymbol{\alpha}'}(t) - \mathbf{u}^{\infty}(\mathbf{X}_{\boldsymbol{\alpha}'}(t))) + (\mathcal{R}^{E}(\mathbf{X}(t)){:}\mathcal{E}^{\infty})_{\boldsymbol{\alpha}}.$$

Here $\mathcal{R}^{E}(\mathbf{X})$ is another resistance tensor (with one particle and two spatial indices) relating to the imposed (constant) strain $\mathcal{E}^{\infty} = \frac{1}{2}(\nabla\mathbf{u}^{\infty} + (\nabla\mathbf{u}^{\infty})^{\dagger})$. We see that this generalization is mathematically equivalent to adding a (nonconservative) force

$$-\sum_{\boldsymbol{\alpha}'\in\mathcal{A}} \mathcal{R}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}(\mathbf{X}(t)) \cdot \mathbf{u}^{\infty}(\mathbf{X}_{\boldsymbol{\alpha}'}(t)) + (\mathcal{R}^{E}(\mathbf{X}(t))\mathcal{E}^{\infty})_{\boldsymbol{\alpha}},$$

which can be handled in the same way as the interparticle force term. To avoid extra burden in the calculations and nondimensionalization, we therefore proceed with no imposed flow $\mathbf{u}^{\infty} \equiv 0$ but comment in subsection 2.3 on how it could be incorporated.

2.2. Nondimensionalization.

2.2.1. Parameters of externally specified functions. We nondimensionalize the potential for the interacting particles by

$$\Phi(\mathbf{X}) = F\ell_f\tilde{\Phi}(\mathbf{X}/\ell_f),$$

where F is a typical force amplitude and ℓ_f is a length scale characterizing the range of the interparticle forces. Note that we are now using a force magnitude rather than a force density magnitude as we did for the IB method in [24]; this is due to the shift from a fluid-based to PB perspective. We let ℓ_0 be a length scale characterizing the typical initial separation between particles, and V_0 a typical magnitude of the initial particle velocities, and write

$$\mathbf{X}_{0,\boldsymbol{\alpha}} = \ell_0 \tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}}, \qquad \mathbf{V}_{0,\boldsymbol{\alpha}} = V_0 \tilde{\mathbf{V}}_{0,\boldsymbol{\alpha}},$$

where $\tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}}$ and $\tilde{\mathbf{V}}_{0,\boldsymbol{\alpha}}$ are nondimensionalized initial data which are to be thought of (at least formally) as order unity.

The resistance tensor $\mathcal{R}$ depends both on the current configuration $\mathbf{X}(t)$ of the particles and on their sizes and shapes. For now, we will assume that all particles are identical, with maximal diameter a and hydrodynamic drag friction γ. In the dilute limit (where the particle separation scale is much larger than the particle size a), we would have $\mathcal{R} = \gamma\mathcal{I}$, where $\mathcal{I}$ is the identity matrix, but the resistance tensor in a nondilute solution will have both diagonal and off-diagonal modifications due to hydrodynamic coupling between different particles. We can express the resistance tensor in terms of a nondimensional function $\tilde{\mathcal{R}}$ (depending on the shape of the particles) [20] which has order unity variations (i.e., is properly normalized):

$$\mathcal{R}(\mathbf{X}) = \gamma\tilde{\mathcal{R}}(\mathbf{X}/a).$$

The amplitude of the thermal forces is similarly expressed in terms of a nondimensional normalized function:

$$\mathcal{S}(\mathbf{X}) = \sqrt{\gamma}\tilde{\mathcal{S}}(\mathbf{X}/a),$$

with $\tilde{\mathcal{S}}\tilde{\mathcal{S}}^\dagger = \tilde{\mathcal{R}}$. We will assume that $\mathcal{R}$ is appropriately mollified to avoid singularities as two or more particles coalesce, as must be done in any numerical implementation.

2.2.2. Reference units. We use the following:

(i) length scale $\ell_T = a$,
(ii) time scale $\tau_T = \sqrt{ma^2/k_BT}$,
(iii) mass scale $m_T = m$.

These units are chosen with motivation similar to those used to nondimensionalize the IB equations in [24]. We have simply chosen $\ell_T = a$, both because physiological systems are typically composed of extended structures such as polymers and membranes which will be modeled by stringing together particles separated by distances comparable to their sizes and because it matches most closely the nondimensionalization choices made for other simulation methods discussed in this paper. The characteristic velocity is again the order of magnitude of the thermal velocity of a particle:

$$V_T = \frac{\ell_T}{\tau_T} = \sqrt{\frac{k_BT}{m}}.$$

2.2.3. Nondimensional groups.

(i) The thermal Kubo number

$$\mathrm{Ku_T} = \frac{\sqrt{mk_BT}}{\gamma a},$$

which describes the ratio of the time scale of frictional decorrelation m/γ to the time scale ℓ_T/V_T over which the particles would move across a spatial structural length $\ell_T = a$. Note that if the fluid density ρ is comparable to the particle mass density m/a^3, we can also think of the thermal Kubo number as a thermal particle Reynolds number $\ell_T V_T/\nu$, because the self-friction constant scales as $\gamma \sim \rho\nu a$. We note, though, that the equations for PB dynamics are most accurate in the limit in which the particle mass density is much greater than that of the fluid [8, 21, 30, 36, 44, 48].

(ii) Nondimensionalized measures of the effects of structural forces and initial velocity,

$$\phi = \frac{Fa}{k_B T}, \qquad \Upsilon = \frac{V_0}{V_T}.$$

(iii) The length scale ratios

$$\tilde{\ell}_f = \frac{\ell_f}{a}, \qquad \tilde{\ell}_0 = \frac{\ell_0}{a}.$$

2.2.4. Nondimensionalized PB dynamics. Nondimensionalizing with respect to the reference units described in section 2.2.2 and denoting nondimensional variables (but not externally prescribed functions) by the same notation as for their dimensional form, we obtain

$$\begin{aligned} d\mathbf{X}_{\boldsymbol{\alpha}}(t) &= \mathbf{V}_{\boldsymbol{\alpha}}(t)\, dt, \\ d\mathbf{V}_{\boldsymbol{\alpha}}(t) &= -\phi \nabla_{\boldsymbol{\alpha}} \tilde{\Phi}(\mathbf{X}(t)/\tilde{\ell}_f)\, dt - \sum_{\boldsymbol{\alpha}' \in \mathcal{A}} \mathrm{Ku}_{\mathrm{T}}^{-1} \tilde{\mathcal{R}}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}(\mathbf{X}(t)) \mathbf{V}_{\boldsymbol{\alpha}'}(t)\, dt \\ &\quad + \sqrt{2} \mathrm{Ku}_{\mathrm{T}}^{-1/2} \tilde{\mathcal{S}}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}(\mathbf{X}(t))\, d\mathbf{W}_{\boldsymbol{\alpha}'}. \end{aligned}$$

The nondimensionalized initial conditions are

$$\mathbf{X}_{\boldsymbol{\alpha}}(t=0) = \tilde{\ell}_0 \tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}}, \qquad \mathbf{V}_{\boldsymbol{\alpha}}(t=0) = \Upsilon \tilde{\mathbf{V}}_{0,\boldsymbol{\alpha}}.$$

2.3. Effective dynamics for PB dynamics at low thermal Kubo number. Elimination of the fast momentum modes yields the following description of the evolution of the positions of the particles.

PROPOSITION 2.1 (PB dynamics at small Kubo number). *In the limit* $\mathrm{Ku_T} \to 0$ *with all other nondimensional quantities held fixed, the solution for the particle positions* $\{\mathbf{X}_{\boldsymbol{\alpha}}(t)\}_{\boldsymbol{\alpha}}$ *obtained from the complete system* (2.1) *and rescaled in time as*

$$\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t) = \lim_{\mathrm{Ku_T} \to 0} \mathbf{X}_{\boldsymbol{\alpha}}(t/\mathrm{Ku_T})$$

converges in law to the solution of the following simplified closed stochastic differential system involving only the particle positions $\{\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t)\}$:

$$\begin{aligned} d\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t) &= \bar{\mathbf{V}}_{\boldsymbol{\alpha}}(\bar{\mathbf{X}}(t))\, dt + \sum_{\boldsymbol{\alpha}' \in \mathcal{A}} \mathcal{S}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(\bar{\mathbf{X}}(t))\, d\mathbf{W}_{\boldsymbol{\alpha}'}(t), \\ \bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t=0) &= \tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}}, \end{aligned} \tag{2.3}$$

where the stochastic real white noise terms $d\mathbf{W}_{\boldsymbol{\alpha}}(t)$ *are defined near* (2.2) *and are given the Itô interpretation. The explicit expression for the drift term is*

$$\bar{\mathbf{V}}_{\boldsymbol{\alpha}}(\mathbf{X}) = \sum_{\boldsymbol{\alpha}' \in \mathcal{A}} \left[-((\tilde{\mathcal{R}}(\mathbf{X}))^{-1})_{\boldsymbol{\alpha},\boldsymbol{\alpha}'} \cdot \nabla_{\boldsymbol{\alpha}'} \tilde{\Phi}(\mathbf{X}/\tilde{\ell}_f) + \nabla_{\boldsymbol{\alpha}'} \cdot (\tilde{\mathcal{R}}^{-1}(\mathbf{X}))_{\boldsymbol{\alpha}',\boldsymbol{\alpha}} \right], \tag{2.4}$$

and the matrix coefficients of the stochastic terms are

$$\mathcal{S}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(\mathbf{X}) = \sqrt{2}(\tilde{\mathcal{R}}^{-1/2}(\mathbf{X}))_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}.$$

The drift term here includes not only the motion of the particles responding to the interparticle forces under the quasi-steady assumption for the fluid state, but a "divergence-drift" term, which appears due to the spatial dependence of $\tilde{\mathcal{R}}$ and is present even in the absence of forces on the particles [10, 19, 41]. The clearest way to understand the divergence-drift is by noting that the diffusion correlation tensor

$$\mathcal{D}(\mathbf{r}) \equiv \frac{1}{2}\frac{d}{dt}\left\langle (\mathbf{X}_{\boldsymbol{\alpha}}(t) - \mathbf{X}_{\boldsymbol{\alpha}}(t')) \otimes (\mathbf{X}_{\boldsymbol{\alpha}'}(t) - \mathbf{X}_{\boldsymbol{\alpha}'}(t')) \,|\mathbf{X}_{\boldsymbol{\alpha}}(t') = \mathbf{x} + \mathbf{r}, \mathbf{X}_{\boldsymbol{\alpha}'}(t') = \mathbf{x}\right\rangle_{t=t'} \tag{2.5}$$

is, for the nondimensionalized PB system, exactly $\mathcal{D}(\mathbf{X}) = (\tilde{\mathcal{R}}(\mathbf{X}))^{-1}$, so that the divergence-drift term is $\nabla \cdot \mathcal{D}$. Roughly speaking, the divergence-drift arises because variable diffusivity induces a bias to the mean particle motion due to the asymmetric strength of fluctuations in opposite directions. This term is important only when two particles happen to be close together relative to their sizes [15, pp. 232–233], and is omitted in simulations based only on Oseen or Rotne–Prager approximations to the resistance tensor [2, 9, 10, 17, 28, 43].

The coarse-grained PB dynamics respect the Einstein relation between mobility and diffusion correlations. Unlike the IB method, it is permissible for the potential in the PB dynamics to include external forces (so that total momentum need not be conserved), and the conclusions of Proposition 2.1 are unchanged.

If a steady background flow $\mathbf{u}^\infty(\mathbf{x})$ is imposed which is not too strong relative to the thermal fluctuations, the only change in the effective dynamics is the addition of the nondimensionalized form of the terms $\mathbf{u}^\infty(\mathbf{X}_{\boldsymbol{\alpha}}(t)) + (\mathcal{R}^{-1} \cdot \mathcal{R}^E{:}\mathcal{E}^\infty)_{\boldsymbol{\alpha}}$ to the right-hand side of the expression (2.4) for the effective drift, in agreement with how these effects are handled within Stokesian dynamics [6].

2.4. Stochastic mode reduction for PB dynamics. We now derive Proposition 2.1, following the general prescription from MTV, which we presented in some detail in [24]. We focus here only on those aspects of the calculation that are particular to the PB equations; the basic notation and formalism are the same as in [24].

2.4.1. Kolmogorov backward equation formalism. We identify $\varepsilon = \mathrm{Ku_T}$ as the small parameter and rescale to a longer time $t \to t/\varepsilon$. We denote the system variables in terms of the collection of particle positions $\mathbf{X}$ and velocities $\mathbf{V}$. Then the Kolmogorov backward equation under this time rescaling can be written as

$$\begin{aligned} -\frac{\partial \rho^\varepsilon(s, \mathbf{X}, \mathbf{V}|t)}{\partial s} &= \varepsilon^{-2}\mathcal{L}_1\rho^\varepsilon + \varepsilon^{-1}\mathcal{L}_2\rho^\varepsilon, \\ \rho^\varepsilon(s = t, \mathbf{X}, \mathbf{V}|t) &= f(\mathbf{X}, \mathbf{V}), \end{aligned}$$

with differential operators

$$\begin{aligned} \mathcal{L}_1 &= \sum_{\boldsymbol{\alpha},\boldsymbol{\alpha}' \in \mathcal{A}} -\mathbf{V}_{\boldsymbol{\alpha}} \cdot \tilde{\mathcal{R}}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}(\mathbf{X}) \cdot \frac{\partial}{\partial \mathbf{V}_{\boldsymbol{\alpha}'}} + \frac{\partial}{\partial \mathbf{V}_{\boldsymbol{\alpha}}} \cdot \tilde{\mathcal{R}}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}(\mathbf{X}) \cdot \frac{\partial}{\partial \mathbf{V}_{\boldsymbol{\alpha}'}}, \\ \mathcal{L}_2 &= -\sum_{\boldsymbol{\alpha} \in \mathcal{A}} \phi \nabla_{\boldsymbol{\alpha}} \tilde{\Phi}\left(\frac{\mathbf{X}(t)}{\tilde{\ell}_f}\right) \cdot \frac{\partial}{\partial \mathbf{V}_{\boldsymbol{\alpha}}} + \mathbf{V}_{\boldsymbol{\alpha}} \cdot \frac{\partial}{\partial \mathbf{X}_{\boldsymbol{\alpha}}}. \end{aligned}$$

2.4.2. Asymptotic expansion of solution. We assume that f depends only on the slow variables $f = f(\mathbf{X})$ to avoid consideration of irrelevant initial transients. The solvability conditions for the equations in the asymptotic hierarchy are obtained by integrating against the invariant measure π_{OU} associated with the operator $\mathcal{L}_1$, which here is

$$\pi_{OU}(\mathbf{V}) = \left(\prod_{\boldsymbol{\alpha}\in\mathcal{A}} \frac{1}{2\pi}\right) \exp\left(-\frac{1}{2}\sum_{\boldsymbol{\alpha}\in\mathcal{A}} |\mathbf{V}_{\boldsymbol{\alpha}}|^2\right).$$

The first two equations in the asymptotic hierarchy are solved in the same way as in [24], and the solvability condition resulting from the third equation is again

$$\begin{aligned} -\frac{\partial \rho_0}{\partial s} &= -\mathbf{E}_{OU}\mathcal{L}_2\mathcal{L}_1^{-1}\mathcal{L}_2\mathbf{E}_{OU}\rho_0, \\ \rho_0(s=t,\mathbf{X}|t) &= f(\mathbf{X}); \end{aligned} \tag{2.6}$$

thus the main task is to compute explicitly

$$\bar{\mathcal{L}} \equiv -\mathbf{E}_{OU}\mathcal{L}_2\mathcal{L}_1^{-1}\mathcal{L}_2\mathbf{E}_{OU}.$$

2.4.3. Explicit computation of limiting PDE. Thus far, our derivation largely coincides with the expansion procedure in [47]. However, instead of diagonalizing $\mathcal{L}_1$ in terms of Hermite polynomials, we follow in the spirit of the procedure developed in Appendix B of MTV. We show how this approach can be modified to directly treat a case such as the present one, where $\mathcal{L}_1$ is not diagonal in the fast variables $\mathbf{V}$.

We pass again to fast Fourier variables $\mathbf{P} = \{\mathbf{p}_{\boldsymbol{\alpha}}\}$ defined through

$$\hat{g}(\mathbf{P}) = \int_{\mathbb{R}^N} \exp\left[i\sum_{\boldsymbol{\alpha}\in\mathcal{A}} \mathbf{p}_{\boldsymbol{\alpha}}\cdot\mathbf{V}_{\boldsymbol{\alpha}}\right] g(\mathbf{V})\, d\mathbf{V}.$$

The differential operators act on functions of $\mathbf{P}$ and $\mathbf{X}$ as follows:

$$\begin{aligned} \hat{\mathcal{L}}_1 &= \frac{\partial}{\partial \mathbf{P}}\cdot\tilde{\mathcal{R}}(\mathbf{X})\cdot\mathbf{P} - \mathbf{P}\cdot\tilde{\mathcal{R}}(\mathbf{X})\cdot\mathbf{P}, \\ \hat{\mathcal{L}}_2 &= i\phi\mathbf{P}\cdot\nabla_{\mathbf{X}}\tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right) - i\frac{\partial}{\partial\mathbf{P}}\cdot\frac{\partial}{\partial\mathbf{X}}. \end{aligned}$$

We have written these operators in a "grand matrix" form, which will facilitate later computations. We just think of quantities like $\mathbf{X}$ and $\mathbf{P}$ as vectors with $3N$ components, and matrices like $\tilde{\mathcal{R}}$ as $3N\times 3N$ matrices. Scalar products between vectors (and vector operators) then involve summing over particle labels in addition to the Cartesian dimensions.

We now need to compute $\hat{\mathcal{L}}_1^{-1}$ and will have to generalize Lemma A.2 in MTV to allow for the nondiagonal structure of $\hat{\mathcal{L}}_1$. Since a similar result will be needed for the DPD calculation below, we present here a lemma sufficient for both calculations.

LEMMA 2.2. *Suppose we are given a first order differential operator of the form*

$$\hat{\mathcal{L}}_1 = \frac{\partial}{\partial\mathbf{P}}\cdot\mathcal{A}\cdot\mathbf{P} - \mathbf{P}\cdot\mathcal{A}\cdot\mathbf{P},$$

where $\mathcal{A}$ is a positive definite symmetric matrix. Then, given any function $\hat{g}(\mathbf{P})$ satisfying $\hat{\mathcal{L}}_1^\dagger \hat{g} = 0$, $\hat{g}$ has an inverse image of $\hat{\mathcal{L}}_1$ of the following form:

$$(\hat{\mathcal{L}}_1^{-1}\hat{g})(\mathbf{P}) = -\int_0^\infty \exp\left(\operatorname{Tr}\mathcal{A}t - \frac{1}{2}\mathbf{P}\cdot(\exp(\mathcal{A}t)\cdot\exp(\mathcal{A}t) - \mathcal{I})\cdot\mathbf{P}\right)\hat{g}(\boldsymbol{\beta}(t))\,dt,$$

where

$$\boldsymbol{\beta}(t) = \exp(\mathcal{A}t)\cdot\mathbf{P}.$$

Note that $\hat{\mathcal{L}}_1$ has a one-dimensional kernel (consisting of constant functions), and thus the inverse of $\hat{\mathcal{L}}_1$ is not uniquely defined. However, all choices of inverse images under $\hat{\mathcal{L}}_1$ differ only by functions of $\mathbf{X}$, which are annihilated in the computation of $\bar{\mathcal{L}}$. Therefore, any choice of inverse image will suffice. Note also that Lemma 2.2 is indeed a generalization of Lemma A.2 in MTV, even though an extra constant parameter appears in the latter; this can always be removed by suitable rescaling of the variables $\mathbf{P}$. We will prove Lemma 2.2 below in section 2.6.

We are now ready to write down the form for the differential operator $\bar{\mathcal{L}}$ describing the effective particle dynamics on time scales $O(\mathrm{Ku}_\mathrm{T}^{-1})$ in the limit $\mathrm{Ku}_\mathrm{T} \to 0$. Modifying the development in Appendix B of MTV in the manner indicated by Lemma 2.2, we have

$$\begin{aligned}\bar{\mathcal{L}}g(\mathbf{X}) &= -\mathbf{E}_{OU}\mathcal{L}_2\mathcal{L}_1^{-1}\mathcal{L}_2\mathbf{E}_{OU}g(\mathbf{X})\\ &= \int_{\mathbb{R}^N} d\mathbf{P}\,\hat{\tilde{P}}_{OU}(\mathbf{P})\hat{\mathcal{L}}_2\int_0^\infty dt\,\left[\hat{\mathcal{L}}_2(g(\mathbf{X})\delta(\mathbf{P}'))\right]_{\mathbf{P}'=\boldsymbol{\beta}(\mathbf{P},t)}\\ &\qquad\times\exp\left((\operatorname{Tr}\tilde{\mathcal{R}})t - \frac{1}{2}\mathbf{P}\cdot(\exp(\tilde{\mathcal{R}}t)\cdot\exp(\tilde{\mathcal{R}}t) - \mathcal{I})\cdot\mathbf{P}\right),\end{aligned} \tag{2.7}$$

where

$$\boldsymbol{\beta}(\mathbf{P},t) = \exp(\tilde{\mathcal{R}}t)\cdot\mathbf{P},$$

and

$$\hat{\tilde{P}}_{OU}(\mathbf{P}) = \exp\left(-\frac{1}{2}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}|\mathbf{p}_{\boldsymbol{\alpha}}|^2\right) \tag{2.8}$$

is the invariant measure for the particle velocities, expressed in terms of the Fourier coordinates $\mathbf{P}$. We compute this expression in pieces:

$$\begin{aligned}\left[\hat{\mathcal{L}}_2(g(\mathbf{X})\delta(\mathbf{P}'))\right]_{\mathbf{P}'=\boldsymbol{\beta}(\mathbf{P},t)} &= -i\left[\frac{\partial g}{\partial\mathbf{X}}\cdot\frac{\partial\delta(\mathbf{P}')}{\partial\mathbf{P}'}\right]_{\mathbf{P}'=\boldsymbol{\beta}(\mathbf{P},t)}\\ &= -i(\operatorname{Det}(\exp(\tilde{\mathcal{R}}t)))^{-1}\frac{\partial g}{\partial\mathbf{X}}\cdot\exp\big(-\tilde{\mathcal{R}}t\big)\cdot\frac{\partial\delta(\mathbf{P})}{\partial\mathbf{P}}.\end{aligned}$$

Thanks to the delta function of $\mathbf{P}$ and the relation $\operatorname{Det}(\exp(\tilde{\mathcal{R}}t)) = \exp(\operatorname{Tr}\tilde{\mathcal{R}}t)$, the integral over t can be evaluated to give

$$\hat{\mathcal{L}}_1^{-1}\hat{\mathcal{L}}_2\mathbf{E}_{OU}g(\mathbf{X}) = i\frac{\partial g}{\partial\mathbf{X}}\cdot\exp\big(-\tilde{\mathcal{R}}t\big)\cdot\frac{\partial\delta(\mathbf{P})}{\partial\mathbf{P}} = i\frac{\partial g}{\partial\mathbf{X}}\cdot\tilde{\mathcal{R}}^{-1}\cdot\frac{\partial\delta(\mathbf{P})}{\partial\mathbf{P}},$$

since $\tilde{\mathcal{R}}$ is a positive definite matrix [6]. Continuing,

$$\hat{\mathcal{L}}_2\hat{\mathcal{L}}_1^{-1}\hat{\mathcal{L}}_2\mathbf{E}_{OU}g(\mathbf{X}) = \operatorname{Tr}\frac{\partial}{\partial\mathbf{X}}\left(\frac{\partial g}{\partial\mathbf{X}}\cdot\tilde{\mathcal{R}}^{-1}\cdot\frac{\partial^2\delta(\mathbf{P})}{\partial\mathbf{P}^{\otimes 2}}\right) + \phi\frac{\partial g}{\partial\mathbf{X}}\cdot\tilde{\mathcal{R}}^{-1}\cdot\nabla_{\mathbf{X}}\tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right)\delta(\mathbf{P}).$$

Finally, integrating against the invariant measure (2.8) for $\mathbf{P}$ and using the symmetry of $\tilde{\mathcal{R}}$, we obtain

$$\bar{\mathcal{L}}g(\mathbf{X}) = \frac{\partial}{\partial\mathbf{X}}\cdot\left(\tilde{\mathcal{R}}^{-1}\cdot\frac{\partial g}{\partial\mathbf{X}}\right) - \phi\left((\tilde{\mathcal{R}})^{-1}\cdot\nabla_{\mathbf{X}}\tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right)\right)\cdot\frac{\partial g}{\partial\mathbf{X}}.$$

Returning to our more concrete notation, this reduced operator has exactly the (nondimensionalized) Smoluchowski form

$$\bar{\mathcal{L}}g(\mathbf{X}) = \sum_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathcal{A}}\left[\frac{\partial}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\cdot\left((\tilde{\mathcal{R}}^{-1})_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}\cdot\frac{\partial g}{\partial\mathbf{X}_{\boldsymbol{\alpha}'}}\right) - \phi\left((\tilde{\mathcal{R}}^{-1})_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}\cdot\nabla_{\boldsymbol{\alpha}'}\tilde{\Phi}\left(\frac{\mathbf{X}}{\tilde{\ell}_f}\right)\right)\cdot\frac{\partial g}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\right].$$

The drift and diffusion coefficients stated in Proposition 2.1 are read off this effective Kolmogorov backward operator in the standard way.

2.5. Generalized mass matrix. The above calculation can be directly generalized to systems of particles with unequal masses or even systems involving nondiagonal mass matrices (as in the case where some coordinates may correspond to rotational degrees of freedom) [4, 5, 6]. The nondimensionalization requires some modification, but this is straightforward for any particular system under consideration. (For example, the reference mass might be chosen as the average mass or mass of the largest particle, etc.) We assume that this nondimensionalization has been done. There will be some other nondimensional parameters characterizing the system (e.g., mass and particle size ratios), but all these will be considered to be held fixed in the low thermal Kubo number limit, and so we do not need to account for them specifically. We restrict attention to the case in which the system variables $\mathbf{X}$ and $\mathbf{V}$ involve only rigid motions of bodies so that the mass matrix is constant.

We start with the nondimensional form of the particle dynamics equations with general nondimensionalized mass matrix $\tilde{\mathcal{M}}$:

$$\begin{aligned} &d\mathbf{X}_{\boldsymbol{\alpha}}(t) = \mathbf{V}_{\boldsymbol{\alpha}}(t)\,dt,\\ &\sum_{\boldsymbol{\alpha}'\in\mathcal{A}}\tilde{\mathcal{M}}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}\cdot d\mathbf{V}_{\boldsymbol{\alpha}'}(t) = -\phi\nabla_{\boldsymbol{\alpha}}\tilde{\Phi}(\mathbf{X}(t)/\tilde{\ell}_f)\,dt\\ &\quad+\sum_{\boldsymbol{\alpha}'\in\mathcal{A}}\left[-\mathrm{Ku}_{\mathrm{T}}^{-1}\tilde{\mathcal{R}}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}(\mathbf{X}(t))\mathbf{V}_{\boldsymbol{\alpha}'}(t)\,dt + \sqrt{2}\mathrm{Ku}_{\mathrm{T}}^{-1/2}\mathcal{S}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}(\mathbf{X}(t))\,d\mathbf{W}_{\boldsymbol{\alpha}'}(t)\right], \end{aligned} \tag{2.9}$$

with $\tilde{\mathcal{S}}\tilde{\mathcal{S}}^{\dagger} = \tilde{\mathcal{R}}$ (see [6]).

We can transform this system with generalized nondimensional mass matrix to the simplified case considered above (i.e., (2.1)), where the nondimensional mass matrix is the identity matrix, as follows,

$$\mathbf{V}^{(\tilde{\mathcal{M}})} = \tilde{\mathcal{M}}^{1/2} \cdot \mathbf{V}, \qquad \mathbf{X}^{(\tilde{\mathcal{M}})} = \tilde{\mathcal{M}}^{1/2} \cdot \mathbf{X},$$

noting that the mass matrix $\tilde{\mathcal{M}}$ is symmetric and positive definite. Multiplying the first equation in (2.9) by $\tilde{\mathcal{M}}^{1/2}$ and the second equation in (2.9) by $\tilde{\mathcal{M}}^{-1/2}$ and making this change of variables, we find that

$$\begin{aligned}
&d\mathbf{X}_{\boldsymbol{\alpha}}^{(\tilde{\mathcal{M}})}(t) = \mathbf{V}_{\boldsymbol{\alpha}}^{(\tilde{\mathcal{M}})}(t)\, dt,\\
&d\mathbf{V}_{\boldsymbol{\alpha}}^{(\tilde{\mathcal{M}})}(t) = -\phi \sum_{\boldsymbol{\alpha}' \in \mathcal{A}} (\tilde{\mathcal{M}}^{-1/2})_{\boldsymbol{\alpha},\boldsymbol{\alpha}'} \cdot \nabla_{\boldsymbol{\alpha}'} \tilde{\Phi}(\mathbf{X}^{(\tilde{\mathcal{M}})}(t)/\tilde{\ell}_f)\, dt\\
&+ \sum_{\boldsymbol{\alpha}' \in \mathcal{A}} \left[-\mathrm{Ku}_{\mathrm{T}}^{-1} \tilde{\mathcal{R}}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}^{(\tilde{\mathcal{M}})}(\mathbf{X}^{(\tilde{\mathcal{M}})}(t)) \mathbf{V}_{\boldsymbol{\alpha}'}^{(\tilde{\mathcal{M}})}(t)\, dt + \sqrt{2}\mathrm{Ku}_{\mathrm{T}}^{-1/2} \mathcal{S}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}^{(\mathcal{M})}(\mathbf{X}^{(\tilde{\mathcal{M}})}(t))\, d\mathbf{W}_{\boldsymbol{\alpha}'}(t) \right],
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\mathcal{R}}^{(\tilde{\mathcal{M}})} &= \tilde{\mathcal{M}}^{-1/2} \cdot \tilde{\mathcal{R}} \cdot \tilde{\mathcal{M}}^{-1/2},\\
\mathcal{S}^{(\mathcal{M})} \cdot \mathcal{S}^{(\mathcal{M})\dagger} &= \tilde{\mathcal{R}}^{(\tilde{\mathcal{M}})}.
\end{aligned}$$

Consequently, the effective coarse-grained dynamics for a system of particles governed by a general nondimensional mass matrix can be obtained by making the following replacements in Proposition 2.1:

$$\begin{aligned}
\bar{\mathbf{X}} &\longrightarrow \tilde{\mathcal{M}}^{1/2} \cdot \bar{\mathbf{X}},\\
\tilde{\mathcal{R}} &\longrightarrow \tilde{\mathcal{M}}^{-1/2} \cdot \tilde{\mathcal{R}} \tilde{\mathcal{M}}^{-1/2},\\
\nabla_{\mathbf{X}} \tilde{\Phi}(\mathbf{X}/\tilde{\ell}_f) &\longrightarrow \tilde{\mathcal{M}}^{-1/2} \cdot \nabla_{\mathbf{X}} \tilde{\Phi}(\mathbf{X}/\tilde{\ell}_f).
\end{aligned}$$

We find in this way that the effective drift and diffusion coefficients for the particle coordinates $\mathbf{X}$ are to be replaced by

$$\begin{aligned}
\bar{\mathbf{V}}_{\boldsymbol{\alpha}}(\mathbf{X}) &= \sum_{\boldsymbol{\alpha}' \in \mathcal{A}} -((\tilde{\mathcal{R}}(\mathbf{X}))^{-1})_{\boldsymbol{\alpha},\boldsymbol{\alpha}'} \cdot \nabla_{\boldsymbol{\alpha}'} \tilde{\Phi}(\mathbf{X}/\tilde{\ell}_f)\\
&\qquad + \sum_{\boldsymbol{\alpha}',\boldsymbol{\alpha}'' \in \mathcal{A}} \nabla_{\boldsymbol{\alpha}'} \cdot (\tilde{\mathcal{R}}^{-1}(\mathbf{X}))_{\boldsymbol{\alpha}',\boldsymbol{\alpha}''} (\tilde{\mathcal{M}}^{1/2})_{\boldsymbol{\alpha}'',\boldsymbol{\alpha}},\\
\mathcal{S}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(\mathbf{X}) &= \sqrt{2}(\tilde{\mathcal{R}}^{-1/2}(\mathbf{X}))_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}.
\end{aligned}$$

In other words, only the divergence-drift term needs modification by the generalized mass matrix.

2.6. Proof of Lemma 2.2. Following Appendix B of MTV, we write

$$\hat{\mathcal{L}}_1^{-1} = -\int_0^\infty e^{\hat{\mathcal{L}}_1 t}\, dt. \tag{2.10}$$

Since $\hat{\mathcal{L}}_1$ is still a first order differential operator, we can compute $\mathcal{T}(t) \equiv e^{\hat{\mathcal{L}}_1 t}$ using the method of characteristics. The characteristic equation is

$$\frac{d\boldsymbol{\Pi}(\boldsymbol{\beta},t)}{dt} = -\mathcal{A}\boldsymbol{\Pi}(\boldsymbol{\beta},t), \qquad \boldsymbol{\Pi}(\boldsymbol{\beta},t=0) = \boldsymbol{\beta},$$

which has solution $\mathbf{\Pi}(\boldsymbol{\beta},t) = \exp(-\mathcal{A}t)\boldsymbol{\beta}$. The transformation to the characteristic coordinate $\boldsymbol{\beta}$ is therefore given by

$$\boldsymbol{\beta}(\mathbf{P},t) = \exp(\mathcal{A}t)\mathbf{P}. \tag{2.11}$$

Along the characteristics, the evolution operator obeys the ODE

$$\frac{\partial}{\partial t}\left(\mathcal{T}(t)\hat{g}|_{\mathbf{P}=\mathbf{\Pi}(\boldsymbol{\beta},t)}\right) = \left(\operatorname{Tr}\mathcal{A} - \mathbf{\Pi}(\boldsymbol{\beta},t)\cdot\mathcal{A}\cdot\mathbf{\Pi}(\boldsymbol{\beta},t)\right)\mathcal{T}(t)\hat{g}|_{\mathbf{P}=\mathbf{\Pi}(\boldsymbol{\beta},t)},$$

which has solution

$$\mathcal{T}(t)\hat{g}|_{\mathbf{P}=\mathbf{\Pi}(\boldsymbol{\beta},t)} = \exp\left(\operatorname{Tr}\mathcal{A}t - \int_0^t \mathbf{\Pi}(\boldsymbol{\beta},s)\cdot\mathcal{A}\cdot\mathbf{\Pi}(\boldsymbol{\beta},s)\,ds\right)\hat{g}\Big|_{\mathbf{P}=\boldsymbol{\beta}}.$$

The integral can be evaluated exactly, using the given fact that $\mathcal{A}$ is a symmetric matrix:

$$\begin{aligned}\int_0^t \mathbf{\Pi}(\boldsymbol{\beta},s)\cdot\mathcal{A}\cdot\mathbf{\Pi}(\boldsymbol{\beta},s)\,ds &= \int_0^t \boldsymbol{\beta}\cdot\exp(-\mathcal{A}s)\mathcal{A}\exp(-\mathcal{A}s)\cdot\boldsymbol{\beta}\\ &= \frac{1}{2}\boldsymbol{\beta}\cdot(\mathcal{I}-\exp(-\mathcal{A}t)\exp(-\mathcal{A}t))\cdot\boldsymbol{\beta}.\end{aligned}$$

Reverting from the characteristic coordinate $\boldsymbol{\beta}$ to the original coordinate $\mathbf{P}$ via (2.11), we have

$$\mathcal{T}(t)\hat{g} = \exp\left(\operatorname{Tr}\mathcal{A}t - \frac{1}{2}\mathbf{P}\cdot(\exp(\mathcal{A}t)\exp(\mathcal{A}t)-\mathcal{I})\cdot\mathbf{P}\right)\hat{g}(\boldsymbol{\beta}(\mathbf{P},t)).$$

Upon substitution into (2.10), we obtain the statement in Lemma 2.2. ∎

3. Application of stochastic mode reduction for DPD. The final simulation method which we will treat with the stochastic mode reduction procedure is DPD [1, 7, 11, 13, 23, 35, 38]. This simulation scheme has several formal similarities to the PB methods mentioned in section 2, but differs fundamentally in interpretation. DPD is intended to be used for "complex fluid" systems such as suspensions of polymers, colloids, or other macromolecules. Rather than attempting to resolve the coordinates and configurations of the macromolecules, DPD instead coarse-grains the system in terms of fluid particles which represent some parcel of the fluid suspension. These fluid particles are mesoscopic in size and interact with each other through some force laws that endeavor to capture constitutively the properties of the multiphase mixture. These forces include both conservative and dissipative components. More precisely, under DPD the equations of motion for the fluid particles can be written in the following form [13, 35]:

$$\begin{aligned} m d\mathbf{V}_{\boldsymbol{\alpha}}(t) &= \left[-\nabla_{\boldsymbol{\alpha}}\Phi(\mathbf{X}(t)) - \sum_{\boldsymbol{\alpha}'\neq\boldsymbol{\alpha}} \gamma\omega(r_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)/\ell_\gamma)(\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)\cdot\mathbf{V}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t))\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)\right] dt\\ &\quad + \sum_{\boldsymbol{\alpha}'\neq\boldsymbol{\alpha}} \sigma\tilde{\omega}(r_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)/\ell_\gamma)\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)\,dW_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t),\end{aligned} \tag{3.1a}$$

$$d\mathbf{X}_{\boldsymbol{\alpha}}(t) = \mathbf{V}_{\boldsymbol{\alpha}}(t)\,dt,$$

with initial conditions

$$\mathbf{X}_{\boldsymbol{\alpha}}(t=0)=\mathbf{X}_{0,\boldsymbol{\alpha}}, \qquad \mathbf{V}_{\boldsymbol{\alpha}}(t=0)=\mathbf{V}_{0,\boldsymbol{\alpha}}. \tag{3.1b}$$

We use the same notation as in other simulation approaches to describe analogous structures and concepts in the DPD framework. The new notation is the following:

(i) $r_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)=|\mathbf{X}_{\boldsymbol{\alpha}}(t)-\mathbf{X}_{\boldsymbol{\alpha}'}(t)|$ is the distance between two DPD particles;

(ii) $\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)=(\mathbf{X}_{\boldsymbol{\alpha}}(t)-\mathbf{X}_{\boldsymbol{\alpha}'}(t))/r_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)$ is the unit vector directed from particle $\boldsymbol{\alpha}'$ to $\boldsymbol{\alpha}$;

(iii) $\mathbf{V}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)=\mathbf{V}_{\boldsymbol{\alpha}}(t)-\mathbf{V}_{\boldsymbol{\alpha}'}(t)$ is the relative velocity between two DPD particles;

(iv) γ describes the frictional coupling constant between two DPD particles;

(v) σ denotes a normalization factor for the thermal forces;

(vi) ℓ_γ denotes a length scale over which the fluid particles exert thermal and frictional forces on each other (this may loosely be viewed as an effective size of the fluid particles);

(vii) ω and $\tilde{\omega}$ are dimensionless functions which describe the strength of the frictional and thermal coupling between particles when they are separated by various distances; these functions typically vanish beyond some particle separation distance;

(viii) $\{W_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)\}_{\boldsymbol{\alpha}\neq\boldsymbol{\alpha}'}$ is a collection of independent standard real-valued one-dimensional Brownian motions (see the discussion near (2.2)) whose differentials describe random thermal force exchanges between particles $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$.

The fluctuation-dissipation theorem, or consistency with Gibbs–Boltzmann statistics for a thermal equilibrium state, implies the following relationships between the above parameters:

$$\gamma=\frac{m\sigma^2}{2k_BT}, \qquad \omega=\tilde{\omega}^2.$$

Equations (3.1) have a discontinuity when two or more particles coalesce, unless $\omega(r)$ and $\tilde{\omega}(r)$ are chosen to vanish at small r. This is physically reasonable, since if the soft fluid particles overlap significantly, the forces they exert on each other should begin to cancel out because of the integration of their interaction over wide angles. The published numerical implementations appear to leave the discontinuity in their simulated equations, which does not matter much because coalescence is a rare event (in three dimensions). For mathematical purposes, it will be convenient to assume that $\omega(r)$ (and therefore also $\tilde{\omega}(r)$) vanishes smoothly at $r=0$.

Note that the thermal forces in DPD conserve momentum because the DPD particles alone compose the complex fluid system, in contrast to the rigid particles in the standard PB simulation method, which can exchange momentum with the fluid medium. Also, note that the friction term in DPD is different from that in the rigid PB formalism (2.1). We have restricted our attention to isothermal systems so that a common temperature T characterizes all of the DPD fluid particles. The generalizations of DPD to systems with thermal gradients [1, 11] involve an extra equation for the dynamics of the internal energy of the fluid particles which does not fit well within the stochastic modeling framework of MTV.

We will obtain here, using the MTV stochastic modeling framework, a simplified formal description of the dynamics of the DPD particles at low thermal Kubo number. This procedure may be viewed as an analogue of the Smoluchowski reduction for the conventional dynamic model of particles in a fluid (section 2), and indeed the results and calculations are rather similar after certain identifications are made. The results

from the stochastic mode reduction procedure will determine the effective drift and diffusivity of the DPD particles. As these are fictitious particles, it is not so important that these quantities correspond to fundamental physical laws (in contrast to the IB method [26, 25]).

We must stress that, unlike in the previous sections where the results had a rigorous foundation, the stochastic mode reduction procedure has only a limited formal validity for the DPD system. The reason is that the finite range of the frictional interactions allows the dynamical and uncontrolled appearance of slow modes of particle motion which do not feel frictional damping. The simplified dynamics can be expected to be relevant only for sufficiently dense systems, as we shall explain in section 3.3.4. However, we are not yet able to make this into a rigorous statement.

The coarse-grained dynamics of the DPD particles have previously been characterized in a different way in [12, 34] through the calculation of kinetic coefficients such as viscosities of the DPD fluid via Chapman–Enskog expansions. Our framework for calculation provides more detail than these kinetic coefficients in one sense because it incorporates the spatial correlation structure of the DPD fluid particle motion. On the other hand, the stochastic mode elimination framework does not seem well suited to computing quantities such as effective viscosity or pressure of the DPD fluid, since the velocities of the DPD particles, which are needed for these calculations, are fast modes which are eliminated.

3.1. Nondimensionalization.

3.1.1. Parameters of externally specified functions. The thermal and frictional coupling has already been expressed in terms of nondimensional functions ω and $\tilde{\omega}$. We nondimensionalize the potential and initial velocities as we did for PB dynamics,

$$\Phi(\mathbf{X}) = F\ell_f \tilde{\Phi}(\mathbf{X}/\ell_f), \qquad \mathbf{V}_{0,\boldsymbol{\alpha}} = V_0 \tilde{\mathbf{V}}_{0,\boldsymbol{\alpha}},$$

and the initial particle locations as in the IB method,

$$\mathbf{X}_{0,\boldsymbol{\alpha}} = \ell_\gamma \tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}};$$

this latter nondimensionalization is naturally suggested by the notion that ℓ_γ is a typical fluid particle size and the fluid particles should be space-filling.

3.1.2. Reference units. Here we will use the following:

(i) length scale $\ell_T = \ell_\gamma$,

(ii) time scale $\tau_T = \sqrt{m\ell_\gamma^2/k_BT}$,

(iii) mass scale $m_T = m$.

We have chosen these reference units based on thermal (and frictional properties) of the particles, as we did for the other methods. Note that the reference velocity scale,

$$v_T \equiv \frac{\ell_T}{\tau_T} = \sqrt{\frac{k_BT}{m}},$$

is the order of magnitude which the velocity of the fluid particles would have in thermal equilibrium in the absence of applied forces.

3.1.3. Nondimensional groups.

(i) The thermal Kubo number

$$\mathrm{Ku_T} = \sqrt{\frac{mk_BT}{\gamma^2\ell_\gamma^2}}.$$

If we were to assume that the frictional constant for the DPD particles scales with respect to physical parameters in the same way as rigid particles in viscously overdamped flows ($\gamma \propto (m/\ell_\gamma^3)\nu\ell_\gamma$), then we can also think of $\mathrm{Ku_T}$ as a thermal particle Reynolds number.

(ii) Nondimensionalized measures of the effects of structural forces and initial velocity

$$\phi = \frac{F\ell_\gamma}{k_B T}, \qquad \Upsilon = \frac{V_0}{v_T}.$$

(iii) The length scale ratio

$$\tilde{\ell}_f = \frac{\ell_f}{\ell_\gamma}.$$

3.1.4. Nondimensionalized DPD equations. Nondimensionalizing with respect to the reference units described in section 3.1.2 and denoting nondimensional variables (but not externally prescribed functions) by the same notation as their dimensional form, we obtain

$$\begin{aligned} d\mathbf{X}_{\boldsymbol{\alpha}}(t) &= \mathbf{V}_{\boldsymbol{\alpha}}(t)\,dt, \\ d\mathbf{V}_{\boldsymbol{\alpha}}(t) &= \Bigg[-\phi \sum_{\boldsymbol{\alpha}' \neq \boldsymbol{\alpha}} \nabla_{\boldsymbol{\alpha}} \tilde{\Phi}(\mathbf{X}(t)/\tilde{\ell}_f) \\ &\qquad - \sum_{\boldsymbol{\alpha}' \neq \boldsymbol{\alpha}} \mathrm{Ku_T}^{-1} \omega(r_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t))(\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t) \cdot \mathbf{V}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t))\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t) \Bigg]\, dt \\ &\quad + \sum_{\boldsymbol{\alpha}' \neq \boldsymbol{\alpha}} \sqrt{2}\mathrm{Ku_T}^{-1/2} \omega^{1/2}(r_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t))\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t)\, dW_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(t), \end{aligned}$$

with initial conditions

$$\mathbf{X}_{\boldsymbol{\alpha}}(t=0) = \tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}}, \qquad \mathbf{V}_{\boldsymbol{\alpha}}(t=0) = \Upsilon\tilde{\mathbf{V}}_{0,\boldsymbol{\alpha}}.$$

3.2. Effective dynamics for DPD particles at low Kubo number. As we shall discuss in section 3.3.4, there is a degeneracy in the DPD equations which does not appear to allow a direct, rigorous application of the theorem by Kurtz [29] to obtain the simplified small Kubo number dynamics. However, it appears that the degeneracy plays only a small role in sufficiently dense systems, and that the formal procedure should be meaningful in these situations. We therefore state the results of stochastic mode reduction on the DPD equations in terms of a formal, nonrigorous proposition.

PROPOSITION 3.1 (DPD at small Kubo number (only formal)). *Consider the DPD system* (3.1) *under the restriction that* $\sum_{\boldsymbol{\alpha}\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}}\Phi(\mathbf{X}) = 0$ *and the system is sufficiently dense. Then in the limit* $\mathrm{Ku_T} \to 0$ *with all other nondimensional quantities held fixed, the solution for the particle positions* $\{\mathbf{X}_{\boldsymbol{\alpha}}(t)\}_{\boldsymbol{\alpha}}$*, obtained from the complete DPD system* (3.1) *and rescaled in time as*

$$\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t) = \lim_{\mathrm{Ku_T}\to 0} \mathbf{X}_{\boldsymbol{\alpha}}(t/\mathrm{Ku_T}),$$

can be approximated in some weak sense by the solution of the following simplified stochastic differential system involving only the particle positions $\{\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t)\}$:

$$\begin{aligned} d\bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t) &= \bar{\mathbf{V}}_{\boldsymbol{\alpha}}(\bar{\mathbf{X}}(t))\,dt + \sum_{\boldsymbol{\alpha}'\in\mathcal{A}} \mathcal{S}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(\bar{\mathbf{X}}(t))\,d\mathbf{W}_{\boldsymbol{\alpha}'}(t), \\ \bar{\mathbf{X}}_{\boldsymbol{\alpha}}(t=0) &= \tilde{\mathbf{X}}_{0,\boldsymbol{\alpha}}, \end{aligned} \tag{3.2}$$

where the stochastic real white noise terms $d\mathbf{W}_{\boldsymbol{\alpha}}(t)$ *are defined near* (2.2) *and are given the Itô interpretation. The explicit expression for the drift term is*

$$\bar{\mathbf{V}}_{\boldsymbol{\alpha}}(\mathbf{X}) = -\sum_{\boldsymbol{\alpha}'\in\mathcal{A}} ((\tilde{\mathcal{Q}}(\mathbf{X}))^{-1})_{\boldsymbol{\alpha},\boldsymbol{\alpha}'} \cdot \nabla_{\boldsymbol{\alpha}'}\tilde{\Phi}(\mathbf{X}/\tilde{\ell}_f) + \sum_{\boldsymbol{\alpha}'\in\mathcal{A}} \nabla_{\boldsymbol{\alpha}'}(\tilde{\mathcal{Q}}^{-1}(\mathbf{X}))_{\boldsymbol{\alpha}',\boldsymbol{\alpha}}, \tag{3.3}$$

and the matrix coefficients of the stochastic terms are

$$\mathcal{S}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}(\mathbf{X}) = \sqrt{2}(\tilde{\mathcal{Q}}^{-1/2}(\mathbf{X}))_{\boldsymbol{\alpha},\boldsymbol{\alpha}'},$$

where

$$\tilde{\mathcal{Q}}_{\boldsymbol{\alpha},\boldsymbol{\alpha}'} \equiv= \begin{cases} \sum_{\boldsymbol{\alpha}''\neq\boldsymbol{\alpha}} \omega(r_{\boldsymbol{\alpha}\boldsymbol{\alpha}''})\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}''}\otimes\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}''} & \text{for } \boldsymbol{\alpha}'=\boldsymbol{\alpha}. \\ -\omega(r_{\boldsymbol{\alpha}\boldsymbol{\alpha}'})\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}\otimes\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'} & \text{for } \boldsymbol{\alpha}''\neq\boldsymbol{\alpha}. \end{cases} \tag{3.4}$$

The inverse of $\tilde{\mathcal{Q}}$ *is to be understood as having domain orthogonal to the null-space of* $\tilde{\mathcal{Q}}$.

On a formal level, the coarse-grained DPD is quite similar to that of the coarse-grained PB dynamics described in Proposition 2.1; the resistance tensor $\tilde{\mathcal{R}}$ is simply replaced by an effective DPD resistance tensor $\tilde{\mathcal{Q}}$ defined in (3.4). However, the restrictions for the validity of the effective coarse-grained DPD stated in Proposition 3.1 are substantial.

First, the total momentum of the system must be conserved, as for IB dynamics. This rules out consideration of external potentials, which would induce $O(1)$ motion of the center of mass of the DPD particle system over $O(1)$ time scales. This center of mass motion would influence the relative motion of the DPD particles because of the forces applied by the external potential. As with the IB method discussed in [24], we do not know how to generalize our simplified stochastic dynamics to incorporate these changes.

Secondly, the system must be sufficiently dense so that, with large probability, the collection of particles does not (or rarely does) break up into two or more clusters which feel no frictional or thermal coupling with each other. Indeed, in such a configuration, the relative position of the centers of mass of these two clusters would be a degree of freedom that would not feel the strong damping at all and cannot be described within the stochastic mode reduction framework. (It acts as a slow mode, but we cannot identify it a priori!) We discuss this issue below in the context of the actual stochastic mode elimination procedure.

3.3. Stochastic mode reduction for DPD. We now derive Proposition 3.1 following the general prescription from MTV, which we presented in some detail in [24]. We focus here only on those aspects of the calculation that are particular to the DPD equations; the basic notation and formalism are the same as for PB dynamics in section 2.4. We will encounter a technical gap, so the calculations here should only be viewed as formal.

3.3.1. Kolmogorov backward equation formalism. We identify $\varepsilon = \mathrm{Ku_T}$ as the small parameter and rescale to a longer time $t \to t/\varepsilon$. We denote the system variables in terms of the collection of particle positions $\mathbf{X}$ and velocities $\mathbf{V}$, as in section 2. Then the Kolmogorov backward equation under this time rescaling can be written as

$$-\frac{\partial \rho^\varepsilon(s,\mathbf{X},\mathbf{V}|t)}{\partial s} = \varepsilon^{-2}\mathcal{L}_1\rho^\varepsilon + \varepsilon^{-1}\mathcal{L}_2\rho^\varepsilon,$$
$$\rho^\varepsilon(s=t,\mathbf{X},\mathbf{V}|t) = f(\mathbf{X},\mathbf{V}),$$

with differential operators

$$\mathcal{L}_1 = \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\alpha}'} \omega(r_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}) \left[\frac{1}{2}\left(\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}\cdot\left(\frac{\partial}{\partial\mathbf{V}_{\boldsymbol{\alpha}}} - \frac{\partial}{\partial\mathbf{V}_{\boldsymbol{\alpha}'}}\right)\right)^2 - \hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}\cdot(\mathbf{V}_{\boldsymbol{\alpha}}-\mathbf{V}_{\boldsymbol{\alpha}'})\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}\cdot\frac{\partial}{\partial\mathbf{V}_{\boldsymbol{\alpha}}}\right],$$

$$\mathcal{L}_2 = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} -\phi\nabla_{\boldsymbol{\alpha}}\tilde{\Phi}\left(\frac{\mathbf{X}(t)}{\tilde{\ell}_f}\right)\frac{\partial}{\partial\mathbf{V}_{\boldsymbol{\alpha}}} + \mathbf{V}_{\boldsymbol{\alpha}}\cdot\frac{\partial}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}.$$

3.3.2. Asymptotic expansion of solution. We follow the same formalism as in [24], again assuming f depends only on the slow variables $f = f(\mathbf{X})$ to avoid consideration of initial transients. The solvability conditions for the equations in the asymptotic hierarchy are obtained by integrating against the invariant measure π_{OU} associated with the operator $\mathcal{L}_1$, which here is

$$\pi_{OU}(\mathbf{V}) = \left(\prod_{\boldsymbol{\alpha}\in\mathcal{A}}\frac{1}{2\pi}\right)\exp\left(-\frac{1}{2}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}|\mathbf{V}_{\boldsymbol{\alpha}}|^2\right).$$

The first two equations in the asymptotic hierarchy are solved in the same way as in [24], and the solvability condition resulting from the third equation is again (2.6), so the main task is to compute explicitly

$$\bar{\mathcal{L}} \equiv -\mathbf{E}_{OU}\mathcal{L}_2\mathcal{L}_1^{-1}\mathcal{L}_2\mathbf{E}_{OU}.$$

3.3.3. Explicit computation of limiting PDE. We again follow the spirit of the calculation in Appendix B of MTV but, as in section 2.4.3, need to make some modifications because $\mathcal{L}_1$ is not diagonal in the fast variables $\mathbf{V}$.

We pass again to fast Fourier variables $\mathbf{P} = \{\mathbf{p}_{\boldsymbol{\alpha}}\}$ defined through

$$\hat{g}(\mathbf{P}) = \int_{\mathbb{R}^N}\exp\left[i\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\mathbf{p}_{\boldsymbol{\alpha}}\cdot\mathbf{V}_{\boldsymbol{\alpha}}\right]g(\mathbf{V})\,d\mathbf{V}.$$

The differential operators act on functions of $\mathbf{P}$ and $\mathbf{X}$ as follows:

$$\hat{\mathcal{L}}_1 = \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\alpha}'}\omega(r_{\boldsymbol{\alpha}\boldsymbol{\alpha}'})\left[-\frac{1}{2}\left(\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}\cdot(\mathbf{p}_{\boldsymbol{\alpha}}-\mathbf{p}_{\boldsymbol{\alpha}'})\right)^2 + \hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}\cdot\left(\frac{\partial}{\partial\mathbf{p}_{\boldsymbol{\alpha}}} - \frac{\partial}{\partial\mathbf{p}_{\boldsymbol{\alpha}'}}\right)\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}\cdot\mathbf{p}_{\boldsymbol{\alpha}}\right]$$
$$= \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\alpha}'}\omega(r_{\boldsymbol{\alpha}\boldsymbol{\alpha}'})\left[-\frac{1}{2}\left(\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}\cdot(\mathbf{p}_{\boldsymbol{\alpha}}-\mathbf{p}_{\boldsymbol{\alpha}'})\right)^2 + \hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}\cdot\frac{\partial}{\partial\mathbf{p}_{\boldsymbol{\alpha}}}\hat{e}_{\boldsymbol{\alpha}\boldsymbol{\alpha}'}\cdot(\mathbf{p}_{\boldsymbol{\alpha}}-\mathbf{p}_{\boldsymbol{\alpha}'})\right],$$

$$\hat{\mathcal{L}}_2 = i\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\phi\mathbf{p}_{\boldsymbol{\alpha}}\cdot\nabla_{\boldsymbol{\alpha}}\tilde{\Phi}\left(\frac{\mathbf{X}(t)}{\tilde{\ell}_f}\right) - i\frac{\partial}{\partial\mathbf{p}_{\boldsymbol{\alpha}}}\cdot\frac{\partial}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}.$$

We proceed again as in section 2.4.3 but need to pay special attention to the structure of $\hat{\mathcal{L}}_1$. Note that $\hat{\mathcal{L}}_1$ has a form satisfying the hypotheses of Lemma 2.2 with the symmetric matrix $\mathcal{A} = \tilde{\mathcal{Q}}$ defined in (3.4), except that we need to take care about positive definiteness. As we discuss in section 3.3.4, we can readily show that $\tilde{\mathcal{Q}}$ is nonnegative definite, but it may possess several nontrivial zero eigenvalues. $\tilde{\mathcal{Q}}$ will always have one zero eigenvalue corresponding to an eigenvector with all entries equal; this simply reflects the fact that the momentum of the center of mass feels no damping. If there are no external forces, this momentum remains constant and can be projected away without difficulty. Of greater concern is the possible presence of additional zero eigenvalues, with eigenvectors depending on the particle configuration. These would imply that certain combinations of the velocities acted as slow rather than fast modes. Since these combinations are configuration-dependent, we cannot simply project them away or reclassify them a priori as slow modes. Zero eigenvalues of $\tilde{\mathcal{Q}}$ beyond the one corresponding to center of mass motion, therefore, would obstruct our ability to obtain simplified dynamics for DPD in the small Kubo number limit. We will refer to particle configurations $\mathbf{X}$ that give rise to multiple zero eigenvalues of $\tilde{\mathcal{Q}}$ as "degenerate configurations." As we shall discuss in section 3.3.4, we cannot exclude the possibility of these degenerate configurations, but the probability that any given realization of the dynamics will pass through a degenerate configuration over some finite time interval should become small as the density of the system increases. (In fact, this probability would appear to remain uniformly small over a fixed interval in the rescaled time $t \to t/\mathrm{Ku_T}$ as $\mathrm{Ku_T} \to 0$.) A larger class of realizations may go through degenerate configurations for sufficiently brief periods of time that the coarse-grained dynamics are unaltered. Therefore, we expect that the simplified stochastic dynamical description should apply in some sense to a large fraction of realizations if the density of the system is sufficiently large. To make such a statement have true mathematical sense would require a more proper formulation. Indeed, over any finite time interval, some nonzero fraction of realizations of the *simplified* dynamics described in Proposition 3.1 will pass through degenerate configurations for which the equations are not well defined; thus it is not appropriate to speak of convergence in law to the formal coarse-grained DPD equations.

We defer further discussion of these important technical matters until section 3.3.4. For now, let us suppose that we are restricting attention to the subset of systems which do not pass through a degenerate configuration, so that $\tilde{\mathcal{Q}}$ has, over the (rescaled) time interval of interest, exactly three zero eigenvalues corresponding to motion of the center of mass. We can project these degrees of freedom away from consideration by working in a frame comoving with the (constant) center of mass velocity, and then, by analogy to (2.7), we compute

$$\begin{aligned}\bar{\mathcal{L}}g(\mathbf{X}) &= -\mathbf{E}_{OU}\mathcal{L}_2\mathcal{L}_1^{-1}\mathcal{L}_2\mathbf{E}_{OU}g(\mathbf{X}) \\ &= \int_{\mathbb{R}^N} d\mathbf{P}\,\hat{\tilde{P}}_{OU}(\mathbf{P})\hat{\mathcal{L}}_2\int_0^\infty dt\,\left[\hat{\mathcal{L}}_2(g(\mathbf{X})\delta(\mathbf{P}'))\right]_{\mathbf{P}'=\boldsymbol{\beta}(\mathbf{P},t)} \\ &\qquad \times \exp\left((\,\mathrm{Tr}\,\tilde{\mathcal{Q}})t - \frac{1}{2}\mathbf{P}\cdot(\exp(\tilde{\mathcal{Q}}t)\cdot\exp(\tilde{\mathcal{Q}}t) - \mathcal{I})\cdot\mathbf{P}\right),\end{aligned}$$

where

$$\boldsymbol{\beta}(\mathbf{P},t) = \exp(\tilde{\mathcal{Q}}t)\cdot\mathbf{P}$$

and

$$\hat{\tilde{P}}_{OU}(\mathbf{P}) = \exp\left(-\frac{1}{2}\sum_{\boldsymbol{\alpha}\in\mathcal{A}}|\mathbf{p}_{\boldsymbol{\alpha}}|^2\right).$$

Then by making the same kind of calculations as we did for the PB dynamics in section 2.4.3, we obtain

$$\bar{\mathcal{L}}g(\mathbf{X}) = \sum_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathcal{A}}\left[\frac{\partial}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\cdot\left((\tilde{\mathcal{Q}}^{-1})_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}\cdot\frac{\partial g}{\partial\mathbf{X}_{\boldsymbol{\alpha}'}}\right)\right.$$
$$\left. - \phi\left((\tilde{\mathcal{Q}}^{-1})_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}\cdot\nabla_{\boldsymbol{\alpha}'}\tilde{\Phi}\left(\frac{\mathbf{X}}{\bar{\ell}_f}\right)\right)\cdot\frac{\partial g}{\partial\mathbf{X}_{\boldsymbol{\alpha}}}\right].$$

This is exactly the Kolmogorov backward differential operator corresponding to the stochastic dynamics recorded in Proposition 3.1.

3.3.4. Positive definiteness of $\tilde{\mathcal{Q}}$. We now discuss in more detail the positive definiteness of $\tilde{\mathcal{Q}}$ in (3.4).

DEFINITION 3.2. *Given a frictional coupling function $\omega(r)$, a particle configuration $\mathbf{X}$ will be said to be* frictionally connected *if for each pair of particles $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ and each pair of unit vectors $\hat{e},\hat{e}'\in\mathbb{R}^3$ there exists a finite chain of particles $\{\boldsymbol{\alpha}^{(i)}\}_{i=1}^n$ with $\boldsymbol{\alpha}^{(1)}=\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^{(n)}=\boldsymbol{\alpha}'$ obeying the following conditions:*

(i) $\omega(|\mathbf{X}_{\boldsymbol{\alpha}^{(i+1)}}-\mathbf{X}_{\boldsymbol{\alpha}^{(i)}}|)>0$ *for* $1\le i\le n-1$,

(ii) $(\mathbf{X}_{\boldsymbol{\alpha}^{(i+1)}}-\mathbf{X}_{\boldsymbol{\alpha}^{(i)}})\cdot(\mathbf{X}_{\boldsymbol{\alpha}^{(i)}}-\mathbf{X}_{\boldsymbol{\alpha}^{(i-1)}})\neq 0$ *for* $2\le i\le n-1$,

(iii) $\hat{e}\cdot(\mathbf{X}_{\boldsymbol{\alpha}^{(2)}}-\mathbf{X}_{\boldsymbol{\alpha}^{(1)}})\neq 0$ *and* $\hat{e}'\cdot(\mathbf{X}_{\boldsymbol{\alpha}^{(n)}}-\mathbf{X}_{\boldsymbol{\alpha}^{(n-1)}})\neq 0$.

Otherwise, the configuration is said to be frictionally disconnected.

Colloquially, a frictionally connected configuration is one for which any relative motion of any pair of particles induces some sort of frictional damping. This need not be a direct frictional coupling. Imagine, for example, a large but dense cluster of particles with cluster diameter larger than the distance over which frictional coupling acts. Choose two particles at opposite edges of the cluster, and pick arbitrary directions $\hat{e}$ and $\hat{e}'$ associated with the respective particles. Suppose that a chain of particles can be found between these two particles such that each pair of successive particles has a nonzero frictional interaction, no two successive pairs have orthogonal separation vectors, and the separation vectors of the pairs at the ends are not orthogonal to $\hat{e}$ and $\hat{e}'$. Then any motion of the first particle along $\hat{e}$ and the second particle along $\hat{e}'$ will create some frictional damping within the chain, no matter how the other particles are chosen to move along with the two given particles. If no such chain can be found between the two particles, then that implies that they lie in frictionally decoupled subclusters, so that there is an extra degree of freedom (rigid motion of each subcluster) beyond center of mass motion which is not frictionally damped. To understand the nonorthogonality restrictions in the definition of frictional connectedness, note that, for any pair of frictionally interacting particles, the motion of one particle along a line normal to their separation vector incurs frictional damping only at second order in its displacement (and so is vanishing to first order).

Direct calculations establish the following result.

PROPOSITION 3.3. *The matrix $\tilde{\mathcal{Q}}$ in* (3.4) *is always nonnegative definite: $\mathbf{V}\cdot\tilde{\mathcal{Q}}\cdot\mathbf{V}\ge 0$ for any vector $\mathbf{V}$. Let K be the null-space of $\tilde{\mathcal{Q}}$: $\mathbf{V}\in K\leftrightarrow\tilde{\mathcal{Q}}\cdot\mathbf{V}=\mathbf{0}$. This null-space always contains a three-dimensional subspace of vectors characterized*

by $\mathbf{V}_{\boldsymbol{\alpha}} = \mathbf{V}_{\boldsymbol{\alpha}'}$ *for all* $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathcal{A}$. *The null-space is strictly larger than this three-dimensional subspace whenever the configuration* $\mathbf{X}$ *is frictionally disconnected.*

The general three-dimensional subspace of null vectors corresponds to the center of mass motion, and we can easily eliminate them from our system by a simple linear projection if there is no external potential. The additional vectors in the null-space from frictionally disconnected configurations represent additional undamped modes in the system involving relative motion. For example, if there exists a cluster $\{\boldsymbol{\alpha} \in \mathcal{A}_c\}$ of particles which are not in frictional contact with any particles not in this cluster, then the null-space of $\tilde{\mathcal{Q}}$ contains all vectors $\mathbf{V}$ with the property $\mathbf{V}_{\boldsymbol{\alpha}} = \mathbf{V}_{\boldsymbol{\alpha}'}$ for all $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathcal{A}_c$, and $\mathbf{V}_{\boldsymbol{\alpha}} = 0$ for all $\boldsymbol{\alpha} \notin \mathcal{A}_c$.

CONJECTURE 3.4. *For frictionally connected configurations* $\mathbf{X}$*, the null-space of the matrix* $\tilde{\mathcal{Q}}$ *is precisely the three-dimensional space described in Proposition* 3.3. *There are no additional null eigenvectors.*

The reason for the conjecture is that, from a physical standpoint, the system has no undamped relative degrees of freedom other than center of mass motion. We have not been able to provide a mathematical proof, however. The matrix $\tilde{\mathcal{Q}}$ has the form of an infinitesimal generator of a continuous time Markov chain [42], except that the off-diagonal components consist of nonnegative dyadic matrices rather than nonnegative scalars. The usual arguments for characterizing the null-space of an infinitesimal generator, however, suffered algebraic defects when we tried to apply them to the case at hand.

If we grant the conjecture, then a sufficiently dense system should remain frictionally connected with high probability, and one could imagine applying Kurtz's theorem [29] only on such realizations. A proper technical framework, however, would need to be constructed to make such a statement mathematically meaningful, and this is beyond the scope of the present work.

4. Conclusion. The MTV stochastic mode reduction procedure provides a way to recover rigorously the Smoluchowski equation for PB models at low thermal Kubo number. In contrast, the governing equations for DPD have an algebraic structure amenable to the MTV procedure, but they incur analytical obstacles due to the finite range of the frictional interaction. This feature introduces a mobile degeneracy which prevents the a priori identification of the fast and slow modes in the system; slow modes can, rather, emerge dynamically. It would be interesting to formulate a mathematically sound way to obtain and describe the simplified dynamics of such a system, particularly when the degeneracies occur rarely.

Acknowledgments. The authors thank Eric Vanden-Eijnden for helpful discussions and suggestions, and the referees for constructive comments which improved the presentation.

REFERENCES

[1] J. B. AVALOS AND A. D. MACKIE, *Dissipative particle dynamics with energy conservation*, Europhys. Lett., 40 (1997), pp. 141–146.

[2] D. A. BEARD AND T. SCHLICK, *Inertial stochastic dynamics.* II. *Influence of inertia on slow kinetic processes of supercoiled DNA*, J. Comput. Phys., 112 (2000), pp. 7323–7338.

[3] L. BOCQUET, *High friction limit of the Kramers equation: The multiple time-scale approach*, Amer. J. Phys., 65 (1997), pp. 140–144.

[4] G. BOSSIS AND J. F. BRADY, *Dynamic simulation of sheared suspensions.* I. *General method*, J. Chem. Phys., 80 (1984), pp. 5141–5154.

[5] J. F. BRADY, *Brownian motion, hydrodynamics, and the osmotic pressure*, J. Chem. Phys., 98 (1993), pp. 3335–3341.

[6] J. F. Brady and G. Bossis, *Stokesian dynamics*, in Annu. Rev. Fluid Mech. 20, Annual Reviews, Palo Alto, CA, 1988, pp. 111–157.
[7] W. K. Den Otter and J. H. R. Clarke, *A new algorithm for dissipative particle dynamics*, Europhys. Lett., 53 (2001), pp. 426–431.
[8] J. M. Deutch and I. Oppenheim, *The concept of Brownian motion in modern statistical mechanics*, in Brownian Motion, Faraday Discussions of the Chemical Society 83, The Royal Society of Chemistry, London, 1987, pp. 1–20.
[9] B. Dünweg, G. S. Grest, and K. Kremer, *Molecular dynamics simulations of polymer systems*, in Numerical Methods for Polymeric Systems (Minneapolis, MN, 1996), Springer, New York, 1998, pp. 159–195.
[10] D. L. Ermak and J. A. McCammon, *Brownian dynamics with hydrodynamic interactions*, J. Chem. Phys., 69 (1978), pp. 1352–1360.
[11] P. Español, *Dissipative particle dynamics with energy conservation*, Europhys. Lett., 40 (1997), pp. 631–636.
[12] P. Español, *Fluid particle model*, Phys. Rev. E, 57 (1998), pp. 2930–2948.
[13] P. Español and P. Warren, *Statistical mechanics of dissipative particle dynamics*, Europhys. Lett., 30 (1995), pp. 191–196.
[14] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley & Sons, New York, 1986.
[15] The Faraday Division, *Concentrated Colloidal Dispersions*, Faraday Discussions of the Chemical Society 76, The Royal Society of Chemistry, London, 1983.
[16] The Faraday Division, *Brownian Motion*, Faraday Discussions of the Chemical Society 83, The Royal Society of Chemistry, London, 1987.
[17] M. Fixman, *Brownian dynamics of chain polymers*, in Brownian Motion, Faraday Discussions of the Chemical Society 83, The Royal Society of Chemistry, London, 1987, pp. 1–20.
[18] C. W. Gardiner, *Handbook of Stochastic Methods*, Springer Ser. Synergetics 13, 2nd ed., Springer-Verlag, Berlin, 1985.
[19] P. S. Grassia, E. J. Hinch, and L. C. Nitsche, *Computer simulations of Brownian motion of complex systems*, J. Fluid Mech., 282 (1995), pp. 373–403.
[20] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics with Special Applications to Particulate Media*, Prentice–Hall, Englewood Cliffs, NJ, 1965.
[21] E. H. Hauge and A. Martin-Löf, *Fluctuating hydrodynamics and Brownian motion*, J. Statist. Phys., 7 (1973), pp. 259–281.
[22] W. Hess and R. Klein, *Dynamical properties of colloidal systems.* 1. *Derivation of stochastic transport equations*, Phys. A, 94 (1978), pp. 71–90.
[23] P. J. Hoogerbrugge and J. M. V. A. Koelman, *Simulating microscopic hydrodynamic phenomena with dissipative particle dynamics*, Europhys. Lett., 19 (1992), pp. 155–160.
[24] P. R. Kramer and A. J. Majda, *Stochastic mode reduction for the immersed boundary method*, SIAM J. Appl. Math., 64 (2003), pp. 369–400.
[25] P. R. Kramer and C. S. Peskin, *An Extension of the Immersed Boundary Method Including Thermal Fluctuations*, manuscript, 2003.
[26] P. R. Kramer and C. S. Peskin, *Incorporating thermal fluctuations into the immersed boundary method*, in Proceedings of the Second MIT Conference on Computational Fluid and Solid Mechanics, Cambridge, MA, 2003, Elsevier Science, Oxford, UK, 2003, pp. 1755–1758.
[27] H. A. Kramers, *Brownian motion in a field of force and the diffusion model of chemical reactions*, Physica, 7 (1940), pp. 284–304.
[28] M. Kröger, A. Alba-Pérez, M. Laso, and H. C. Öttinger, *Variance reduced Brownian simulation of a bead-spring chain under steady shear flow considering hydrodynamic interaction effects*, J. Chem. Phys., (2000), pp. 4767–4773.
[29] T. G. Kurtz, *A limit theorem for perturbed operator semigroups with applications to random evolutions*, J. Funct. Anal., 12 (1973), pp. 55–67.
[30] H. A. Lorentz, *Lessen über Theoretische Natuurkunde: Kinetische Problemen*, Vol. 5, E. J. Brill, 1921.
[31] A. Majda, I. Timofeyev, and E. Vanden-Eijnden, *A priori tests of a stochastic mode reduction strategy*, Phys. D, 170 (2002), pp. 206–252.
[32] A. J. Majda, I. Timofeyev, and E. Vanden-Eijnden, *Models for stochastic climate prediction*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 14687–14691.
[33] A. J. Majda, I. Timofeyev, and E. Vanden-Eijnden, *A mathematical framework for stochastic climate models*, Comm. Pure Appl. Math., 54 (2001), pp. 891–974.
[34] C. A. Marsh, G. Backx, and M. H. Ernst, *Fokker-Planck-Boltzmann equation for dissipative particle dynamics*, Europhys. Lett., (1997), pp. 411–415.
[35] C. A. Marsh and J. M. Yeomans, *Dissipative particle dynamics: The equilibrium for finite*

time steps, Europhys. Lett., 37 (1997), pp. 511–516.
[36] A. J. Masters, *Time-scale separations and the validity of the Smoluchowski, Fokker-Planck and Langevin equations as applied to concentrated particle suspensions*, Molec. Phys., 57 (1986), pp. 303–317.
[37] T. J. Murphy and J. L. Aguirre, *Brownian motion of N interacting particles.* I. *Extension of the Einstein diffusion relation to the N-particle case*, J. Chem. Phys., 57 (1972), pp. 2098–2104.
[38] I. Pagonabarraga, M. H. J. Hagen, and D. Frenkel, *Self-consistent dissipative particle dynamics algorithm*, Europhys. Lett., 42 (1998), pp. 377–382.
[39] G. C. Papanicolaou, *Some probabilistic problems and methods in singular perturbations*, Rocky Mountain J. Math., 6 (1976), pp. 653–674.
[40] C. S. Peskin, *The immersed boundary method*, Acta Numer., 11 (2002), pp. 479–517.
[41] P. N. Pusey and R. J. A. Tough, *Hydrodynamic interactions and diffusion in concentrated particle suspensions*, in Concentrated Colloidal Suspensions, Faraday Discussions of the Chemical Society 76, The Royal Society of Chemistry, London, 1983, pp. 123–136.
[42] S. Resnick, *Adventures in Stochastic Processes*, Birkhäuser-Boston, Boston, MA, 1992.
[43] J. Rotne and S. Prager, *Variational treatment of hydrodynamic interaction in polymers*, J. Chem. Phys., 50 (1969), pp. 4831–4837.
[44] J.-N. Roux, *Brownian particles at different times scales: A new derivation of the Smoluchowski equation*, Phys. A, 188 (1992), pp. 526–552.
[45] J. Schröter, *The complete Chapman-Enskog procedure for the Fokker-Planck equation*, Arch. Ration. Mech. Anal., 66 (1977), pp. 183–199.
[46] A. Sierou and J. F. Brady, *Accelerated Stokesian dynamics simulations*, J. Fluid Mech., 448 (2001), pp. 115–146.
[47] U. M. Titulaer, *Corrections to the Smoluchowski equation in the presence of hydrodynamic interactions*, Phys. A, 100 (1980), pp. 251–265.
[48] M. Tokuyama and I. Oppenheim, *Statistical-mechanical theory of Brownian motion—Translational motion in an equilibrium fluid*, Phys. A, 94 (1978), pp. 501–520.
[49] N. G. van Kampen and I. Oppenheim, *Brownian motion as a problem of eliminating fast variables*, Phys. A, 138 (1986), pp. 231–248.
[50] G. Wilemski, *On the derivation of Smoluchowski equations with corrections in the classical theory of Brownian motion*, J. Statist. Phys., 14 (1976), pp. 153–169.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 423–441

THE ELECTROPHORETIC MOBILITY OF A CLOSELY FITTING SPHERE IN A CYLINDRICAL PORE*

EHUD YARIV† AND HOWARD BRENNER†

Abstract. We analyze the electrophoretic motion of a freely suspended closely fitting sphere, eccentrically positioned within an infinitely long cylindrical pore, when subjected to a uniform electric field acting parallel to the pore. The thin Debye-layer approximation is employed. Using singular perturbation expansions, the fluid domain is separated into an "inner" gap region around the sphere's equator, wherein electric field and velocity gradients are large, and an "outer" region, consisting of the remaining fluid domain, wherein field variations are moderate. Laplace's equation is solved within the gap region using stretched coordinates, whereby matching with the outer solution is facilitated by use of an integral conservation equation for the electric field flux. Using a reciprocal theorem, the electrokinetic contributions to the force (torque) on the sphere are represented as quadratures of the electric field over the sphere surface, with the respective stress fields pertaining to purely translational (rotational) motions appearing as Green's functions. The translational velocity of a concentrically positioned sphere is found to be half that for a sphere in an unbounded fluid. Both the translational and rotational sphere mobilities increase in magnitude with increasing eccentricity.

Key words. singular perturbations, Stokes flow, reciprocal theorem

AMS subject classifications. 76W05, 76D07, 76D08, 41A60

DOI. 10.1137/S0036139902411119

1. Introduction.

The effects of boundaries upon the electrophoretic motion of colloidal particles have been the subject of many studies, as these effects are inevitable in any practical system. The complete system of electrokinetic equations is strongly coupled and nonlinear, making analytic treatment difficult even for the simplest geometries [17]. This has led to extensive use of the "thin Debye layer" approximation, which results in linear equations in the outer "bulk" region, lying outside of the Debye double layers surrounding the various surfaces, say $\mathcal{S}$, in contact with the electrolytic fluid (i.e., particle and wall surfaces). Indeed, in many practical cases the Debye-layer thickness, say λ_D, is of the order of nanometers [20], which allows use of this asymptotic approximation, even for micron-size particles. Within the framework of this approximation [9], the flow in the bulk region is governed by the conventional Stokes equations (without any electrical body force), albeit satisfying a slip condition on $\mathcal{S}$. This condition reflects the finite velocity jump across the layer, which is proportional to the electric field existing at its outer edge (in the bulk domain). This field is derived from an electric potential, which satisfies a Neumann-type boundary-value problem in the bulk region.

The simplicity of the former model allows the solution of electrophoretic problems posed for geometric configurations that involve both a particle and a wall. It is common to use a sphere as a simple geometric model for particle shape in various applications.[1] In cases where the sphere is remote from the wall compared with its radius, the electrostatic and flow problems can be solved using reflection techniques.

*Received by the editors July 10, 2002; accepted for publication (in revised form) January 15, 2003; published electronically December 31, 2003. This work was supported by a postdoctoral grant from Eli Lilly and Company.
http://www.siam.org/journals/siap/64-2/41111.html
†Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02478 (yariv@mit.edu, hbrenner@mit.edu).
[1]Even long molecules, such as DNA, tend to coil into spherical equilibrium shapes.

This has been done for the cases of a sphere moving far from a planar wall [9], as well as for a relatively small sphere moving within an infinitely long circular cylinder [9, 21]. First-order electrophoretic wall effects are found in such circumstances to be proportional to $(a/b)^3$, where a is the sphere's radius and b is a typical sphere-wall distance. As such, these effects are rather weak compared with the $O(a/b)$ hydrodynamic interactions (in the absence of electrophoretic effects) occurring during sedimentation of a sphere which interacts with a distant boundary [8].

Exact sphere-boundary solutions, valid for $a/b \sim O(1)$, may be obtained using eigenfunction expansions. Such solutions are difficult to derive, except for rather simple configurations involving elementary geometrical symmetries. Roughly speaking, these types of solutions are classified into two categories. The first consists of two-sphere interactions [16, 11], as well as sphere motion in proximity to a plane [10, 13]. The second involves sphere motion in bounded regions, as in the case of a concentrically positioned sphere translating within an infinitely long circular cylinder [12]. More complex geometries necessitate the use of strictly numerical methods, such as collocation [14] or boundary integrals [18].

While these exact solutions are valid in principle for all $O(1)$ sphere-sphere or sphere-boundary separations, numerical convergence of the pertinent series solutions becomes poor for small separations. It is exactly in such circumstances that electrophoretic flows exhibit a unique behavior, qualitatively different from that occurring during conventional, electrolyte-free Stokes flows. In the latter case, particle drag becomes unbounded (and, consequently, particle mobility goes to zero) as the sphere-wall gap thickness shrinks to zero, reflecting the increasingly large viscous stresses existing within the narrow gap. However, this is not the general rule in electrophoretic motions. The most notable difference occurs during the movement of a sphere parallel to a wall, wherein the results of [10] predict increasingly *large* magnitudes of sphere velocities as the sphere-wall separation vanishes. In the case of sphere motion within a cylindrical pore, the results of [12] predict sphere velocities *comparable* with those for an unbounded domain.

It is important to emphasize that all previously mentioned results were obtained under the thin-Debye-layer approximation, and thus do not represent interactions between the respective particle and wall double layers. As such, this singular behavior has to do with gap separations that are small compared with particle size but still large relative to the Debye-layer thickness, λ_D. Mobility models for such thin gap geometries are important in the analysis of colloidal dispersion phenomena, especially in the formation of electrophoretic aggregates. Furthermore, the continuous progress achieved to date in microfabrication techniques has resulted in microfluidic devices whose channel depth is comparable to colloidal particle size [7]. The importance of the availability of accurate mobility models covering the entire range of separations has led to the development of improved numerical series solution schemes [24], as well as to the proposal of empirical formulae.

Analytic, closed-form analyses of near-contact geometries may be affected by using singular perturbation schemes. Usually, the range of validity of such asymptotic solutions overlaps those of the "exact" solutions. Thus, the combination of asymptotic approximation together with exact solutions provides uniformly valid hybrid mobility models. Several asymptotic solutions have already been obtained for geometries of the first category. In this context, analysis of the motion of a sphere parallel to a plane wall was presented in [22]. This study clearly demonstrates the strong effect upon particle mobility arising from the intense electric field existing within the narrow gap region.

The present investigation deals with a singular case associated with the second category, namely the motion of a closely fitting sphere in a direction parallel to the wall of an infinitely long circular cylinder. We consider the general case of an eccentrically positioned sphere. This geometry, which is extremely difficult to analyze in the case of arbitrary sphere-to-channel radius ratios (for which only the concentric case has been solved; see [12]), lends itself to analytic treatment within an appropriate asymptotic framework. The conventional, electrolyte-free, Stokes motion of a closely fitting sphere has already been analyzed in [4]. The same geometry was also used in the analysis of van der Waals and double-layer sphere-channel interactions [1]. During the course of the subsequent analysis, we solve the electrostatic problem for this geometry in order to obtain the electric field distribution. It is convenient to separately solve Laplace's equation in the narrow gap region between the sphere and wall using scaled variables. The matching procedure is based upon the balance of net electric flux along the channel, thereby avoiding the difficult solution of Laplace's equation in the "outer" region. With the electrostatic solution in hand, the linear Stokes flow problem is well defined. This problem is decomposed into four distinct parts, labeled (a)–(d), reflecting different boundary condition contributions. The first two parts constitute "simple" Stokes flows, with no electrokinetic effects, already available in [4]. The remaining problems, (c) and (d), reflect the electrokinetically driven portion of the flow.

In principle, the Stokes equations governing the latter problems need to be solved in analytic detail, subject to the pertinent slip conditions imposed by the electric field at the respective channel wall and sphere surfaces. However, by virtue of the existence of a reciprocal theorem [2], such detailed solutions become unnecessary, since the theorem allows the force and torque acting on the sphere to be expressed directly in terms of appropriate quadratures of the electric field over its surface, wherein the purely hydrodynamic translational and rotational stress fields available from [4] appear as Green's functions.

The present paper furnishes leading-order asymptotic solutions of the electrostatic and Stokes flow problems. These solutions rationalize the $O(1)$ velocities obtained in [12] for closely fitting geometries. While the present asymptotic scheme may be utilized to systematically derive higher-order corrections, such corrections would have but small practical significance, as the basic channel-sphere geometry itself already constitutes a rather simplified model of the actual systems encountered in practice.

This paper is organized as follows: In the next section we formulate the mathematical problem governing the electric potential and flow fields. The flow problem is then decomposed into the above-mentioned parts, (a)–(d). Geometrical descriptions of the pertinent equations and boundary conditions appropriate to small gap widths are outlined. The electrostatic problem, governing the electrical potential, is solved in section 3. Following that, a reciprocal theorem is used in section 4 to compute the "electrokinetic" flow contribution, part (c). In section 5 we demonstrate that the "potential" part, (d), does not contribute to the force and torque acting on the sphere. Finally, the separate force and torque contributions of the preceding sections are combined in section 6, and the linear and angular velocities of a freely suspended sphere are derived therefrom.

2. Problem formulation. Consider a spherical particle (radius a) eccentrically positioned within an infinitely long circular cylinder [radius $(1+\varepsilon)a$; $\varepsilon > 0$] filled with an electrolyte solution. The cylinder wall, $\mathcal{W}$, as well as the sphere surface, $\mathcal{P}$, are assumed to possess uniform surface charge densities, with respective zeta potentials

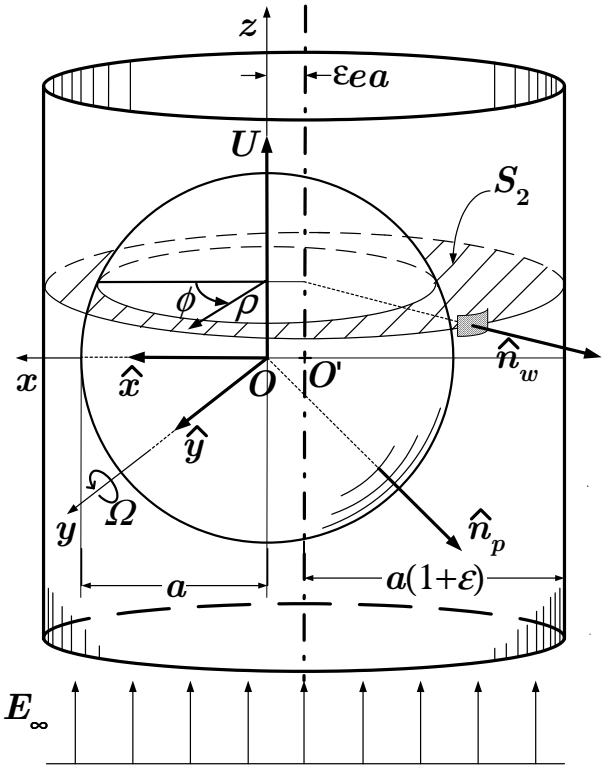


FIG. 2.1. *Schematic of the sphere-cylinder geometry.*

ζ_w and ζ_p. A uniform electric field, E_∞, is applied parallel to the cylinder walls (see Figure 2.1), causing translation and rotation of the sphere relative to $\mathcal{W}$ with respective velocities $\boldsymbol{U}$ and $\boldsymbol{\Omega}$.

We focus on the case of thin Debye layers [9] around both $\mathcal{W}$ and $\mathcal{P}$. As the gap separation, εa, constitutes the smallest bulk scale, the usual limit, namely $\lambda_D/a \to 0$, is here modified to $\lambda_D/\varepsilon a \to 0$. Thus, the electric potential φ is governed by Laplace's equation in the bulk region and satisfies a "no-flux" boundary condition on $\mathcal{W} \cup \mathcal{P}$, namely $\hat{\boldsymbol{n}} \cdot \nabla\varphi = 0$ ($\hat{\boldsymbol{n}}$ being a generic vector normal to $\mathcal{W} \cup \mathcal{P}$). Since the fluid in the bulk region is electrically neutral, no electrical body forces appear in the pertinent Stokes equations. The electric field affects the bulk flow only through the finite velocity slip, $\epsilon_{\rm el}\zeta_\alpha \nabla\varphi/\mu$ (with μ and $\epsilon_{\rm el}$, respectively, denoting the presumably uniform fluid viscosity and dielectric permittivity coefficients), experienced by the fluid on the respective wall ($\alpha = w$) and particle ($\alpha = p$) surfaces. This slip embodies Smoluchowski's result governing the velocity jump across the Debye layer.

In nondimensionalizing the pertinent equations, we normalize length variables with a, the electric potential with $E_\infty a$, linear velocities with the characteristic electrophoretic velocity $U_0 = \epsilon_{\rm el} E_\infty \zeta_p/\mu$, angular velocities with U_0/a, stresses (and pressure) with $\mu U_0/a$, forces with $\mu U_0 a$, and torques with $\mu U_0 a^2$. Thus, the dimensionless electric potential satisfies (i) Laplace's equation in the fluid domain,

$$\nabla^2\varphi = 0; \tag{2.1}$$

(ii) Neumann-type boundary conditions on the solid boundaries,

$$\hat{\boldsymbol{n}}_w \cdot \nabla\varphi = 0 \quad \text{on} \quad \mathcal{W}, \tag{2.2}$$

$$\hat{\boldsymbol{n}}_p \cdot \nabla\varphi = 0 \quad \text{on} \quad \mathcal{P} \tag{2.3}$$

(with unit normal vectors $\hat{\boldsymbol{n}}_w$ and $\hat{\boldsymbol{n}}_p$ on $\mathcal{W}$ and $\mathcal{P}$ pointing in the respective directions depicted in Figure 2.1); and (iii) the far-field condition

$$\nabla\varphi \to -\hat{\boldsymbol{z}} \quad \text{as} \quad |z| \to \infty, \tag{2.4}$$

wherein $\hat{\boldsymbol{z}}$ is a unit vector pointing in the direction of the applied electric field.

The fluid velocity and pressure fields in the bulk region, $(\boldsymbol{v}, p)$, satisfy the incompressible Stokes equations, namely

$$\nabla \cdot \boldsymbol{v} = 0, \quad \nabla p = \nabla^2 \boldsymbol{v}, \tag{2.5}$$

as well as the boundary conditions

$$\boldsymbol{v} = \gamma \nabla\varphi \quad \text{on} \quad \mathcal{W}, \tag{2.6}$$

$$\boldsymbol{v} = \nabla\varphi + \boldsymbol{U} + \boldsymbol{\Omega} \times \boldsymbol{r} \quad \text{on} \quad \mathcal{P}. \tag{2.7}$$

In the above, $\gamma = \zeta_w/\zeta_p$, and $\boldsymbol{r}$ is a position vector measured relative to the sphere center O. These conditions incorporate both rigid-body motion (on $\mathcal{P}$) and electrokinetic slip on both surfaces. Far from the sphere, the velocity field approaches an electroosmotic "plug flow,"

$$\boldsymbol{v} \to -\gamma\hat{\boldsymbol{z}} \quad \text{as} \quad |z| \to \infty, \tag{2.8}$$

with z a rectilinear coordinate measuring distance parallel to the cylinder walls (see Figure 2.1). This condition isolates the present motion from effects giving rise to *mechanically* driven motion of the fluid (and sphere), thus rendering the fluid motion purely "electrophoretic." (In circumstances where the motion is driven by both mechanical and electrical agencies, the force and torque associated with the pressure- or net-flow-driven motions [4] may be superimposed upon the present results.)

Once the velocity and pressure fields are obtained, the hydrodynamic stress field in the bulk region may be calculated by the Newtonian expression

$$\boldsymbol{\pi} = -p\mathbf{I} + \left[(\nabla\boldsymbol{v}) + (\nabla\boldsymbol{v})^\dagger\right]. \tag{2.9}$$

The force and torque acting upon the particles are then given by the respective expressions

$$\boldsymbol{F} = \oint_{\mathcal{P}} dA\, \hat{\boldsymbol{n}}_p \cdot \boldsymbol{\pi}, \tag{2.10a}$$

$$\boldsymbol{T} = \oint_{\mathcal{P}} dA\, \boldsymbol{r} \times (\hat{\boldsymbol{n}}_p \cdot \boldsymbol{\pi}). \tag{2.10b}$$

The mathematical problem governing the flow on the bulk scale, namely (2.5)–(2.10), retains the conventional form of a Stokes problem, albeit with electrokinetic slip conditions on $\mathcal{P}$ and $\mathcal{W}$. This allows for the use of general creeping-flow theorems (cf.

section 4). Moreover, given the linearity of the flow problem, it proves convenient to decompose the flow field $(\boldsymbol{v}, p)$ into four distinct parts, labeled (a)–(d), each satisfying the linear Stokes equations (cf. [22]). The linear boundary conditions (2.6)–(2.7) and the up- or downstream velocity profile (far from the sphere) are separated as follows:

	(a)	(b)	(c)	(d)
on $\mathcal{P}$:	$\boldsymbol{v} = \boldsymbol{U}$	$\boldsymbol{v} = \boldsymbol{\Omega} \times \boldsymbol{r}$	$\boldsymbol{v} = (1-\gamma)\nabla\varphi$	$\boldsymbol{v} = \gamma\nabla\varphi$
on $\mathcal{W}$:	$\boldsymbol{v} = \boldsymbol{0}$	$\boldsymbol{v} = \boldsymbol{0}$	$\boldsymbol{v} = \boldsymbol{0}$	$\boldsymbol{v} = \gamma\nabla\varphi$
as $\lvert z\rvert \to \infty$:	$\boldsymbol{v} \to \boldsymbol{0}$	$\boldsymbol{v} \to \boldsymbol{0}$	$\boldsymbol{v} \to \boldsymbol{0}$	$\boldsymbol{v} \to -\gamma\hat{\boldsymbol{z}}$

The total hydrodynamic force (and torque) on the sphere is obtained by superposing the comparable contributions resulting from flow fields (a)–(d). Flow field (a) is identical to that resulting from translation (sans rotation) of the sphere with linear velocity $\boldsymbol{U}$ within the cylinder and with no slip on the walls. Similarly, flow field (b) arises from the rotation (sans translation) of the sphere with angular velocity $\boldsymbol{\Omega}$ within the cylinder and with no slip on the walls. Parts (c) and (d) constitute the "electrokinetic" portions of the overall flow. Each individual flow field contributes, in general, to the force and torque acting on the sphere. Imposition of the requirement of zero net force and torque serves to determine $\boldsymbol{U}$ and $\boldsymbol{\Omega}$.

2.1. Sphere-cylinder geometry; closely fitting sphere. In dimensionless form, the cylinder possesses a radius $1+\varepsilon$. The sphere center O is positioned at a distance εe from the cylinder axis (see Figure 2.1). Clearly, $0 < e < 1$, the respective lower and upper bounds corresponding to concentric and fully eccentric positions. The cylinder-sphere configuration possesses two symmetry planes, both passing through O: a cross-sectional plane, say π_1, normal to the cylinder axis, and a meridian plane, say π_2. It is convenient to employ particle-fixed Cartesian (x, y, z) and circular cylindrical (ρ, ϕ, z) coordinate systems, centered about O. The z axis lies parallel to the cylinder's symmetry axis, the x axis lies along the line of intersection formed from π_1 and π_2 (with the unit vector $\hat{\boldsymbol{x}}$ pointing from the cylinder axis towards O), and the direction of the unit vector $\hat{\boldsymbol{y}}$ is taken such that the (x, y, z) axes constitute a right-handed orthogonal system. The azimuthal angle ϕ is measured from π_2. Thus, π_1 coincides with the plane $z = 0$, whereas π_2 is given by $\phi = (0, \pi)$. The cylinder wall surface $\mathcal{W}$ and sphere surface $\mathcal{P}$ are described in these coordinates by the respective relations (see Figure 2.2(a))

$$\rho_w(\phi) = \left[(1+\varepsilon)^2 - \varepsilon^2 e^2 \sin^2\phi\right]^{1/2} - \varepsilon e \cos\phi, \tag{2.11}$$

$$\rho_p(z) = (1-z^2)^{1/2}. \tag{2.12}$$

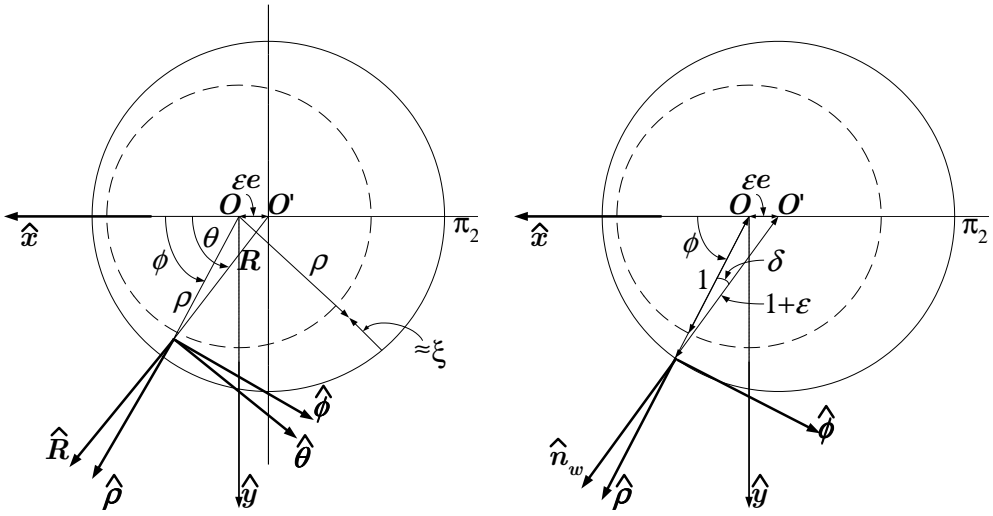
In terms of the system of circular cylindrical coordinates, the electric potential satisfies the following relations:

$$\frac{\partial^2\varphi}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial\varphi}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2\varphi}{\partial\phi^2} + \frac{\partial^2\varphi}{\partial z^2} = 0 \quad \text{for} \quad \rho_p(z) < \rho < \rho_w(\phi), \tag{2.13}$$

$$\hat{\boldsymbol{n}}_p \cdot \left(\hat{\boldsymbol{\rho}}\frac{\partial\varphi}{\partial\rho} + \hat{\boldsymbol{z}}\frac{\partial\varphi}{\partial z}\right) = 0 \quad \text{at} \quad \rho = \rho_p(z), \tag{2.14}$$

$$\hat{\boldsymbol{n}}_w \cdot \left(\hat{\boldsymbol{\rho}}\frac{\partial\varphi}{\partial\rho} + \hat{\boldsymbol{\phi}}\frac{1}{\rho}\frac{\partial\varphi}{\partial\phi}\right) = 0 \quad \text{at} \quad \rho = \rho_w(\phi), \tag{2.15}$$

$$\frac{\partial\varphi}{\partial z} \to -1 \quad \text{as} \quad \lvert z\rvert \to \infty. \tag{2.16}$$



(a) (b)

FIG. 2.2. *The cross-sectional plane* $z = 0$: (a) *particle- and wall-fixed coordinate systems.* (b) *geometrical relations between the unit vectors* $\hat{\boldsymbol{\rho}}$, $\hat{\boldsymbol{\phi}}$, *and* $\hat{\mathbf{n}}_w$.

Owing to the cylinder-sphere configurational symmetry, it is obvious (and, indeed, consistent with (2.11)–(2.16)) that φ is an odd function of z and, moreover, that it possesses symmetry with respect to reflection in the plane π_2; that is, it is invariant under the transformation $\phi \mapsto 2\pi - \phi$. A dynamical consequence of the underlying geometric symmetry is that the sphere translates parallel to the wall and rotates in a direction normal to π_2; that is, $\boldsymbol{U} = \hat{\boldsymbol{z}}U$ and $\boldsymbol{\Omega} = \hat{\boldsymbol{y}}\Omega$.[2]

In what follows, we focus upon the case of a closely fitting sphere, $\varepsilon \ll 1$. The purely hydrodynamic, potential-independent flows, (a) and (b), respectively corresponding to a translating and rotating sphere, have already been obtained [4] for that limiting case. These "mechanical" flows are highly singular, with large velocity and pressure variations occurring in the gap region, $|z| \ll 1$. The same will be seen to be true for the "electrokinetic" flows, (c) and (d). Indeed, consider the flux of electric field through the following pair of cross-sectional planes, the first passing through the sphere ($-1 < z < 1$), the second being remote ($|z| \gg 1$). As the electric vector field is solenoidal, as well as tangent to both $\mathcal{P}$ and $\mathcal{W}$ (see (2.1)–(2.3)), the uniform flux in the remote plane has to pass through the accessible part of the first plane, say S_2 (see Figure 2.1), requiring that

$$\int_0^{2\pi} d\phi \int_{\rho_p(z)}^{\rho_w(\phi)} d\rho\, \rho\, \frac{\partial \varphi}{\partial z} = -\pi\,(1+\varepsilon)^2\,. \tag{2.17}$$

[2]This rectilinear motion is consistent with the linear structure of the problem formulation. It is expected that both nonlinear electrokinetic effects, as well as Brownian diffusion, may result in a lateral migration of the particle. As such, the incorporation of both mechanisms is essential for either stability or averaging analysis of the present problem.

For $|z| \ll 1$, the annular integration domain becomes small, leading to a large electric field magnitude. As the electric field is the driver of the fluid motion (cf. (2.6)–(2.7)), it is expected that the concomitant flow fields, (c) and (d), will be singular as well. It is therefore convenient to analyze the electric field (and, subsequently, the flow) problems in the gap "inner" region independently of the more distant behavior of the pertinent fields.

For $\varepsilon \ll 1$, the cylinder wall is described by the relation

$$\rho_w(\phi) \sim 1 + \varepsilon(1 - e\cos\phi) - \varepsilon^2 \frac{1}{2}\left(e\sin\phi\right)^2 + \varepsilon^3 \frac{1}{2}\left(e\sin\phi\right)^2 + O\left(\varepsilon^4\right). \tag{2.18}$$

Define a new radial coordinate ξ as

$$\xi = 1 + \varepsilon(1 - e\cos\phi) - \rho. \tag{2.19}$$

Since $\xi \sim O(\varepsilon^2)$ on $\mathcal{W}$, this variable approximates the distance measured from the wall (along a line normal to the z axis; see Figure 2.2(a)). In terms of this new variable, and for $|z| \ll 1$, the equation describing $\mathcal{P}$ (cf. (2.11)) adopts the form

$$\xi = \varepsilon(1 - e\cos\phi) + \frac{1}{2}z^2 + \frac{1}{8}z^4 + \frac{1}{16}z^6 + \cdots. \tag{2.20}$$

In what follows, we rescale the coordinates in accordance with the scheme of [4]: Defining the "stretched" variables,

$$X = \xi/\varepsilon, \quad Z = z/\varepsilon^{1/2}, \tag{2.21}$$

leads to the following representation of the surface $\mathcal{P}$:

$$X = H(\phi, Z) + \frac{1}{8}\varepsilon Z^4 + \frac{1}{16}\varepsilon^2 Z^6 + \cdots \equiv X_p(\phi, Z; \varepsilon). \tag{2.22}$$

Here,

$$H(\phi, Z) = \left[\tau(\phi)\right]^2 + \frac{1}{2}Z^2 \tag{2.23}$$

constitutes the leading-order scaled gap separation width, with $\tau(\phi) = (1 - e\cos\phi)^{1/2}$. The comparable expansion of $\mathcal{W}$ is

$$X = \varepsilon\frac{1}{2}\left(e\sin\phi\right)^2 - \varepsilon^2\frac{1}{2}\left(e\sin\phi\right)^2 + \cdots \equiv X_w(\phi, Z; \varepsilon). \tag{2.24}$$

For future reference, we note that the gradient operator is given in terms of the stretched coordinates by the expression

$$\nabla = -\varepsilon^{-1}\hat{\boldsymbol{\rho}}\frac{\partial}{\partial X} + \varepsilon^{-1/2}\hat{\boldsymbol{z}}\frac{\partial}{\partial Z} + \hat{\boldsymbol{\phi}}\left(\frac{\partial}{\partial\phi} + e\sin\phi\frac{\partial}{\partial X}\right), \tag{2.25}$$

wherein $(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{z}})$ denote the respective unit vectors in the circular cylindrical coordinate system (see Figure 2.2(a)).

3. The electric potential within the gap. The potential $\varphi(X,\phi,Z;\varepsilon)$ in the gap region is governed by (i) Laplace's equation in the fluid domain, $X_w(\phi,Z;\varepsilon) < X < X_p(\phi,Z;\varepsilon)$,

$$\frac{\partial^2\varphi}{\partial X^2}+\varepsilon\left[\frac{\partial^2\varphi}{\partial Z^2}-\frac{1}{1+\varepsilon\,(1-\varepsilon\cos\phi-X)}\frac{\partial\varphi}{\partial X}\right] + \varepsilon^2\frac{1}{[1+\varepsilon\,(1-\varepsilon\cos\phi-X)]^2}\left(\frac{\partial^2\varphi}{\partial\phi^2}+e^2\sin\phi\frac{\partial^2\varphi}{\partial X^2}+2e\sin\phi\frac{\partial^2\varphi}{\partial X\partial\phi}+e\cos\phi\frac{\partial\varphi}{\partial X}\right)=0; \tag{3.1}$$

(ii) the boundary condition on $\mathcal{P}$,

$$\frac{\partial\varphi}{\partial X}+\varepsilon\left[(1-e\cos\phi-X)\,\frac{\partial\varphi}{\partial X}-Z\frac{\partial\varphi}{\partial Z}\right]=0 \quad \text{at} \quad X=X_p(\phi,Z;\varepsilon); \tag{3.2}$$

and (iii) the boundary condition on $\mathcal{W}$,

$$\hat{\boldsymbol{n}}_w\cdot\left[\hat{\boldsymbol{\rho}}\frac{\partial\varphi}{\partial X}-\varepsilon\hat{\boldsymbol{\phi}}\frac{1}{1+\varepsilon\,(1-e\cos\phi-X)}\left(\frac{\partial\varphi}{\partial\phi}+e\sin\phi\frac{\partial\varphi}{\partial X}\right)\right] = 0 \quad \text{at} \quad X=X_w(\phi,Z;\varepsilon). \tag{3.3}$$

From (2.24) and (2.25) we obtain the following expression for the direction of a vector normal to $\mathcal{W}$:

$$\hat{\boldsymbol{\rho}}-\varepsilon\hat{\boldsymbol{\phi}}\,e\sin\phi\,[1+O(\varepsilon)]\,. \tag{3.4}$$

Thus, to first order in ε, the corresponding unit vector $\hat{\boldsymbol{n}}_w$ is given by $\hat{\boldsymbol{\rho}}-\varepsilon\hat{\boldsymbol{\phi}}e\sin\phi$. This result could have been anticipated from Figure 2.2(b), wherein to leading order the small angle δ is given by $\varepsilon e\sin\phi$.

The inner electric field may be related to the uniform field (at the "infinity" of the *outer* problem) via the global flux condition (2.17), written in terms of the stretched coordinates[3] as

$$\varepsilon^{1/2}\int_0^{2\pi}d\phi\int_{X_w(\phi,Z;\varepsilon)}^{X_p(\phi,Z;\varepsilon)}dX\;\;[1+\varepsilon\,(1-e\cos\phi-X)]\,\frac{\partial\varphi}{\partial Z}=-\pi\,(1+\varepsilon)^2\,. \tag{3.5}$$

This condition, serving to illustrate the singular nature of the electric field, suggests the trial expansion

$$\varphi(X,\phi,Z;\varepsilon)\sim\varepsilon^{-1/2}\left[\varphi^{(0)}(X,\phi,Z)+\varepsilon\varphi^{(1)}(X,\phi,Z)+\cdots\right]. \tag{3.6}$$

The sequence of boundary-value problems governing the respective fields $\{\varphi^{(i)}\}$ $(i\geq 0)$ are obtained from (3.1)–(3.4), with the boundary conditions transferred from the actual boundaries, $X_w(\phi,Z;\varepsilon)$ and $X_p(\phi,Z;\varepsilon)$, to the corresponding virtual boundaries, $X=0$ and $X=H(\phi,Z)$, using standard techniques. The leading-order term, $\varphi^{(0)}(X,\phi,Z)$, satisfies the following equations:

$$\frac{\partial^2\varphi^{(0)}}{\partial X^2}=0 \quad \text{for} \quad 0<X<H(\phi,Z), \tag{3.7}$$

$$\frac{\partial\varphi^{(0)}}{\partial X}=0 \quad \text{at} \quad X=0,H(\phi,Z). \tag{3.8}$$

[3]Obviously, if the exact problem governing φ were to be solved, this condition would prove redundant.

The comparable $O(\varepsilon)$ balance yields

$$\frac{\partial^2 \varphi^{(1)}}{\partial X^2} = -\frac{\partial^2 \varphi^{(0)}}{\partial Z^2} \quad \text{for} \quad 0 < X < H(\phi, Z), \tag{3.9}$$

$$\frac{\partial \varphi^{(1)}}{\partial X} = 0 \quad \text{at} \quad X = 0, \tag{3.10}$$

$$\frac{\partial \varphi^{(1)}}{\partial X} = Z \frac{\partial \varphi^{(0)}}{\partial Z} \quad \text{at} \quad X = H(\phi, Z). \tag{3.11}$$

As noted above, the electric field in the gap region is related to that in the outer region through the integral flux balance (3.5), which, to leading order, yields

$$\int_0^{2\pi} d\phi \int_0^{H(\phi,Z)} dX \; \frac{\partial \varphi^{(0)}}{\partial Z} = -\pi. \tag{3.12}$$

It is obvious from (3.7)–(3.8) that $\varphi^{(0)} = \varphi^{(0)}(\phi, Z)$. This function is to be obtained from the solvability conditions (cf. [19]) governing the first-order problem (3.9)–(3.11), namely

$$\frac{\partial^2 \varphi^{(0)}}{\partial Z^2} + \frac{Z}{H(\phi, Z)} \frac{\partial \varphi^{(0)}}{\partial Z} = 0. \tag{3.13}$$

This second-order differential equation possesses the solution

$$\frac{\partial \varphi^{(0)}}{\partial Z} = \frac{C(\phi)}{H(\phi, Z)}, \quad \varphi^{(0)} = \frac{\sqrt{2} C(\phi)}{\tau(\phi)} \arctan \frac{Z}{\tau(\phi)\sqrt{2}} + D(\phi). \tag{3.14}$$

The functions of integration, $C(\phi)$ and $D(\phi)$, cannot be determined from the formulation of the gap problem alone. (Indeed, as the far-field condition (2.16) is inapplicable within the gap region, the boundary-value problem posed by (3.1)–(3.3) does not possess a unique solution.) In standard asymptotic approaches, such functions are determined by matching of the inner gap solution with an appropriate outer one. To leading order, the outer problem consists of a sphere positioned within a circular cylinder of an equal radius. Since these two geometric surfaces do not belong to a common orthogonal family of coordinates, the solution of this problem would appear to be extremely difficult to obtain.

In what follows we present an alternative matching procedure to obtain $\varphi^{(0)}$, invoking a line of reasoning similar to that of [4]. The *ansatz* lies in the existence of the limit

$$\Delta \varphi^{(0)}(\phi) = \lim_{Z \to -\infty} \varphi^{(0)}(\phi, Z) - \lim_{Z \to \infty} \varphi^{(0)}(\phi, Z) = -\frac{\sqrt{2}\pi C(\phi)}{\tau(\phi)}. \tag{3.15}$$

Since the electric field is $O(1)$ in the outer region, it is obvious that this limit constitutes the leading-order term in the expression for the outer potential difference (namely between the values in the region just below the sphere equator, $z < 0$, and that immediately above, $z > 0$). But this term cannot depend upon ϕ, for otherwise it would produce $O(\varepsilon^{-1/2})$ electric fields in the outer region. We therefore conclude that

$$C(\phi) = -\frac{\Delta \varphi^{(0)}}{\pi\sqrt{2}} \tau(\phi), \tag{3.16}$$

wherein the potential difference $\Delta\varphi^{(0)}$ is a constant which depends only upon the position of the sphere within the cylinder cross section. (As another consequence, we find that $D(\phi)$ cannot depend upon ϕ and may thus be taken to be zero.) This constant is obtained from (3.12), yielding

$$\Delta\varphi^{(0)} = \frac{\pi\mu_0}{\sqrt{2}}, \tag{3.17}$$

wherein the parameter

$$\mu_0 = \frac{2\pi}{\int_0^{2\pi} \tau(\phi)\, d\phi} \tag{3.18}$$

is a function of e. Back substitution into (3.16) yields

$$C(\phi) = -\frac{\mu_0 \tau(\phi)}{2}, \tag{3.19}$$

which completes the calculation of the leading-order electric potential in the gap region. With this latter field available, we now turn to analyze the flow fields derived by this field, namely parts (c) and (d).

4. Flow field (c). The boundary-value problem governing the flow field (c) consists of (i) Stokes equations in the fluid domain,

$$\nabla \cdot \boldsymbol{v} = 0, \quad \nabla^2 \boldsymbol{v} = \nabla p; \tag{4.1}$$

(ii) the nonhomogeneous boundary condition on the sphere,

$$\boldsymbol{v} = (1-\gamma)\,\nabla\varphi \quad \text{on} \quad \mathcal{P}; \tag{4.2}$$

and (iii) the homogeneous boundary conditions on the wall and at infinity,

$$\boldsymbol{v} = \boldsymbol{0} \quad \text{on} \quad \mathcal{W}, \tag{4.3}$$

$$\boldsymbol{v} \to \boldsymbol{0} \quad \text{as} \quad |z| \to \infty. \tag{4.4}$$

In principle, once this problem is solved, the accompanying force and torque may be obtained via integration (2.10) of the resulting stress field, obtained from (2.9). However, this detailed calculation can be avoided by utilizing a result, obtained [2] for the case of a particle positioned in an arbitrary Stokes flow which vanishes at infinity. Specifically, the hydrodynamic force and torque acting on the particle are expressed as linear functionals of the prescribed velocity field at its surface, say $\widetilde{\boldsymbol{v}}$. This result, in the present dimensionless notation, is formulated in the following relations:

$$\boldsymbol{F}_k = \oint_{\mathcal{P}} dA\, \hat{\boldsymbol{n}}_p \cdot \boldsymbol{\Pi}^{\dagger}_{\mathrm{tr}} \cdot \widetilde{\boldsymbol{v}}, \qquad \boldsymbol{T} = \oint_{\mathcal{P}} dA\, \hat{\boldsymbol{n}}_p \cdot \boldsymbol{\Pi}^{\dagger}_{\mathrm{rot}} \cdot \widetilde{\boldsymbol{v}}, \tag{4.5}$$

wherein $\boldsymbol{\Pi}_{\mathrm{tr}}$ and $\boldsymbol{\Pi}_{\mathrm{rot}}$ denote the respective translational and rotational triadic "stress" fields [8]. Explicitly, the dyadic stress field arising from translational motion of the particle with an arbitrary velocity $\boldsymbol{U}$ is given by $\boldsymbol{\Pi}_{\mathrm{tr}} \cdot \boldsymbol{U}$; similarly, the dyadic stress field arising from rotational motion of the particle with an arbitrary angular velocity $\boldsymbol{\Omega}$ is given by $\boldsymbol{\Pi}_{\mathrm{rot}} \cdot \boldsymbol{\Omega}$.

The original derivation [2] of this result was given for a particle in an unbounded fluid. However, equations (4.5) are equally applicable to bounded systems so long

as the velocity field vanishes on the walls (cf. [5, 22]). As such, it directly applies to flow field (c) (which is actually the reason for the separation of the electrokinetic contribution into the two distinct parts, (c) and (d)). It is readily verified that the respective nonvanishing components of the force (in the z direction) and torque (about the y direction) acting on the sphere are, respectively, given by

$$F = \oint_{\mathcal{P}} dA\, \hat{\boldsymbol{n}}_p \cdot \boldsymbol{\pi}_{\rm tr} \cdot \widetilde{\boldsymbol{v}}, \quad T = \oint_{\mathcal{P}} dA\, \hat{\boldsymbol{n}}_p \cdot \boldsymbol{\pi}_{\rm rot} \cdot \widetilde{\boldsymbol{v}}, \tag{4.6}$$

wherein $\boldsymbol{\pi}_{\rm tr}$ ($\boldsymbol{\pi}_{\rm rot}$) denotes the stress field associated with pure translation (rotation) of the sphere with a unit linear (angular) velocity in the z (y) direction. The translational and rotational stress fields, $\boldsymbol{\pi}_{\rm tr}$ and $\boldsymbol{\pi}_{\rm rot}$, may be obtained by substitution of the respective velocity fields into (2.9). In the present context, the outward unit vector $\hat{\boldsymbol{n}}_p$ is simply given by $\rho\hat{\boldsymbol{\rho}}+z\hat{\boldsymbol{z}}$, whereas the velocity field on $\mathcal{P}$ is $\widetilde{\boldsymbol{v}} = (1-\gamma)\nabla\varphi$.

Consider first the force acting on the sphere. The translational velocity field, say $\boldsymbol{v}_{\rm tr}$, for the case of a closely fitting sphere, was evaluated by [4]. The corresponding tractions on $\mathcal{P}$ are symmetric with respect to z, as too is the electric field (cf. (3.14)). We thus obtain

$$F = 2(1-\gamma)\int_0^{2\pi} d\phi \int_0^1 dz\, \hat{\boldsymbol{n}}_p \cdot \boldsymbol{\pi}_{\rm tr} \cdot \nabla\varphi. \tag{4.7}$$

Introduction of an intermediate parameter χ ($\varepsilon^{1/2} \ll \chi \ll 1$) enables the decomposition of this integral into respective "inner" (i.e., gap) and "outer" contributions

$$F_{\rm inner} = 2(1-\gamma)\int_0^{2\pi} d\phi \int_0^{\chi} dz\, \hat{\boldsymbol{n}}_p \cdot \pi_{\rm tr} \cdot \nabla\varphi, \tag{4.8}$$

$$F_{\rm outer} = 2(1-\gamma)\int_0^{2\pi} d\phi \int_{\chi}^1 dz\, \hat{\boldsymbol{n}}_p \cdot \pi_{\rm tr} \cdot \nabla\varphi. \tag{4.9}$$

Within the gap region, the translational velocity and pressure fields were shown [4] to possess the following expansions:

$$u \sim \varepsilon^{-1/2}\left[u^{(0)}(X,\phi,Z) + \varepsilon u^{(1)}(X,\phi,Z) + \cdots\right], \tag{4.10}$$

$$v \sim \varepsilon^{-1/2}\left[v^{(0)}(X,\phi,Z) + \varepsilon v^{(1)}(X,\phi,Z) + \cdots\right], \tag{4.11}$$

$$w \sim \varepsilon^{-1}\left[w^{(0)}(X,\phi,Z) + \varepsilon w^{(1)}(X,\phi,Z) + \cdots\right], \tag{4.12}$$

$$p \sim \varepsilon^{-5/2}\left[p^{(0)}(X,\phi,Z) + \varepsilon p^{(1)}(X,\phi,Z) + \cdots\right]. \tag{4.13}$$

In the above, (u,v,w) denote the respective velocity components relative to the circular cylindrical system. The corresponding tractions within the gap region are obtained via substitution of the above expansions into (2.9), yielding

$$\hat{\boldsymbol{n}}_p \cdot \boldsymbol{\pi}_{\rm tr} \sim -\left\{\varepsilon^{-5/2}\hat{\boldsymbol{\rho}}\, p^{(0)} - \varepsilon^{-3/2}\hat{\boldsymbol{\phi}}\frac{\partial v^{(0)}}{\partial X} - \varepsilon^{-2}\hat{\boldsymbol{z}}\left[Zp^{(0)} + \frac{\partial w^{(0)}}{\partial X}\right]\right\}\left[1 + O(\varepsilon)\right]. \tag{4.14}$$

The leading-order expression for $\nabla\varphi$ is obtained using (2.25) and (3.6) (note that $\varphi^{(0)}$ is independent of X):

$$\nabla\varphi \sim \left[-\varepsilon^{-1/2}\hat{\boldsymbol{\rho}}\frac{\partial\varphi^{(1)}}{\partial X} + \varepsilon^{-1/2}\hat{\boldsymbol{\phi}}\frac{\partial\varphi^{(0)}}{\partial\phi} + \varepsilon^{-1}\hat{\boldsymbol{z}}\frac{\partial\varphi^{(0)}}{\partial Z}\right]\left[1 + O(\varepsilon)\right]. \tag{4.15}$$

Thus, the leading-order inner contribution to the force is

$$F_{\text{inner}} \sim 2\varepsilon^{-5/2}(1-\gamma)\int_0^{2\pi} d\phi \int_0^{\chi/\varepsilon^{1/2}} dZ \left[p^{(0)} \frac{\partial \varphi^{(1)}}{\partial X} - Z p^{(0)} \frac{\partial \varphi^{(0)}}{\partial Z} - \frac{\partial w^{(0)}}{\partial X} \frac{\partial \varphi^{(0)}}{\partial Z} \right]_{X=H}. \tag{4.16}$$

Use of (3.11) reveals that the first two terms in the brackets cancel one another at $X = H$. The expressions of [4] give $\left[\partial w^{(0)}/\partial X\right]_{X=H} = 3\tau^5 \eta_0 / H^2$, wherein the parameter

$$\eta_0 = \frac{2\pi}{\int_0^{2\pi} \left[\tau(\phi)\right]^5 \, d\phi} \tag{4.17}$$

is a function of e. Since the bracketed expression in (4.16) is integrable over $0 < Z < \infty$ (cf. (3.14)), and since χ is an arbitrary parameter, the upper range of the integral over Z may be set equal to ∞:

$$F_{\text{inner}} \sim 3(1-\gamma)\mu_0 \eta_0 \varepsilon^{-5/2} \int_0^{2\pi} d\phi \, \left[\tau(\phi)\right]^6 \int_0^{\infty} \frac{dZ}{\left[H(\phi, Z)\right]^3}. \tag{4.18}$$

Performing the integration [6] yields

$$F_{\text{inner}} \sim (1-\gamma)\varepsilon^{-5/2} f_{\text{el}}^{(0)}(e), \tag{4.19}$$

wherein

$$f_{\text{el}}^{(0)}(e) = \frac{9\sqrt{2}\pi^2 \eta_0(e)}{8}. \tag{4.20}$$

It is obvious that the leading-order outer region contribution is asymptotically smaller, since to leading order the outer region geometry is independent of ε (cf. [22]). The total force resulting from the "electric" flow field (c) thus possesses the expansion

$$F = (1-\gamma) f_{\text{el}}(e; \varepsilon), \tag{4.21}$$

wherein

$$f_{\text{el}}(e) \sim \varepsilon^{-5/2} f_{\text{el}}^{(0)}(e) \left[1 + O(\varepsilon)\right]. \tag{4.22}$$

Next, consider the torque acting on the sphere. The rotational flow field, associated with rotation of the sphere at unit angular velocity, possesses the same structure as the translational one, namely (4.10)–(4.13), except that all leading-order fields vanish identically. (This is equivalent to the pertinent expansions beginning at one higher-order exponent in ε.) Thus, the counterpart of (4.16) is

$$T_{\text{inner}} \sim 2\varepsilon^{-3/2}(1-\gamma)\int_0^{2\pi} d\phi \int_0^{\chi/\varepsilon^{1/2}} dZ \left[p^{(1)} \frac{\partial \varphi^{(1)}}{\partial X} - Z p^{(1)} \frac{\partial \varphi^{(0)}}{\partial Z} - \frac{\partial w^{(1)}}{\partial X} \frac{\partial \varphi^{(0)}}{\partial Z} \right]_{X=H}, \tag{4.23}$$

with $p^{(1)}$ and $w^{(1)}$ now denoting flow variables of the rotational problem. As with the comparable expression (4.16), the first two terms in the brackets cancel one another.

The detailed solution of the rotational problem appears in [3], from which it is found that

$$\left[\frac{\partial w^{(1)}}{\partial X}\right]_{X=H} = \frac{2e\eta_0\left[\tau\left(\phi\right)\right]^5}{\left[H\left(\phi,Z\right)\right]^2} + \frac{4\left[\tau\left(\phi\right)\right]^2\cos\phi}{\left[H\left(\phi,Z\right)\right]^2} - \frac{4\cos\phi}{H\left(\phi,Z\right)}.$$

Substitution into (4.23) of this expression, in conjunction with (3.14), yields

$$T_{\text{inner}} \sim 2(1-\gamma)\mu_0\varepsilon^{-3/2}\int_0^{2\pi}d\phi\int_0^{\infty}dZ\left\{\frac{e\eta_0\left[\tau\left(\phi\right)\right]^6}{\left[H\left(\phi,Z\right)\right]^3} + \frac{2\left[\tau\left(\phi\right)\right]^3\cos\phi}{\left[H\left(\phi,Z\right)\right]^3} - \frac{2\left[\tau\left(\phi\right)\right]\cos\phi}{\left[H\left(\phi,Z\right)\right]^2}\right\}. \tag{4.24}$$

Effecting the integration [6] gives

$$T_{\text{inner}} \sim \varepsilon^{-3/2}\, g_{\text{el}}^{(0)}(e), \tag{4.25}$$

wherein

$$g_{\text{el}}^{(0)}(e) = \frac{3\sqrt{2}\pi^2}{4}e\,\eta_0(e) + \frac{\left(1-e^2\right)^{1/2}-1}{e\left(1-e^2\right)^{1/2}}\frac{\sqrt{2}\pi^2\mu_0(e)}{2}. \tag{4.26}$$

Again, as with the comparable expression governing the force, the contribution of the outer region to the torque is asymptotically smaller than that given by the preceding expression. The total torque resulting from flow field (c) is thus

$$T = (1-\gamma)g_{\text{el}}(e), \tag{4.27}$$

wherein

$$g_{\text{el}}(e;\varepsilon) \sim \varepsilon^{-3/2} g_{\text{el}}^{(0)}(e)\left[1+O\left(\varepsilon\right)\right]. \tag{4.28}$$

5. Flow field (d). Here, the flow field satisfies the following boundary conditions:

$$\boldsymbol{v} = \gamma\nabla\varphi \quad \text{on} \quad \mathcal{P}, \tag{5.1}$$

$$\boldsymbol{v} = \gamma\nabla\varphi \quad \text{on} \quad \mathcal{W}, \tag{5.2}$$

$$\boldsymbol{v} \to \gamma\nabla\varphi = -\gamma\hat{\boldsymbol{z}} \quad \text{as} \quad |z| \to \infty. \tag{5.3}$$

As any irrotational flow field satisfies the Stokes equations identically, it is obvious that $\boldsymbol{v} \equiv \gamma\nabla\varphi$ (cf. [22]). Moreover, following (2.1), no pressure variations are associated with this flow field. The corresponding stress field (see (2.9)) is

$$\boldsymbol{\pi} = 2\gamma\nabla\nabla\varphi. \tag{5.4}$$

Consider the resulting force on the sphere, given by (see (2.10a))

$$\boldsymbol{F} = 2\gamma\oint_{\mathcal{P}} dA\,\hat{\boldsymbol{n}}_p\cdot\nabla\nabla\varphi. \tag{5.5}$$

As the stress field is divergence free, the surface of integration may be deformed into any conveniently configured surface within the fluid domain, say $\mathcal{P}'$, that completely encloses the sphere; that is,

$$\boldsymbol{F} = 2\gamma \oint_{\mathcal{P}'} dA\, \hat{\boldsymbol{n}} \cdot \nabla\nabla\varphi, \tag{5.6}$$

wherein $\hat{\boldsymbol{n}}$ denotes a generic unit vector pointing out of $\mathcal{P}'$. Thus, we choose a finite cylinder, say of height $2h$, whose envelope, which coincides laterally with $\mathcal{W}$, possesses two flat ends, $z = \pm h$. In effecting the requisite integration, it is convenient to employ a second circular cylindrical coordinate system, (R, θ, z), centered about the intersection O' of the cylinder axis with π_1 (see Figure 2.2). The gradient operator in this system is

$$\nabla = \hat{\boldsymbol{R}}\frac{\partial}{\partial R} + \hat{\boldsymbol{\theta}}\frac{1}{R}\frac{\partial}{\partial \theta} + \hat{\boldsymbol{z}}\frac{\partial}{\partial z}, \tag{5.7}$$

wherein $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{\theta}}$ denote the respective radial and azimuthal unit vectors. Note that the unit vectors $(\hat{\boldsymbol{R}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{z}})$ are independent of both R and Z. In terms of the new coordinates, the boundary condition (2.2) on the cylinder wall, $R = 1 + \varepsilon$, adopts the simple form

$$\frac{\partial \varphi}{\partial R} = 0. \tag{5.8}$$

Now, consider the force tractions acting on this wall, wherein $\hat{\boldsymbol{n}} = \hat{\boldsymbol{R}}$. The integrand of (5.5) possesses the form

$$\hat{\boldsymbol{n}} \cdot \nabla\nabla\varphi = \hat{\boldsymbol{R}}\frac{\partial^2 \varphi}{\partial R^2} + \hat{\boldsymbol{\theta}}\left(\frac{1}{R}\frac{\partial^2 \varphi}{\partial \theta \partial R} - \frac{1}{R^2}\frac{\partial \varphi}{\partial \theta}\right) + \hat{\boldsymbol{z}}\frac{\partial^2 \varphi}{\partial Z \partial R}. \tag{5.9}$$

In light of the boundary condition (5.8), both terms involving mixed derivatives vanish identically. The other two terms are antisymmetric in z, which results in the respective tractions for the regions $z > 0$ and $z < 0$ cancelling one another. Thus, the wall portion of $\mathcal{P}'$ does not contribute to the force. On the two ends of $\mathcal{P}'$, whereon $z = \pm h$ and $\hat{\boldsymbol{n}} = \pm\hat{\boldsymbol{z}}$, the integrand possesses the form

$$\hat{\boldsymbol{n}} \cdot \nabla\nabla\varphi = \pm\left(\hat{\boldsymbol{R}}\frac{\partial^2 \varphi}{\partial Z \partial R} + \hat{\boldsymbol{\theta}}\frac{1}{R}\frac{\partial^2 \varphi}{\partial \theta \partial Z} + \hat{\boldsymbol{z}}\frac{\partial^2 \varphi}{\partial Z^2}\right). \tag{5.10}$$

As h becomes large, all the derivatives in this expression become small (see (2.16)), while the area of the integration domain, namely $(1 + \varepsilon)^2$, remains constant. As the force is independent of h, it is clear that the contributions from the two ends also vanish.

The torque acting on the sphere is given by

$$\boldsymbol{T} = 2\gamma \oint_{\mathcal{P}} dA\, \boldsymbol{r} \times (\hat{\boldsymbol{n}}_p \cdot \nabla\nabla\varphi). \tag{5.11}$$

To evaluate this integral, we use the spherical form of the gradient operator,

$$\nabla = \hat{\boldsymbol{r}}\frac{\partial}{\partial r} + \frac{1}{r}\nabla_{\boldsymbol{e}}, \tag{5.12}$$

wherein $r = |\boldsymbol{r}|$, with $\hat{\boldsymbol{r}}$ a radial unit vector pointing away from O (that is, $\boldsymbol{r} = r\hat{\boldsymbol{r}}$) and the operator $\nabla_{\boldsymbol{e}} = r(\mathbf{I} - \hat{\boldsymbol{r}}\hat{\boldsymbol{r}}) \cdot \nabla$ representing differentiation on the surface of the unit sphere $\mathcal{P}$. (In principle, this operator may be expressed in terms of any two polar angles about O.) Note that $\boldsymbol{r} = \hat{\boldsymbol{r}} = \hat{\boldsymbol{n}}_p$ on $\mathcal{P}$ and that $\hat{\boldsymbol{r}}$ is independent of r. Hence,

$$\begin{aligned} \boldsymbol{T}/2\gamma &= \oint_{\mathcal{P}} dA\, \hat{\boldsymbol{r}} \times \frac{\partial}{\partial r} (\nabla\varphi) \\ &= \oint_{\mathcal{P}} dA\, \hat{\boldsymbol{r}} \times \left[\hat{\boldsymbol{r}} \frac{\partial^2 \varphi}{\partial r^2} + \frac{1}{r} \nabla_{\boldsymbol{e}} \left(\frac{\partial \varphi}{\partial r} \right) - \frac{1}{r^2} \nabla_{\boldsymbol{e}} \varphi \right] \\ &= \oint_{\mathcal{P}} dA\, \hat{\boldsymbol{r}} \times \left[\nabla_{\boldsymbol{e}} \left(\frac{\partial \varphi}{\partial r} \right) - \nabla_{\boldsymbol{e}} \varphi \right]. \end{aligned} \tag{5.13}$$

In terms of the spherical coordinate r, the potential φ satisfies the following condition on $\mathcal{P}$ (see (2.3)):

$$\left(\frac{\partial \varphi}{\partial r} \right)_{r=1} = 0. \tag{5.14}$$

As the operator $\nabla_{\boldsymbol{e}}$ is r-independent, it commutes with evaluation at $r = 1$. Thus, the first term in the brackets of (5.13) vanishes; explicitly,

$$\begin{aligned} \boldsymbol{T}/2\gamma &= -\oint_{\mathcal{P}} dA\, \hat{\boldsymbol{r}} \times \nabla_{\boldsymbol{e}} \varphi \\ &= -\oint_{\mathcal{P}} dA\, \hat{\boldsymbol{r}} \times \nabla\varphi = -\oint_{\mathcal{P}} dA\, \hat{\boldsymbol{n}}_p \times \nabla\varphi. \end{aligned} \tag{5.15}$$

Since $\nabla\varphi$ is a curl-free field, the last integral may be evaluated over any surface within the fluid domain for which we again employ the finite cylinder $\mathcal{P}'$. On the wall portion of $\mathcal{P}'$, wherein $\hat{\boldsymbol{n}} = \hat{\boldsymbol{R}}$, we obtain the contribution

$$\int_0^{2\pi} d\theta \int_{-h}^{h} dz \left(\hat{\boldsymbol{\theta}} \frac{\partial \varphi}{\partial z} - \hat{\boldsymbol{z}} \frac{1}{R} \frac{\partial \varphi}{\partial \theta} \right). \tag{5.16}$$

In view of the antisymmetric dependence of φ upon z, the second term does not contribute to this integral. Performing the inner integration furnishes the intermediate expression

$$\int_0^{2\pi} [\varphi(R = 1 + \varepsilon, \theta, h) - \varphi(R = 1 + \varepsilon, \theta, -h)]\, \hat{\boldsymbol{\theta}}(\theta)\, d\theta. \tag{5.17}$$

As h becomes large, the dependence of the bracketed terms upon θ becomes exponentially weak. Since $\int_0^{2\pi} \hat{\boldsymbol{\theta}}(\theta)\, d\theta = \mathbf{0}$, the above integral must vanish. Finally, on the two ends of the cylinder, wherein $\hat{\boldsymbol{n}} = \pm\hat{\boldsymbol{z}}$, $\nabla\varphi$ tends to $-\hat{\boldsymbol{z}}$ as h becomes large; hence, $\hat{\boldsymbol{n}} \times \nabla\varphi$ approaches zero there. We thus conclude that the surface integral of $\hat{\boldsymbol{n}} \times \nabla\varphi$ over $\mathcal{P}'$ must vanish, and, consequently, that flow field (d) does not generate a torque on the sphere.

6. Sphere mobility. With the contributions of flows (c) and (d) evaluated, the various terms in the expressions for the force and torque on the sphere may be assembled. The force and torque resulting from the translational flow field (a) have already been calculated [4] for $\varepsilon \ll 1$ and are represented in the respective forms

$$F = U f_{\mathrm{tr}}(e; \varepsilon), \quad T = U g_{\mathrm{tr}}(e; \varepsilon). \tag{6.1}$$

Similarly, the force and torque resulting from the rotational flow field (b) are represented as

$$F = \Omega f_{\rm rot}(e;\varepsilon), \quad T = \Omega g_{\rm rot}(e;\varepsilon). \tag{6.2}$$

The respective force and torque coefficients appearing above have been shown [4] to possess the following asymptotic expansions:

$$f_{\rm tr}(e;\varepsilon) \sim \varepsilon^{-5/2} f_{\rm tr}^{(0)}(e)\left[1+O(\varepsilon)\right], \quad g_{\rm tr}(e;\varepsilon) \sim \varepsilon^{-3/2} g_{\rm tr}^{(0)}(e)\left[1+O(\varepsilon)\right], \tag{6.3}$$

$$f_{\rm rot}(e;\varepsilon) \sim \varepsilon^{-3/2} f_{\rm rot}^{(0)}(e)\left[1+O(\varepsilon)\right], \quad g_{\rm rot}(e;\varepsilon) \sim \varepsilon^{-1/2} g_{\rm rot}^{(0)}(e)\left[1+O(\varepsilon)\right]. \tag{6.4}$$

The force and torque resulting from the "electric" portion (c) were obtained in section 4 (see (4.21)–(4.22), (4.27)–(4.28)). Also, as was demonstrated in section 5, part (d) does not contribute to either the force or the torque.

The total force and torque exerted on the sphere, resulting from superposition of contributions (a)–(d), are thus, respectively, given by

$$\sum F = U f_{\rm tr} + \Omega f_{\rm rot} + (1-\gamma) f_{\rm el}, \tag{6.5}$$

$$\sum T = U g_{\rm tr} + \Omega g_{\rm rot} + (1-\gamma) g_{\rm el}. \tag{6.6}$$

Hence, the respective particle velocities of a freely suspended sphere are

$$U = (1-\gamma)\,\frac{f_{\rm rot}\, g_{\rm el} - g_{\rm rot}\, f_{\rm el}}{f_{\rm tr}\, g_{\rm rot} - f_{\rm rot}\, g_{\rm tr}}, \quad \Omega = (1-\gamma)\,\frac{g_{\rm tr}\, f_{\rm el} - f_{\rm tr}\, g_{\rm el}}{f_{\rm tr}\, g_{\rm rot} - f_{\rm rot}\, g_{\rm tr}}, \tag{6.7}$$

which, to leading order, yield

$$U \sim (1-\gamma)\, U^{(0)}(e)\left[1+O(\varepsilon)\right], \quad \Omega \sim (1-\gamma)\,\Omega^{(0)}(e)\left[1+O(\varepsilon)\right], \tag{6.8}$$

wherein

$$U^{(0)}(e) = \frac{f_{\rm rot}^{(0)}\, g_{\rm el}^{(0)} - g_{\rm rot}^{(0)} f_{\rm el}^{(0)}}{f_{\rm tr}^{(0)} g_{\rm rot}^{(0)} - f_{\rm rot}^{(0)}\, g_{\rm tr}^{(0)}}, \quad \Omega^{(0)}(e) = \frac{g_{\rm tr}^{(0)}\, f_{\rm el}^{(0)} - f_{\rm tr}^{(0)} g_{\rm el}^{(0)}}{f_{\rm tr}^{(0)} g_{\rm rot}^{(0)} - f_{\rm rot}^{(0)}\, g_{\rm tr}^{(0)}}. \tag{6.9}$$

These velocities reflect a balance between the electrical driving forces[4] and the retarding hydrodynamic forces arising from the fluid viscosity. It may appear surprising that despite the large $O(\varepsilon^{-5/2})$ drag acting on the sphere, the latter still moves with an $O(1)$ speed. This result obviously arises from the large $O(\varepsilon^{-1/2})$ potential drop across the gap, which has to be supplied by the external voltage supply.[5] Both sphere velocities are proportional to the potential difference $1-\gamma$, a result common to other types of sphere-wall geometries [9, 10, 23, 22, 21]. Indeed, were the zeta potentials of the sphere and wall equal (corresponding to the case $\gamma = 1$), the boundary-value problem (2.5)–(2.7) would possess the trivial irrotational solution, $\boldsymbol{v} = \nabla\varphi$, which results in neither a force nor a torque acting on the sphere.

Consider the dependence of $U^{(0)}$ upon e. Explicit expressions for $f_{\rm el}^{(0)}$ and $g_{\rm el}^{(0)}$ are given in (4.20) and (4.26), whereas those for the comparable translational and rotational coefficients are provided in [4] as

$$f_{\rm tr}^{(0)} = -\frac{9}{4}\pi^2\sqrt{2}\eta_0, \quad g_{\rm tr}^{(0)} = -\frac{3}{2}\pi^2\sqrt{2}e\eta_0, \quad g_{\rm rot}^{(0)} = -\pi^2\sqrt{2}e^2\eta_0 - 2\sqrt{2}\eta_3, \tag{6.10}$$

[4]The electric field acts directly on the electrically nonneutral fluid elements within the Debye layer. On the bulk scale, this action is manifested by the slip condition (2.6)–(2.7).

[5]This added potential drop constitutes, however, only a small portion of the applied voltage, the latter also maintaining an electric field along the "infinite" channel length.

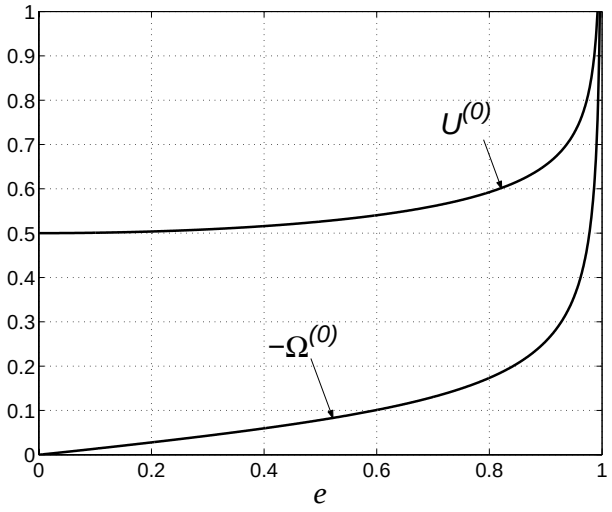


FIG. 6.1. *Variation with e of leading-order linear and angular velocities.*

and, obviously, $f_{\rm rot}^{(0)} = g_{\rm tr}^{(0)}$. In the above, the e-dependent parameter η_3 is defined as

$$\eta_3 = \frac{1}{\pi}\int_0^{2\pi} \frac{\cos^2\phi}{\tau(\phi)} d\phi, \tag{6.11}$$

whereas η_0 was defined in (4.17). These parameters are easily expressed in terms of complete elliptic functions [4]. Using similar means, it readily follows that

$$\mu_0(e) = \frac{\pi(1-m/2)^{1/2}}{2E\left(m^{1/2}\right)}, \tag{6.12}$$

wherein $m = 2e/(1+e)$, with E the complete elliptic integral of the second kind [6]. Note that $\mu_0(0) = 0$, whereas $\mu_0(1) = \pi/2\sqrt{2}$.

Using the limiting values of the functions η_0 and η_3 [4], one finds that $U^{(0)}(0) = 1/2$, corresponding to exactly half of the value of the electrophoretic sphere velocity in the absence of wall effects. This result is in agreement with the numerical trend observed by Keh and Chiou [12], who provide the value 0.50177 for the largest ratio of sphere-to-channel radius studied by them, namely $(1+\varepsilon)^{-1} = 0.999$. For a highly eccentric sphere, $1 - e \ll 1$, we find that

$$U^{(0)}(e) \sim \frac{\pi^2}{24\,(1-e^2)^{1/2}\left(\ln\frac{32}{1-e} - \frac{8}{3}\right)}, \tag{6.13}$$

representing only a very moderate divergence. (Recall, however, that the present calculation breaks down as the minimum separation distance within the gap becomes comparable with the Debye length.)

The variation of $U^{(0)}$ with e is presented in Figure 6.1. Also presented is the variation of the angular velocity $\Omega^{(0)}$: in the axisymmetric case $e = 0$ this velocity vanishes, whereas for $1 - e \ll 1$ it becomes increasingly large in magnitude, with $\Omega^{(0)}/U^{(0)} \to -3/2$. (The cross-over of the $U^{(0)}$ and $\Omega^{(0)}$ curves occurs for e very close to unity, and is not shown in the figure.)

Acknowledgments. We thank Dr. Sangtae Kim of Eli Lilly and Company for his encouragement of this work.

REFERENCES

[1] P. M. Adler, *Influence of colloidal forces on a closely-fitting sphere in a fluid-filled tube*, PCH, 4 (1983), pp. 1–10.
[2] H. Brenner, *The Stokes resistance of an arbitrary particle—*IV. *Arbitrary fields of flow*, Chem. Eng. Sci., 19 (1964), pp. 703–727.
[3] P. M. Bungay, *The Motion of Closely-Fitting Particles Through Fluid-Filled Tubes*, Ph.D. thesis, Department of Chemical Engineering, Carnegie-Mellon University, Pittsburgh, PA, 1971.
[4] P. M. Bungay and H. Brenner, *The motion of a closely-fitting sphere in a fluid-filled tube*, Int. J. Multiphase Flow, 1 (1973), pp. 25–56.
[5] A. J. Goldman, R. Cox, and H. Brenner, *Slow viscous motion of a sphere parallel to a plane wall—*II. *Couette flow*, Chem. Eng. Sci., 22 (1967), pp. 663–660.
[6] I. S. Gradshteyn and I. M. Rhyzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, London, Toronto, 1980.
[7] J. Han, S. W. Turner, and H. G. Craighead, *Entropic trapping and escape of long DNA molcules at submicron size constriction*, Phys. Rev. Lett., 83 (2001), pp. 1688–1691.
[8] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics*, Prentice–Hall, Englewood Cliffs, NJ, 1965.
[9] H. J. Keh and J. L. Anderson, *Boundary effects on electrophoretic motion of colloidal spheres*, J. Fluid Mech., 153 (1985), pp. 417–439.
[10] H. J. Keh and S. B. Chen, *Electrophoresis of a colloidal sphere parallel to a dielectric plane*, J. Fluid Mech., 194 (1988), pp. 377–390.
[11] H. J. Keh and S. B. Chen, *Particle interactions in electrophoresis* I. *Motion of two spheres along their line of centers*, J. Colloid Interface Sci., 138 (1989), pp. 542–555.
[12] H. J. Keh and J. Y. Chiou, *Electrophoresis of a colloidal sphere in a circular cylindrical pore*, AIChE J., 42 (1996), pp. 1397–1406.
[13] H. J. Keh and L. C. Lien, *Electrophoresis of a dielectric sphere normal to a large conducting plane*, J. Chinese Inst. Chem. Engng., 20 (1989), p. 283.
[14] H. J. Keh and L. C. Lien, *Electrophoresis of a colloidal sphere along the axis of a circular orifice or a circular disk*, J. Fluid Mech., 224 (1991), pp. 305–333.
[15] M. Loewenberg and R. H. Davis, *Near-contact electrophoretic particle motion*, J. Fluid Mech., 288 (1995), pp. 103–122.
[16] L. D. Reed and F. A. Morrison, *Hydrodynamic interactions in electrophoresis*, J. Colloid Interface Sci., 54 (1976), pp. 117–133.
[17] W. B. Russel, D. A. Saville, and W. R. Schowalter, *Colloidal Dispersions*, Cambridge University Press, Cambridge, UK, 1989.
[18] A. Sellier, *Electrophoretic motion of two solid particles embedded in an unbounded and viscous electrolyte*, Comput. Mech., 28 (2002), pp. 202–211.
[19] Y. Solomentsev, D. Velegol, and J. Anderson, *Conduction in the small gap between two spheres*, Phys. Fluids, 9 (1997), pp. 1209–1217.
[20] H. A. Stone and S. Kim, *Microfluidics: Basic issues, applications, and challenges*, AIChE J., 47 (2001), pp. 1250–1254.
[21] E. Yariv and H. Brenner, *The electrophoretic mobility of an eccentrically-positioned spherical particle in a cylindrical pore*, Phys. Fluids, 14 (2002), pp. 3354–3357.
[22] E. Yariv and H. Brenner, *Near-contact electrophoretic motion of a sphere parallel to a planar wall*, J. Fluid. Mech., 484 (2003), pp. 85–111.
[23] E. Yariv, H. Brenner, and S. Kim, *Curvature induced dispersion in electroosmotic microfluidic flows*, SIAM J. Appl. Math., to appear.
[24] S. Zeng, A. Z. Zinchenko, and R. H. Davis, *Electrophoretic motion of two interacting particles*, J. Colloid Interface Sci., 209 (1999), pp. 282–301.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 442–467

A SECOND ORDER SHAPE OPTIMIZATION APPROACH FOR IMAGE SEGMENTATION*

MICHAEL HINTERMÜLLER† AND WOLFGANG RING†

Abstract. The problem of segmentation of a given image using the active contour technique is considered. An abstract calculus to find appropriate speed functions for active contour models in image segmentation or related problems based on variational principles is presented. The speed method from shape sensitivity analysis is used to derive speed functions which correspond to gradient or Newton-type directions for the underlying optimization problem. The Newton-type speed function is found by solving an elliptic problem on the current active contour in every time step. Numerical experiments comparing the classical gradient method with Newton's method are presented.

Key words. segmentation, active contours, level set method, shape sensitivity analysis, Newton's method

AMS subject classifications. 68U10, 49Q10, 49Q12

DOI. 10.1137/S0036139902403901

1. Introduction.

Identifying curve-like objects in images is one of the fundamental tasks in image analysis. In image segmentation we are interested in finding boundary curves for regions with approximately constant color or gray values. These curves usually represent boundaries of objects in the image. Image segmentation is therefore often the starting point for the treatment of other, more involved problems in image analysis such as automatic object recognition or image registration.

In recent years, image segmentation has been greatly influenced by two different ideas. On the one hand, global energy principles which should be satisfied for the optimal contour have been introduced and successfully applied. On the other hand, deformable (active) contours, which are represented as zero level sets of a time-dependent function $u : \mathbb{R}^2 \times [0,T] \to \mathbb{R}$, have been used to describe the geometric variable. Kass, Witkin, and Terzopoulos [22] introduced parametrized curves (now referred to as classical snakes) which evolve in such a way that the sum of an internal energy, comprising an elasticity and a rigidity term, and an external energy, indicating the presence of edges in the image, is minimized. Caselles, et al. [6] introduced a geometrically intrinsic (parametrization-independent) formulation of active contours, treating the propagating curve as the zero level set of a function $u : \mathbb{R}^2 \times [0,T] \to \mathbb{R}$. The propagation of the level set function u is driven by an appropriate speed function $F : \mathbb{R}^2 \to \mathbb{R}$, which occurs in the level set equation

$$u_t + F\,|\nabla u| = 0 \text{ on } \mathbb{R}^2. \tag{1.1}$$

The speed function F proposed in [6] is given by

$$F = g\left(\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right) + \nu\right), \tag{1.2}$$

*Received by the editors March 11, 2002; accepted for publication (in revised form) December 20, 2002; published electronically December 31, 2003.
http://www.siam.org/journals/siap/64-2/40390.html
†Special Research Center on Optimization and Control, Institute of Mathematics, University of Graz, Heinrichstrasse 36, A-8010 Graz, Austria (michael.hintermueller@uni-graz.at, wolfgang.ring@kfunigraz.ac.at).

where $g : \mathbb{R}^2 \to \mathbb{R}$ is an edge detector and ν is a constant. The edge detector is chosen in such a way that $g = 0$ at ideal edges of the image and $g > 0$ otherwise. The construction of the speed function F is such that the active contour propagates according to a curve-shortening mean curvature flow (see, e.g., [19]) with an additional constant deflation velocity ν. The motion of the curve is stopped at points which are located on (strong enough) edges where $g \sim 0$. Thus, g functions as a stopping criterion.

Several authors (see, e.g., [8, 26, 27, 37, 30]) have observed that a similar speed function, given by

$$F = \operatorname{div}\left(g\frac{\nabla u}{|\nabla u|}\right) = g \operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right) + \frac{1}{|\nabla u|}\langle \nabla g, \nabla u\rangle, \tag{1.3}$$

can be interpreted as the gradient direction for the cost functional

$$J(\Gamma) = \int_\Gamma g\, dS \tag{1.4}$$

with respect to the contour Γ, where S denotes the arclength measure on Γ. The flow in the negative gradient direction with respect to the cost functional (1.4) can therefore be considered as a geodesic flow with respect to the Riemannian metric $g : \mathbb{R}^2 \to \mathbb{R}$. Thus, the intrinsic curve propagation (1.1) with speed function (1.3) can also be derived from variational principles. In fact, it can be proved (see [8, 4]) that minimizing the classical snake model and the geodesic model (1.4) are (in some sense) equivalent.

It has turned out to be useful to add a domain integral term to the cost functional (1.4) to speed up the propagation. A corresponding cost functional has the form

$$J(\Gamma) = \int_\Gamma g\, dS + \nu \int_\Omega g\, d\mathbf{x}. \tag{1.5}$$

Many variants of the speed functions (1.2) and (1.3) or their corresponding variational principle have been considered in the literature, including affine invariant geodesic flow [30, 29], generalizations to three-dimensional situations [25, 9], region-based active contours [31, 20, 21], segmentation of moving objects [7], and active contour models based on the Mumford–Shah functional [13, 11, 12, 10]. We also refer to the recent monographs [5, 32], which treat the image segmentation problem extensively and provide numerous references to further literature on the subject.

Usually, in the image processing literature, parametrized contours and methods from classical calculus of variations (see [8, 37, 4]) are used to derive the Euler–Lagrange equation for a cost functional of type (1.4), even if the propagation of the contour is treated in the intrinsic level set formulation. We propose (and advertise) the use of an alternative technique with which we can calculate sensitivities with respect to geometric variables on a purely intrinsic basis. We shall use the speed method, which is commonly applied for the sensitivity analysis of shape optimization problems but, as it seems, is not very well known in the image processing community. The speed method has several advantages over the use of parametrized curves (or surfaces). It is intrinsic, i.e., independent of the chosen parametrization, it can treat the case where the current contour consists of several disjoint closed curves in a unified way, and it has (via the Hadamard–Zolésio structure theorem [15, sect. 3.3, p. 348]) a very natural link to the level set formulation of curve propagation. We also stress

the fact that the Euler–Lagrange equations for many of the cost functionals discussed in the references, which we listed in the previous paragraph, can be easily derived using the speed method. Application of Lemmas 1 and 3 will do the job for most cases. Most of the results related to shape sensitivity calculus (with the exception of the usage of the shape derivative of the signed distance function) can be found in the book by Sokołowski and Zolésio [36] and in the new book by Delfour and Zolésio [15]. The advantage of utilizing shape sensitivity analysis in combination with the level set method as motivated above was previously observed in [24] in the context of inverse problems.

We also want to stress the nature of the segmentation problem as a (nonlinear) optimization problem. It is our goal to find the optimal contour in the least possible number of time steps and to achieve maximal descent in each individual step. This objective is quite different from aiming for a smooth propagation of a contour, which is often the focus of attention for level set–based propagating interface problems. For this reason, we propose applying and adapting ideas from nonlinear programming to active contour propagation. In the following we shall employ line search methods and preconditioning of the gradient direction. To realize the latter idea, we calculate a Newton-type speed function for the level set formulation of the variational problem (1.4). It turns out that the calculation of the Newton-type speed function involves the solution of an elliptic equation on the active contour Γ. That is, we have to track the zero level set at every step of the propagation, and we have to assemble appropriate (geometry-dependent) stiffness and mass matrices. This implementational and computational effort is repaid by a significant reduction of the number of iterations.

The structure of the paper is the following. In section 2, we recall basic facts and formulas from shape sensitivity analysis. Section 3 deals with the calculation of certain useful identities concerning the shape derivative of the signed distance function of a smooth open domain. These identities will prove to be very helpful in section 5. In section 4, we explain how the gradient of a shape functional of the form (1.4) can be interpreted as a normal vector field to the boundary of the current shape via the Hadamard–Zolésio structure theorem and how a connection to the level set formulation can be drawn. In this section, we also introduce the concept of second order (Newton-type) preconditioning of the shape gradient. Section 5 deals with the calculation of the Newton direction for a shape functional of type (1.4). In this context it turns out that, if we restrict our consideration to a certain class of possible speed functions, we get a symmetric shape Hessian, which depends only on intrinsic properties of the current contour. Moreover, this restricted class of speed functions has desirable properties for the stable propagation of the level set function according to the level set equation (1.1). In section 6, we derive the elliptic equation on the actual contour which defines the Newton-type speed function. This equation involves the Laplace–Beltrami operator. The Newton-type algorithm and its numerical realization are the subject of section 7. Also, a numerical technique for relaxing the CFL-condition for the time step size in the level set equation is considered. Finally, in section 8, a report on numerical test runs of the new algorithm and a comparison with the method based on the negative gradient as the speed function are given.

2. Shape sensitivity analysis via the speed method. We briefly recall the speed method from shape optimization, which can be used to calculate sensitivities of a functional with respect to a geometric variable such as a domain or the boundary of an open domain. Our main references for this section are the books [15, 36]. This

section also contains basic facts from tangential calculus on smooth boundaries of open sets.

In image analysis, sensitivities with respect to geometric objects such as contours are usually calculated using parametrized curves and techniques from classical calculus of variations as, e.g., in [8, 37]. We now present a technique for sensitivity analysis which works on boundaries of open sets instead of parametrized curves.

Let $\Gamma = \partial\Omega$ be the boundary of an open set $\Omega \subset \mathbb{R}^2$. We call such a boundary Γ a *contour* in $\mathbb{R}^2$. Suppose $V : \mathbb{R}^2 \to \mathbb{R}^2$ is a given smooth vector field with compact support in $\mathbb{R}^2$. We consider the initial value problem

$$\begin{cases} \mathbf{X}'(t) = V\big(\mathbf{X}(t)\big), \\ \mathbf{X}(0) = \mathbf{x}, \end{cases} \tag{2.1}$$

with $\mathbf{x} \in \mathbb{R}^2$ given. The flow (or time-t map) with respect to V is defined as the mapping $T_t : \mathbb{R}^2 \to \mathbb{R}^2$, with

$$T_t(\mathbf{x}) = \mathbf{X}(t), \tag{2.2}$$

where $\mathbf{X}(t)$ is the solution to (2.1) at time t. If Γ is a contour, we define

$$\Gamma_t = \{T_t(\mathbf{x}) \,:\, \mathbf{x} \in \Gamma\} = T_t(\Gamma). \tag{2.3}$$

In an analogous way, we define $\Omega_t = T_t(\Omega)$ for an arbitrary open set Ω. Note that, if $V \in \mathcal{C}_0^k(\mathbb{R}^2, \mathbb{R}^2)$, then $T_t \in \mathcal{C}^k(\mathbb{R}^2, \mathbb{R}^2)$; thus, smoothness properties of Γ are inherited by Γ_t, provided that the vector field V is smooth enough.

Suppose we are given a functional $J : G \to \mathbb{R}$, where G is an appropriate set of contours. We define the *Eulerian derivative* of J at a contour Γ in the direction of a perturbation vector field V by

$$dJ(\Gamma; V) = \lim_{t \downarrow 0} \frac{1}{t}\Big(J(\Gamma_t) - J(\Gamma)\Big). \tag{2.4}$$

Let B be a Banach space of perturbation vector fields. We say that the functional J is shape differentiable at Γ in B if $dJ(\Gamma; V)$ exists for all $V \in B$ and the mapping $V \mapsto dJ(\Gamma; V)$ is linear and continuous on B. We use the analogous definition for functionals $J(\Omega)$ which depend on an open set Ω as an independent variable instead of on a contour Γ.

We now present a series of lemmas which cover some results from shape calculus which will become useful later on. We start with the Eulerian derivative of a domain integral.

LEMMA 1. *Suppose $\phi \in W^{1,1}_{\mathrm{loc}}(\mathbb{R}^2)$ and $\Omega \subset \mathbb{R}^2$ is open and bounded. Then, the functional*

$$J(\Omega) = \int_\Omega \phi \, d\mathbf{x}$$

is shape differentiable for perturbation vector fields $V \in \mathcal{C}_0^1(\mathbb{R}^2; \mathbb{R}^2)$. The Eulerian derivative of J is given by

$$dJ(\Omega; V) = \int_\Omega \operatorname{div}(\phi\, V)\, d\mathbf{x}. \tag{2.5a}$$

If $\Gamma = \partial\Omega$ is of class $\mathcal{C}^1$, then

$$dJ(\Omega; V) = \int_\Gamma \phi \, \langle V, \mathbf{n} \rangle \, dS, \tag{2.5b}$$

where $\mathbf{n}$ denotes the exterior unit normal vector to Ω, $\langle \cdot, \cdot \rangle$ the inner product on $\mathbb{R}^2$, and dS the arclength measure on Γ.

Proof. See Propositions 2.45 and 2.46 in [36, p. 77]. ∎

For a vector field $V \in \mathcal{C}_0^1(\mathbb{R}^2; \mathbb{R}^2)$ and an open set of class $\mathcal{C}^2$ with boundary Γ, we define the tangential divergence of V by

$$\operatorname{div}_\Gamma V = (\operatorname{div} V - \langle DV \cdot \mathbf{n}, \mathbf{n} \rangle)\big|_\Gamma, \tag{2.6}$$

where DV denotes the Jacobian matrix of V. If the vector field V is defined only on Γ, we can still define the tangential divergence of V as the tangential divergence of an arbitrary extension of V. It can be shown (cf. [36, Prop. 2.51, p. 82]) that the definition does not depend on the particular choice of the extension. With this, we are able to state the following result on boundary integrals.

LEMMA 2. *Suppose $\phi \in W^{2,1}_{\mathrm{loc}}(\mathbb{R}^2)$ and Γ is a contour of class $\mathcal{C}^1$. Then, the functional*

$$J(\Gamma) = \int_\Gamma \phi \, dS \tag{2.7}$$

is shape differentiable for perturbation vector fields $V \in \mathcal{C}_0^1(\mathbb{R}^2; \mathbb{R}^2)$ with

$$dJ(\Gamma; V) = \int_\Gamma \big(\langle \nabla \phi, V \rangle + \phi \operatorname{div}_\Gamma V \big) \, dS. \tag{2.8}$$

Proof. See sections 2.18 and 2.19 in [36]. ∎

Using tangential calculus (see sections 2.19 and 2.20 in [36] or the results in [16]), we can simplify the expression (2.8). We define the tangential gradient of a function $h \in \mathcal{C}^2(\Gamma)$ as

$$\nabla_\Gamma h = \nabla \tilde{h}\big|_\Gamma - \frac{\partial \tilde{h}}{\partial n} \mathbf{n} \tag{2.9}$$

on Γ, where $\tilde{h}$ denotes an arbitrary smooth extension of h. It can be shown that the definition (2.9) does not depend on the specific choice of the extension. We have the following Green's formula on Γ.

PROPOSITION 1 (Green's theorem on Γ). *Suppose Γ is a contour of class $\mathcal{C}^2$, $h \in \mathcal{C}^2(\Gamma)$, and $V \in \mathcal{C}_0^1(\mathbb{R}^2; \mathbb{R}^2)$ with $\langle V, \mathbf{n} \rangle = 0$ for every point $\mathbf{x} \in \Gamma$. Then, we have*

$$\int_\Gamma \langle \nabla_\Gamma h, V \rangle \, dS = - \int_\Gamma h \operatorname{div}_\Gamma V \, dS. \tag{2.10}$$

Remark 1. Green's formula also holds for functions h in the Sobolev space $H^1(\Gamma)$. In this case, (2.10) acts as a definition for the tangential gradient $\nabla_\Gamma h$.

Suppose we are given a smooth vector field V. We set $V_\tau = V - \langle V, \mathcal{N} \rangle \mathcal{N}$ as the tangential component of V with respect to Γ. Here $\mathcal{N}$ denotes an extension of the normal vector field $\mathbf{n}$ on Γ. We have

$$\begin{aligned}
\operatorname{div}_\Gamma V &= \operatorname{div}_\Gamma V_\tau + \operatorname{div}_\Gamma \big(\langle V, \mathcal{N} \rangle \mathcal{N} \big) \\
&= \operatorname{div}_\Gamma V_\tau + \big(\operatorname{div} \big(\langle V, \mathcal{N} \rangle \mathcal{N} \big) - \big\langle D\big(\langle V, \mathcal{N} \rangle \mathcal{N} \big) \cdot \mathcal{N}, \mathcal{N} \big\rangle \big)\big|_\Gamma \\
&= \operatorname{div}_\Gamma V_\tau + \big(\langle V, \mathcal{N} \rangle \big(\operatorname{div} \mathcal{N} - \langle D\mathcal{N} \cdot \mathcal{N}, \mathcal{N} \rangle \big) \big)\big|_\Gamma \\
&= \operatorname{div}_\Gamma V_\tau + \langle V, \mathcal{N} \rangle \operatorname{div}_\Gamma \mathcal{N}\big|_\Gamma.
\end{aligned}$$

The term $\operatorname{div}_\Gamma \mathcal{N}|_\Gamma$ is usually denoted by κ and is called the *curvature* of Γ. Thus, we find

$$\operatorname{div}_\Gamma V = \operatorname{div}_\Gamma V_\tau + \kappa \langle V, \mathbf{n} \rangle$$

on Γ. We thus obtain an equivalent expression for the Eulerian derivative of the cost functional (2.7). We have

$$\begin{aligned} dJ(\Gamma; V) &= \int_\Gamma \big(\langle \nabla \phi, V \rangle + \phi \operatorname{div}_\Gamma V \big)\, dS \\ &= \int_\Gamma \left(\left\langle \nabla_\Gamma \phi + \frac{\partial \phi}{\partial n}\, \mathbf{n}, V \right\rangle + \phi \operatorname{div}_\Gamma (V_\tau) + \phi\, \kappa\, \langle V, \mathbf{n} \rangle \right) dS \\ &= \int_\Gamma \Big(\frac{\partial \phi}{\partial n} + \phi\, \kappa \Big) \langle V, \mathbf{n} \rangle\, dS + \int_\Gamma \Big(\langle \nabla_\Gamma \phi, V_\tau \rangle + \phi \operatorname{div}_\Gamma (V_\tau) \Big)\, dS. \end{aligned}$$

The last integral is zero due to Proposition 1. We therefore obtain the following lemma.

LEMMA 3. *Under the assumptions of Lemma* 2 *the Eulerian derivative of the cost functional* (2.7) *is equivalently given by*

$$dJ(\Gamma; V) = \int_\Gamma \Big(\frac{\partial \phi}{\partial n} + \phi\, \kappa \Big) \langle V, \mathbf{n} \rangle\, dS. \tag{2.11}$$

It is also useful to be able to calculate sensitivities for more general functionals of the form

$$J(\Omega) = \int_\Omega \phi(\Omega, \mathbf{x})\, d\mathbf{x} \tag{2.12}$$

or

$$J(\Gamma) = \int_\Gamma \psi(\Gamma, \mathbf{x})\, dS(\mathbf{x}), \tag{2.13}$$

where the functions $\phi(\Omega) : \Omega \to \mathbb{R}$ and $\psi(\Gamma) : \Gamma \to \mathbb{R}$ themselves depend on the geometric variables Ω and Γ, respectively. In this case, formulas (2.5) and (2.11) have to be corrected by terms which take care of the derivatives of ϕ and ψ with respect to Ω or Γ. We define the following two variants of derivatives of a geometry-dependent function with respect to the geometry.

DEFINITION 1. *Suppose $\psi(\Gamma) \in B(\Gamma)$ for all $\Gamma \in G$, where $B(\Gamma)$ is some appropriate Banach space of functions on Γ, and let $V \in \mathcal{C}_0^1(\mathbb{R}^2, \mathbb{R}^2)$. We set $\psi^t = \psi(\Gamma_t) \circ T_t(V)$ and $\psi^0 = \psi(\Gamma)$, and we assume that $\psi^t \in B(\Gamma_t)$ for all $0 < t < T$ with some $T > 0$. If the limit*

$$\dot{\psi}(\Gamma; V) = \lim_{t \downarrow 0} \frac{1}{t} \big(\psi^t - \psi^0 \big) \tag{2.14}$$

exists in the strong (weak) topology on $B(\Gamma)$, then $\dot{\psi}(\Gamma; V)$ is called the strong (weak) material derivative *of ψ at Γ in direction V.*

The analogous definition holds for functions $\phi(\Omega)$ which are defined on open sets and not on contours.

The material derivative is the derivative of ϕ (or ψ) with respect to the geometry for a moving (Lagrangian) coordinate system. Let us first consider the case of a

domain function $\phi : \Omega \to \mathbb{R}$. It is easily seen that, for the special case where ϕ is independent of Ω, we find

$$\dot{\phi}(\Omega; V) = \dot{\phi}(V) = \langle \nabla\phi, V\rangle.$$

For a function which does not depend on Ω, any reasonable derivative with respect to Ω in a fixed (Eulerian) coordinate system must be 0. It is therefore natural to subtract the term $\langle \nabla\phi, V\rangle$ from $\dot{\phi}$ to define a derivative of ϕ with respect to Ω in a stationary coordinate system. This is the idea of the following definition.

DEFINITION 2. *Suppose that the weak material derivative $\dot{\phi}(\Omega; V)$ and the expression $\langle \nabla\phi(\Omega), V\rangle$ exist in $B(\Omega)$. Then, we set*

$$\phi'(\Omega; V) = \dot{\phi}(\Omega; V) - \langle \nabla\phi, V\rangle \tag{2.15}$$

and we call $\phi'(\Omega; V)$ the shape derivative *of ϕ at Ω in direction V.*

Note that

$$\phi'(\Omega; V) = \phi'(V) = 0$$

for any function ϕ which does not depend on Ω.

For boundary functions $\psi(\Gamma) : \Gamma \to \mathbb{R}$, the expression $\langle \nabla\psi, V\rangle$ does not make sense. In this case, we define the shape derivative as

$$\psi'(\Gamma; V) = \dot{\psi}(\Gamma; V) - \langle \nabla_\Gamma \psi, V\rangle\big|_\Gamma. \tag{2.16}$$

With these definitions we are able to calculate the Eulerian derivatives for the shape functionals (2.12) and (2.13).

PROPOSITION 2. *Suppose $\phi = \phi(\Omega)$ is given such that the weak L^1-material derivative $\dot{\phi}(\Omega; V)$ and the shape derivative $\phi'(\Omega; V) \in L^1(\Omega)$ exist. Then, the cost functional* (2.12) *is shape differentiable and we have*

$$dJ(\Omega; V) = \int_\Omega \phi'(\Omega; V)\, d\mathbf{x} + \int_\Gamma \phi\, \langle V, \mathbf{n}\rangle\, dS. \tag{2.17}$$

For boundary functions $\psi(\Gamma)$ we get, under the same technical assumptions for the cost functional (2.13)*,*

$$dJ(\Gamma; V) = \int_\Gamma \psi'(\Gamma; V)\, dS + \int_\Gamma \kappa\, \psi\, \langle V, \mathbf{n}\rangle\, dS. \tag{2.18}$$

If $\psi(\Gamma) = \phi(\Omega)\big|_\Gamma$, then we have

$$dJ(\Gamma; V) = \int_\Gamma \phi'(\Omega; V)\big|_\Gamma\, dS + \int_\Gamma \Big(\frac{\partial\phi}{\partial n} + \kappa\,\phi\Big)\, \langle V, \mathbf{n}\rangle\, dS. \tag{2.19}$$

Suppose that $\phi(\Omega)$ satisfies $\phi(\Omega)|_\Gamma = 0$ for all (admissible) domains Ω, and let $\vartheta \in \mathcal{D}(\mathbb{R}^2)$ be given. We define the cost functional

$$J_0(\Gamma) = \int_\Gamma \vartheta\, \phi(\Omega)\, dS = 0$$

for arbitrary Γ. Thus,

$$0 = dJ_0(\Gamma; V) = \int_\Gamma \vartheta\, \phi'(\Omega, V)\, dS + \int_\Gamma \frac{\partial}{\partial n}\Big(\vartheta\, \phi(\Omega)\Big)\, \langle V, \mathbf{n}\rangle\, dS.$$

If we choose ϑ such that

$$\frac{\partial \vartheta}{\partial n} = 0 \quad \text{on } \Gamma \tag{2.20}$$

and if we use the fact that the set of test functions which satisfy (2.20) is dense in $L^2(\Gamma)$, we get

$$\phi'(\Omega; V)\big|_\Gamma = -\frac{\partial \phi}{\partial n} \langle V, \mathbf{n}\rangle\big|_\Gamma$$

on Γ. We have therefore proved the following lemma.

LEMMA 4. *Suppose that $\phi(\Omega) \in H^{\frac{3}{2}+\epsilon}(\Omega)$ satisfies $\phi(\Omega)|_\Gamma = 0$ for all (admissible) domains Ω and that the shape derivative $\phi'(\Omega; V)$ exists in $H^{\frac{1}{2}+\epsilon}(\Omega)$ for some $\epsilon > 0$. Then, we have*

$$\phi'(\Omega; V)\big|_\Gamma = -\frac{\partial \phi}{\partial n} \langle V, \mathbf{n}\rangle\big|_\Gamma. \tag{2.21}$$

Remark 2. The Hadamard–Zolésio structure theorem [15, Thm. 3.6 and Cor. 1, p. 348f] states that the Eulerian derivative of a domain or boundary functional always has a representation of the form

$$dJ(\Omega; V) = \big\langle G, \langle V, \mathbf{n}\rangle \big\rangle_{\mathcal{C}^{-k}(\Gamma), \mathcal{C}^k(\Gamma)} = \big\langle G\,\mathbf{n}, V \big\rangle_{\mathcal{C}_2^{-k}(\Gamma), \mathcal{C}_2^k(\Gamma)}; \tag{2.22}$$

that is, the Eulerian derivative is concentrated on Γ and can be identified with the *normal vector field* $G\,\mathbf{n}$ on Γ. We set

$$D_\Gamma J(\Omega) = G\,\mathbf{n}, \tag{2.23}$$

and we call this expression the *shape gradient* of J at Ω.

3. Shape derivative of the signed distance function. The signed (or oriented) distance function is a useful tool in shape analysis. Many differential geometric quantities such as the normal vector field of a contour Γ or its curvature can be easily expressed in terms of the signed distance function b_Γ of Γ. We shall now apply the techniques introduced in the previous section to calculate the shape derivative of the signed distance function of a given (open, bounded) set Ω. This will be helpful later on when we have to calculate Eulerian derivatives of functionals which depend also on geometric properties of Γ such as normal direction or curvature. The following definitions and facts are taken from [15, Chap. 5]. The *distance function* d_A of a subset $A \subset \mathbb{R}^2$ is defined as

$$d_A(\mathbf{x}) = \inf_{\mathbf{y}\in A} |\mathbf{y} - \mathbf{x}|. \tag{3.1}$$

The *signed distance function* b_Ω of a bounded open set $\Omega \subset \mathbb{R}^2$ is defined as

$$b_\Omega(\mathbf{x}) = d_\Omega(\mathbf{x}) - d_{\mathbb{R}^2\setminus\Omega}(\mathbf{x}). \tag{3.2}$$

If we set $\Gamma = \partial\Omega$, we can express d_Ω in terms of Γ. We have

$$b_\Omega(\mathbf{x}) = \begin{cases} d_\Gamma(\mathbf{x}) & \text{for } \mathbf{x} \in \operatorname{int}(\mathbb{R}^2 \setminus \Omega), \\ 0 & \text{for } \mathbf{x} \in \Gamma, \\ -d_\Gamma(\mathbf{x}) & \text{for } \mathbf{x} \in \Omega. \end{cases} \tag{3.3}$$

We shall use the notation $b_\Gamma = b_\Omega$. Note in particular that

$$b_\Gamma\big|_\Gamma = 0. \tag{3.4}$$

It can be shown that b_Γ is uniformly Lipschitz continuous on $\mathbb{R}^2$ and hence, by Rademacher's theorem, differentiable a.e. on $\mathbb{R}^2$ with $|\nabla b_\Gamma| = 1$ a.e. on $\mathbb{R}^2 \setminus \Gamma$. If $\mathrm{meas}(\Gamma) = 0$, then we have

$$|\nabla b_\Gamma|^2 = 1 \quad \text{a.e. on } \mathbb{R}^2. \tag{3.5}$$

If Γ is smooth and compact ($C^{1,1}$ is enough), then ∇b_Γ is Lipschitz continuous, and we have $\nabla b_\Gamma(\mathbf{x}) = \mathbf{n}(p_\Gamma(\mathbf{x}))$ for all $\mathbf{x}$ in some neighborhood of Γ, where p_Γ denotes the projection onto Γ. Thus, ∇b_Γ can be considered as an extension of the unit normal vector field $\mathbf{n}$ onto a neighborhood of Γ, and we have

$$\nabla b_\Gamma\big|_\Gamma = \mathbf{n}. \tag{3.6}$$

Moreover, the second fundamental form of Γ can be expressed in terms of b_Γ. For a C^2-submanifold $\Gamma \subset \mathbb{R}^2$ we have

$$\Delta b_\Gamma\big|_\Gamma = \kappa. \tag{3.7}$$

See [15, p. 369] for the last relation. Taking the gradient on both sides of the Eikonal equation (3.5) yields

$$D^2 b_\Gamma \cdot \nabla b_\Gamma = 0 \quad \text{on } \Gamma. \tag{3.8}$$

Let $W \in \mathcal{C}_0^1(\mathbb{R}^2, \mathbb{R}^2)$ be a given perturbation vector field. We shall derive certain properties of the shape derivative $b'_\Gamma = b'_\Gamma(\Gamma; W)$ of the signed distance function. The signed distance function satisfies the Eikonal equation (3.5) together with the boundary condition (3.4). The weak form of (3.5) is given by

$$\int_{\mathbb{R}^2} |\nabla b_\Gamma|^2 \, \psi \, d\mathbf{x} = \int_{\mathbb{R}^2} \psi \, d\mathbf{x} \tag{3.9}$$

for all test functions $\psi \in \mathcal{D}(\mathbb{R}^2)$. Taking the Eulerian derivative on both sides of (3.9) and using (2.17), we get

$$2 \int_{\mathbb{R}^2} \langle \nabla b'_\Gamma, \nabla b_\Gamma \rangle \, \psi \, d\mathbf{x} = 0 \tag{3.10}$$

for all $\psi \in \mathcal{D}(\mathbb{R}^2)$. Note that, since the functional (3.9) is defined on a fixed domain and depends on Γ only via b_Γ in the integral, the boundary term in (2.17) vanishes. This can be see by writing

$$\int_{\mathbb{R}^2} |\nabla b_\Gamma|^2 \, \psi \, d\mathbf{x} = \int_\Omega |\nabla b_\Gamma|^2 \, \psi \, d\mathbf{x} + \int_{\mathbb{R}^2 \setminus \Omega} |\nabla b_\Gamma|^2 \, \psi \, d\mathbf{x}$$

and applying (2.17) to both terms on the right-hand side. The boundary integrals from both contributions sum up to zero. Equation (3.10) implies that

$$\langle \nabla b'_\Gamma, \nabla b_\Gamma \rangle = 0. \tag{3.11}$$

Equation (3.11) holds at least on some neighborhood of Γ on which b_Γ is smooth enough to guarantee the existence of a (weak) material derivative and hence the applicability of (2.17).

If we apply Lemma 4 to b_Γ and use (3.5), we get $b'_\Gamma|_\Gamma = -\frac{\partial b_\Gamma}{\partial n}\langle V, \mathbf{n}\rangle|_\Gamma = -\langle \nabla b_\Gamma, \nabla b_\Gamma\rangle \cdot \langle V, \mathbf{n}\rangle = -\langle V, \mathbf{n}\rangle$ by (3.5). With $v_n = \langle V, \mathbf{n}\rangle$, we obtain

$$b'_\Gamma\big|_\Gamma = -v_n. \tag{3.12}$$

Since $\nabla b'_\Gamma$ is orthogonal to $\mathbf{n}$ by (3.11), we have

$$\nabla b'_\Gamma = \nabla_\Gamma b'_\Gamma = -\nabla_\Gamma v_n \text{ on } \Gamma. \tag{3.13}$$

Moreover, we have $\Delta b'_\Gamma\big|_\Gamma = \operatorname{div}(\nabla b'_\Gamma)\big|_\Gamma = \operatorname{div}_\Gamma(\nabla_\Gamma b'_\Gamma) + \langle D^2 b'_\Gamma \cdot \nabla b_\Gamma, \nabla b_\Gamma\rangle\big|_\Gamma$. Because $0 = \nabla\langle \nabla b'_\Gamma, \nabla b_\Gamma\rangle = D^2 b'_\Gamma \cdot \nabla b_\Gamma + D^2 b_\Gamma \cdot \nabla b'_\Gamma$ and since $D^2 b_\Gamma$ is symmetric, we can conclude that $\langle D^2 b'_\Gamma \cdot \nabla b_\Gamma, \nabla b_\Gamma\rangle = -\langle D^2 b_\Gamma \cdot \nabla b'_\Gamma, \nabla b_\Gamma\rangle = \langle \nabla b'_\Gamma, D^2 b_\Gamma \cdot \nabla b_\Gamma\rangle = 0$ due to (3.8). Therefore, we obtain

$$\Delta b'_\Gamma\big|_\Gamma = -\Delta_\Gamma v_n. \tag{3.14}$$

4. Gradient and Newton-type level set flow for a shape optimization problem. For the numerical solution of a shape optimization problem one can use shape sensitivity information to move the geometric variable step by step in the direction of the negative gradient. Alternatively, one can use some other descent direction, which can be obtained from the gradient by applying a Newton-type preconditioner to the negative gradient. Like the shape gradient, the descent direction should have the form of a normal vector field on Γ (see Remark 2). Suppose $F\,\mathbf{n}$ is such a descent direction, where $F : \Gamma \to \mathbb{R}$ is a scalar function which depends on Γ. If we embed the discrete iterative optimization procedure in a continuous flow $\Gamma(t)$ which propagates in direction $F\,\mathbf{n}$, we get the following propagating front formulation for $\Gamma(t)$:

$$\dot{\mathbf{x}}(t) = F\big(\mathbf{x}(t), \Gamma(t)\big)\,\mathbf{n}(\mathbf{x}(t)) \quad \text{for } \mathbf{x}(t) \in \Gamma(t). \tag{4.1}$$

An equivalent formulation is given by the level set equation

$$u_t + \tilde{F}|\nabla u| = 0 \quad \text{on } \mathbb{R}^2 \times (0, T), \tag{4.2}$$

where the propagating front is the zero level set of the function u, i.e.,

$$\Gamma(t) = \{\mathbf{x} \in \mathbb{R}^2 \,:\, u(\mathbf{x}, t) = 0\}. \tag{4.3}$$

In (4.2), the scalar function $\tilde{F} : \mathbb{R}^2 \times [0, T) \to \mathbb{R}$ is chosen such that $\tilde{F}|_{\Gamma(t)} = F(\Gamma(t))$. See [35] for an extensive exposition of propagating front problems and their analytical and numerical treatment in the level set context. Note that by the Hadamard–Zolésio structure theorem (see Remark 2), the shape gradient $D_\Gamma J$ can always be interpreted as a scalar speed function G on Γ, which can be used in the level set formulation. Thus, shape sensitivity analysis and the level set method can be combined in a very natural way.

We want to use a speed function $F : \Gamma \to \mathbf{R}$, which represents a Newton-type descent direction for the shape optimization problem (1.4). This function is determined in the following way. Let $F : \Gamma \to \mathbb{R}$ and $G : \Gamma \to \mathbb{R}$ be given functions. We now establish a one-to-one correspondence between scalar speed functions and a

certain class of perturbation vector fields. Let $\tilde{F}$ and $\tilde{G}$ denote extensions of F and G, respectively, which are constructed as solutions to the transport equations

$$\langle \nabla \tilde{F}, \nabla b_\Gamma \rangle = 0 \quad \text{on } \mathbb{R}^2, \quad \tilde{F}\big|_\Gamma = F, \tag{4.4}$$

and

$$\langle \nabla \tilde{G}, \nabla b_\Gamma \rangle = 0 \quad \text{on } \mathbb{R}^2, \quad \tilde{G}\big|_\Gamma = G. \tag{4.5}$$

Note that Γ is noncharacteristic with respect to the transport equation; thus, (4.4) and (4.5) have unique solutions, at least locally in some neighborhood of Γ, which is small enough such that the characteristics of (4.4) (which are straight lines) do not intersect. With these solutions, we define the vector fields

$$V_F = \tilde{F}\,\nabla b_\Gamma \quad \text{and} \quad V_G = \tilde{G}\,\nabla b_\Gamma \tag{4.6}$$

on some neighborhood of Γ on which $\tilde{F}$, $\tilde{G}$, and ∇b_Γ are smooth. Outside this neighborhood we assume that V_F and V_G are extended in some smooth way. Note that the construction of V_F and V_G is such that

$$\langle V_F, \mathbf{n} \rangle = F \text{ and } \langle V_G, \mathbf{n} \rangle = G \quad \text{on } \Gamma. \tag{4.7}$$

Now we consider a cost functional of type (1.4). Let $d^2J(\Gamma;V;W) = d(dJ(\Omega;V))(\Omega;W)$ be the second Eulerian derivative of the cost functional (1.4). In general, the second Eulerian derivative is not symmetric in the two arguments V and W and does not depend only on $V|_\Gamma$ and $W|_\Gamma$. From the subsequent computation we shall see, however, that for perturbation vector fields of the form (4.6), the second Eulerian derivative is symmetric in (V_F, V_G) and depends only on F and G.

We propose the following optimization algorithm. We define a Newton-type speed function $F : \Gamma \to \mathbf{R}$ as the solution to

$$d^2J(\Gamma; V_F; V_G) = -dJ(\Gamma; V_G) \quad \text{for all } G : \Gamma \to \mathbb{R}. \tag{4.8}$$

We then find the extension $\tilde{F}$ of F onto some neighborhood of Γ by solving the transport equation (4.4). Finally, we use $\tilde{F}$ as speed function for one time step in the level set equation

$$u_t + \tilde{F}|\nabla u| = 0. \tag{4.9}$$

The step size is chosen such that some line search criterion is satisfied. With the updated geometry, we start the procedure over again until some stopping criterion is reached.

5. Calculation of the Newton-type speed function. In this section, we calculate the second Eulerian derivative for shape functionals of the form

$$J_1(\Gamma) = \int_\Gamma \phi \, dS \tag{5.1}$$

and

$$J_2(\Omega) = \int_\Omega \phi \, d\mathbf{x} \tag{5.2}$$

for some fixed $\phi \in W^{2,1}_{loc}(\mathbb{R}^2)$. For the following calculations we assume that Γ is of class $\mathcal{C}^2$, which implies that $b_\Gamma \in \mathcal{C}^2$ on some neighborhood of Γ (see [15, Thm. 4.3, p. 219]).

When it is clear from the context, we omit "$|_\Gamma$." We start with the calculation for J_1. Using (2.11), (3.6), and (3.7), we obtain

$$\begin{aligned} dJ_1(\Gamma, V_F) &= \int_\Gamma \Big(\frac{\partial \phi}{\partial n} + \phi\, \kappa \Big) \, \langle V_F, \mathbf{n} \rangle \, dS \\ &= \int_\Gamma \big(\langle \nabla \phi, \nabla b_\Gamma \rangle + \phi\, \Delta b_\Gamma \big) \, \langle V_F, \nabla b_\Gamma \rangle \, dS. \end{aligned} \tag{5.3}$$

In this section, we consider only perturbation vector fields V_F, which satisfy (4.6). Note that $dJ_1(\Gamma; V_F)$ (for fixed V_F) is a shape functional of type (2.13). Therefore, the second Eulerian derivative can be calculated by applying formula (2.19) to $dJ_1(\Gamma; V_F)$. With $b'_\Gamma = b'_\Gamma(\Gamma; V_G)$ we get

$$\begin{aligned} d^2 J_1(\Gamma; V_F; V_G) &= \int_\Gamma \frac{\partial}{\partial n} \Big[\big(\langle \nabla \phi, \nabla b_\Gamma \rangle + \phi\, \Delta b_\Gamma \big) \, \langle V_F, \nabla b_\Gamma \rangle \Big] \, \langle V_G, \mathbf{n} \rangle \, dS \\ &\quad + \int_\Gamma \kappa \Big(\frac{\partial \phi}{\partial n} + \phi\, \kappa \Big) \, \langle V_F, \mathbf{n} \rangle \, \langle V_G, \mathbf{n} \rangle \, dS \\ &\quad + \int_\Gamma \big(\langle \nabla \phi, \nabla b'_\Gamma \rangle + \phi\, \Delta b'_\Gamma \big) \, \langle V_F, \mathbf{n} \rangle \, dS \\ &\quad + \int_\Gamma \Big(\frac{\partial \phi}{\partial n} + \phi\, \kappa \Big) \, \langle V_F, \nabla b'_\Gamma \rangle \, dS \\ &= I_1 + I_2 + I_3 + I_4. \end{aligned}$$

Using (4.7), (3.13), (3.14), and Green's formula (2.10), the integral I_3 simplifies to

$$\begin{aligned} I_3 &= - \int_\Gamma \big(\langle \nabla \phi, \nabla_\Gamma G \rangle + \phi \Delta_\Gamma G \big) \, F \, dS \\ &= - \int_\Gamma \big(\langle \nabla \phi, \nabla_\Gamma G \rangle \, F - \langle \nabla_\Gamma (\phi\, F), \nabla_\Gamma G \rangle \big) \, dS \\ &= \int_\Gamma \phi \, \langle \nabla_\Gamma F, \nabla_\Gamma G \rangle \, dS. \end{aligned} \tag{5.4}$$

Now let us consider I_1. We have

$$\begin{aligned} I_1 &= \int_\Gamma \Big(\frac{\partial^2 \phi}{\partial n^2} + \frac{\partial \phi}{\partial n}\, \kappa + \phi \langle \nabla (\Delta b_\Gamma), \nabla b_\Gamma \rangle \Big) \, F \, G \, dS \\ &\quad + \int_\Gamma \Big(\frac{\partial \phi}{\partial n} + \kappa\, \phi \Big) \, \frac{\partial}{\partial n} \langle V_F, \nabla b_\Gamma \rangle \, G \, dS \\ &= K_1 + K_2. \end{aligned}$$

From (3.5), we conclude

$$0 = \Delta \langle \nabla b_\Gamma, \nabla b_\Gamma \rangle = 2 \langle \nabla (\Delta b_\Gamma), \nabla b_\Gamma \rangle + 2 D^2 b_\Gamma : D^2 b_\Gamma,$$

where $A : B = \sum_{i,j} a_{i,j}\, b_{i,j}$ denotes the tensor product of matrices $A = (a_{i,j})$ and $B = (b_{i,j})$. Thus,

$$\langle \nabla (\Delta b_\Gamma), \nabla b_\Gamma \rangle = - \| D^2 b_\Gamma \|_F^2,$$

where $\| \cdot \|_F$ denotes the Frobenius norm of a matrix.

In I_2 we find a term of the form $\phi\,\kappa^2\,F\,G$. We have $\kappa^2 = (\mathrm{trace}(D^2 b_\Gamma))^2$. In two dimensions, the relation

$$(\mathrm{trace} D^2 b_\Gamma)^2 - \|D^2 b_\Gamma\|_F^2 = 2(b_\Gamma)_{x_1,x_1}(b_\Gamma)_{x_2,x_2} - 2(b_\Gamma)_{x_1,x_2}^2 = 2\det(D^2 b_\Gamma) = 0$$

holds due to (3.8). With this, we obtain

$$I_2 + K_1 = \int_\Gamma \Big(\frac{\partial^2\phi}{\partial n^2} + 2\frac{\partial\phi}{\partial n}\,\kappa\Big)\,F\,G\,dS. \tag{5.5}$$

The remaining term is $K_2 + I_4$. We find

$$\begin{aligned} K_2 + I_4 &= \int_\Gamma \Big(\frac{\partial\phi}{\partial n} + \kappa\,\phi\Big)\,\Big(\frac{\partial}{\partial n}\langle V_F, \nabla b_\Gamma\rangle\,G - \langle V_F, \nabla_\Gamma G\rangle\Big)\,dS \\ &= \int_\Gamma \Big(\frac{\partial\phi}{\partial n} + \kappa\,\phi\Big)\,\Big(\frac{\partial}{\partial n}\langle V_F, \nabla b_\Gamma\rangle\,G - \langle V_F, \nabla_\Gamma G\rangle\Big)\,dS. \end{aligned}$$

For the second expression in the above integral, we obtain, using (4.7) and definition (2.9),

$$\langle V_F, \nabla_\Gamma G\rangle = \Big\langle V_F, \nabla\langle V_G, \nabla b_\Gamma\rangle - \frac{\partial}{\partial n}\langle V_G, \nabla b_\Gamma\rangle\,\mathbf{n}\Big\rangle.$$

Thus, we get

$$\begin{aligned} &K_2 + I_4 \\ &\quad = \int_\Gamma \Big(\frac{\partial\phi}{\partial n} + \kappa\,\phi\Big)\,\Big(\frac{\partial}{\partial n}\langle V_F, \nabla b_\Gamma\rangle\,G + \frac{\partial}{\partial n}\langle V_G, \nabla b_\Gamma\rangle\,F - \big\langle \nabla\langle V_G, \nabla b_\Gamma\rangle, V_F\big\rangle\Big)\,dS. \end{aligned} \tag{5.6}$$

Note that the first two terms $\frac{\partial}{\partial n}\langle V_F, \nabla b_\Gamma\rangle\,G + \frac{\partial}{\partial n}\langle V_G, \nabla b_\Gamma\rangle\,F$ in (5.6) are symmetric in V_F and V_G, but they cannot be determined just from the restrictions $F = \langle V_F, \mathbf{n}\rangle|_\Gamma$ and $G = \langle V_G, \mathbf{n}\rangle|_\Gamma$. The term $\langle \nabla\langle V_G, \nabla b_\Gamma\rangle, V_F\rangle$ has the same nonintrinsic behavior and is not even symmetric in V_F and V_G. Now let us assume that V_F and V_G satisfy (4.6) on some neighborhood of Γ. Using this assumption, together with (3.8), we get

$$\begin{aligned} \nabla\langle V_F, \nabla b_\Gamma\rangle &= DV_F\cdot\nabla b_\Gamma + D^2 b_\Gamma\cdot V_F = D(\tilde F\,\nabla b_\Gamma)\cdot\nabla b_\Gamma + \tilde F D^2 b_\Gamma\cdot\nabla b_\Gamma \\ &= \langle\nabla\tilde F, \nabla b_\Gamma\rangle\cdot\nabla b_\Gamma + 2\langle V_F, D^2\nabla b_\Gamma\cdot\nabla b_\Gamma\rangle = 0, \end{aligned}$$

hence

$$\frac{\partial}{\partial n}\langle V_F, \nabla b_\Gamma\rangle = \big\langle\nabla\langle V_F, \nabla b_\Gamma\rangle, \nabla b_\Gamma\big\rangle = 0,$$

and, with the same reasoning,

$$\frac{\partial}{\partial n}\langle V_G, \nabla b_\Gamma\rangle = 0.$$

Thus, if we restrict our attention to perturbation vector fields of the form (4.4)–(4.6), the nonintrinsic and asymmetric terms in $d^2 J_1(\Gamma; V_F, V_G)$ vanish.

Taking all intermediate results together, we obtain the following expression for the second Eulerian derivative of J_1:

$$d^2 J_1(\Gamma; V_F; V_G) = \int_\Gamma \Big[\Big(\frac{\partial^2\phi}{\partial n^2} + 2\frac{\partial\phi}{\partial n}\,\kappa\Big)\,F\,G + \phi\langle\nabla_\Gamma F, \nabla_\Gamma G\rangle\Big]\,dS. \tag{5.7}$$

For J_2 we obtain, using Lemma 1,

$$dJ_2(\Omega; V_F) = \int_\Gamma \phi \, \langle V_F, \nabla b_\Gamma \rangle \, dS. \tag{5.8}$$

If we apply (2.19) in Proposition 2 and Lemma 4, we get

$$\begin{aligned} d^2 J_2(\Omega; V_F, V_G) = &\int_\Gamma \Big(\frac{\partial \phi}{\partial n} + \kappa \phi \Big) \, \langle V_F, \mathbf{n} \rangle \, \langle V_G, \mathbf{n} \rangle \, dS \\ &+ \int_\Gamma \phi \, \Big(\frac{\partial}{\partial n} \langle V_F, \mathbf{n} \rangle \, G - \langle V_F, \nabla_\Gamma G \rangle \Big) \, dS. \end{aligned}$$

As in the discussion of expression (5.6), we find that the second integral is zero if V_F and V_G satisfy (4.6). We therefore get

$$d^2 J_2(\Omega; V_F; V_G) = \int_\Gamma \Big(\frac{\partial \phi}{\partial n} + \kappa \phi \Big) \, F \, G \, dS. \tag{5.9}$$

6. Gradient and Newton-type flow for variational image segmentation. In this section, we apply the results of sections 2 and 4 to cost functionals of the form (1.5). We consider a grayscale image given by its intensity map $I : \mathbb{R}^2 \to \mathbb{R}$, which assigns each point $\mathbf{x}$ its gray value $I(\mathbf{x}) \in \mathbb{R}$. For simplicity (to avoid special treatment of the boundary), we assume that the image is defined on all of $\mathbb{R}^2$. Let $\tilde{g} : [0, \infty) \to (0, \infty)$ be a given decreasing function which satisfies $\tilde{g}(r) \to 0$ as $r \to \infty$. The function $g_I(\mathbf{x}) = \tilde{g}(|\nabla I|(\mathbf{x}))$ acts as an edge detector in the sense that $g_I(\mathbf{x}) = 0$ if $\mathbf{x}$ lies on an ideal edge of I. In this paper, we use

$$g_I(\mathbf{x}) = \frac{1}{1 + \big(|\nabla I|(\mathbf{x})\big)^k} \text{ with } k = 1, 2. \tag{6.1}$$

To suppress the influence of noise, we replace ∇I in the above expression by a smoothed version $\nabla \hat{I}$. For the sake of simplicity, we use Gaussian smoothing, but other, more effective geometric smoothers (see, e.g., [3]) can be used as well. The method we propose also works for other edge detectors of the form $g_I : \mathbb{R}^2 \to \mathbb{R}$ for which $g_I \sim 0$ on edges and $g \sim c$ with $c > 0$ otherwise, provided that they satisfy the necessary smoothness requirements for performing shape sensitivity analysis of a functional of type (6.2) as exposed in sections 2–5.

Segmentation of an image is the task of partitioning a given image into disjoint parts of approximately constant gray value. Let Γ be the union of the boundaries of these homogeneous regions. Since homogeneous regions with different gray values are separated by edges, it is likely that the boundary Γ is located at points where the edge detector g_I has small values. In Figure 1 it is seen that the edges of the image coincide with the deep valley in the edge detector. This motivates the following variational approach. We seek the final segmenting contour Γ as the minimizer of the functional

$$J(\Gamma) = \int_\Gamma g_I \, dS + \nu \int_\Omega g_I \, d\mathbf{x}, \tag{6.2}$$

where $\Gamma = \partial\Omega$ and $\nu > 0$. We find the minimizer of (6.2) as the steady state of a family of propagating contours $\Gamma(t)$ which approach the minimal contour from the outside. Note that a contour of length zero (a point) is a global minimizer for (6.2). This,

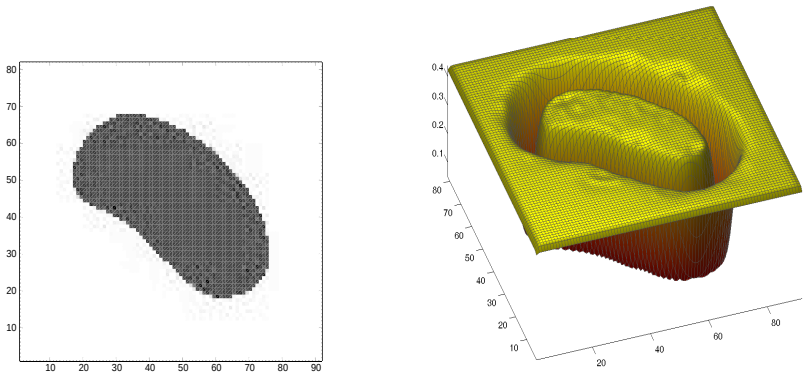


FIG. 1. *Grayscale image I and corresponding edge-map g.*

however, is not the desired segmenting contour. Rather, we want the propagation of the active contour to get stuck at the bottom of the valley of the edge detector, which is a *local minimizer* for (6.2). The second term in (6.2) is a regularization term, which helps shrink the active contour in the homogeneous regions where the influence of the edges is not very strong. The parameter ν must not be chosen too large, because otherwise the algorithm might overshoot the edge and end up at the global minimizer.

The Euler–Lagrange equation for the cost functional (6.2) with respect to the geometrical variable Γ was derived, e.g., in [8, 37] using parametrized curves $\Gamma = \Gamma(s,t)$, where s denotes the curve parameter and t is a time variable describing the movement of the curve. We derive the same result applying the speed method described in section 2. Applying Lemma 3 and (2.5b) in Lemma 1 immediately yields

$$dJ(\Gamma;V) = \langle D_\Gamma J, V\rangle = \int_\Gamma \left\langle \Big(\frac{\partial g_I}{\partial n} + g_I\,(\kappa+\nu)\Big)\,\mathbf{n}, V\right\rangle dS.$$

Thus, the flow for a contour $\Gamma(t)$ which propagates in the direction of the negative gradient with respect to the functional (6.2) is given by

$$\Gamma_t = -\big(g_I\,(\kappa+\nu) + \langle \nabla g_I, \mathbf{n}\rangle\big)\,\mathbf{n}. \tag{6.3}$$

Note that in our case $\mathbf{n}$ denotes the *exterior* normal vector to the region enclosed by Γ, so we have different signs in the expression (6.3) as, e.g., in [8, 37]. The level set formulation corresponding to (6.3) (see [35]) is given by

$$u_t = g_I\left(\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right) + \nu|\nabla u|\right) + \langle \nabla g_I, \nabla u\rangle = \left(\operatorname{div}\left(g_I\,\frac{\nabla u}{|\nabla u|}\right) + \nu\right)|\nabla u|. \tag{6.4}$$

For $\nu = 0$, (6.3) or (6.4) can be interpreted as geodesic curve-shortening flow with respect to an image-dependent metric g_I (cf. [8, 30]).

We now use the results from section 5 for the setup of a Newton-type algorithm as described in section 4, specifically in (4.8), (4.6), (4.5), and (4.4) for variational image segmentation. Using (5.7) and (5.9), we find that (4.8) for the Newton-type

speed function F has the form of an elliptic equation on Γ: Find $F : \Gamma \to \mathbb{R}$ such that

$$\int_\Gamma \Big[\Big(\frac{\partial^2 g_I}{\partial n^2} + (2\kappa+\nu)\,\frac{\partial g_I}{\partial n} + \nu\,\kappa\,g_I\Big)\,F\,G + g_I\langle\nabla_\Gamma F, \nabla_\Gamma G\rangle\Big]\,dS = -\int_\Gamma \Big(\frac{\partial g_I}{\partial n} + (\kappa+\nu)\,g_I\Big)\,G\,dS \tag{6.5}$$

for all test functions $G : \Gamma \to \mathbb{R}$.

The elliptic problem (6.5) has a unique solution, and the solution to (6.5) is a descent direction with respect to (6.2) if the bilinear form on the left-hand side is coercive on $H^1(\Gamma)$. Since $g_I > 0$ on $\mathbb{R}^2$, this is the case if

$$\Big(\frac{\partial^2 g_I}{\partial n^2} + (2\kappa+\nu)\,\frac{\partial g_I}{\partial n} + \nu\,\kappa\,g_I\Big) > 0 \quad \text{on } \Gamma. \tag{6.6}$$

See [36, section 2.21] for a comprehensive treatment of elliptic problems on contours. We give some heuristic arguments why condition (6.6) is likely to be satisfied in a neighborhood of the optimal contour. Let us consider the case $\nu = 0$. If the optimal contour is located at the bottom of a valley for the edge detector g_I and if the contour is approximately aligned with the direction of the valley, we have $\frac{\partial g_I}{\partial n} \sim 0$ and g_I is convex in the direction normal to the contour, i.e., $\frac{\partial^2 g_I}{\partial n^2} > 0$. Thus, (6.6) is satisfied for such a contour.

The positive definiteness of (6.5) is an important computational issue. For the actual computations, the Hessian has to be modified such that positive definiteness is maintained also for contours outside the (possibly very small) neighborhood of the optimal contour, where the convexity of g_I in the normal direction is strong enough to guarantee coercivity. This topic (among others) is addressed in the next section.

7. Implementation of an active contour algorithm for image segmentation based on the Newton-type speed function. It is often said that the speed function

$$F = -\big(g_I\,\kappa + \langle\nabla g_I, \mathbf{n}\rangle\big)$$

corresponds to the negative gradient direction with respect to the cost functional (6.2) with $\nu = 0$ and, therefore, propagation with this speed function decreases the cost functional as fast as possible. On the other hand, it is observed that the decrease of J along the propagation of the level set function u is not very fast and that the time steps in the numerical implementation of the level set algorithm must be chosen relatively small. Otherwise, zig-zagging trajectories for the points on the contour are observed. In the worst case, the time-stepping procedure can even become unstable. In other words, the numerical realization of the front-propagation problem (6.4) suffers from many drawbacks which are well known for gradient-based algorithms in nonlinear programming. Usually, the constant (expanding or shrinking) term, i.e., $\nu > 0$ in (6.2), is added to the speed function F to speed up the propagation. This procedure has the disadvantage that an additional parameter (the constant deflation or inflation speed) is introduced into the algorithm. It is a difficult task to choose this parameter in a reasonable way. If it is too large, it is possible that weak edges in the image are not recognized and the propagation of the contour does not stop at the edge. If it is chosen too small, the desired speed-up cannot be achieved (see Table 2).

We propose a speed-up method for the propagating interface problem which—in ideal cases—is even parameter free, i.e., $\nu = 0$ is set in all iterations. The method

can be considered as a preconditioned gradient method or, alternatively, as a Newton-type technique. As speed function F in the level set equation (1.1), we choose the Newton-type direction with respect to the cost functional (6.2) as calculated in section 5. Additionally, we use a line search technique in order to relax the restriction on the time step size in the discretization of the level set equation given by the CFL-condition. We consider the following algorithm.

ALGORITHM 1.

(1) **Initialization.** *Choose an initial (closed) contour* Γ_0. *Initialize the level set function* u^0 *such that* Γ_0 *is the zero level set of* u^0*; set* $k = 0$. *Choose a bandwidth* $w \in \mathbb{N}$ *and* $\nu \in \mathbb{R}$.

(2) **Newton direction.** *Find the zero level set* Γ_k *of the actual level set function* u^k. *Solve* (6.5) *to obtain the Newton-type direction* F^k.

(3) **Extension.** *Extend* F^k *to a band around the actual zero level set* Γ_k *with bandwidth* w *yielding* F^k_{ext}.

(4) **Update.** *Perform a time step in the level set equation with speed function* F^k_{ext} *to update* u^k *on the band. Let* $\hat{u}^{k+1}$ *denote this update.*

(5) **Reinitialization.** *Reinitialize* $\hat{u}^{k+1}$ *in order to obtain a signed distance function* u^{k+1} *with zero level set given by the zero level set of* $\hat{u}^{k+1}$. *Set* $k = k+1$ *and go to* (2).

Before we discuss steps (1)–(5) of Algorithm 1 in detail, we note that the algorithm operates only on a band around the actual zero level set (or contour) Γ_k. This so-called narrow band approach was introduced by Chopp [14]. The key aspect is the fact that typically only knowledge around the actual contour is of importance in the propagation of the contour through the level set equation (which, then, is also considered only on the band). Clearly, in our situation the shape gradient and the shape Hessian used in (6.5) are both defined only on Γ_k. Thus, the narrow band approach is appropriate. In the discrete setting, the restriction to a band around the actual zero level set reduces the computational time and the memory requirement. In [1] a fixed band is chosen with respect to the contour and then, as soon as the propagated contour approaches the boundary of the band, the band is reinitialized with respect to the actual contour. In contrast to this technique, we allow a continuously moving band, i.e., the band is moved together with the contour. This enables us to take larger time steps while preserving a low computational cost. For more details on the narrow band approach in level set methods we refer to [14, 1, 35].

Now let us discuss the steps of Algorithm 1 and the respective numerical realization.

Initialization. In the literature there exist different characteristic choices with respect to ν in the cost functional (6.2). In the following discussion, we decide to consider deflation, i.e., we choose $\nu \geq 0$ and Γ_0 such that the objects which should be segmented are within the area enclosed by Γ_0. Depending on Γ_0, a signed distance function u^0 is computed with Γ_0 as zero level set. This is done by utilizing the fast marching technique [33, 34] on the band around Γ_0 for solving the Eikonal equation

$$|\nabla u^0| = 1 \quad \text{with} \quad u^0 = 0 \text{ on } \Gamma_0.$$

Unless it is chosen too small, the algorithm is not sensitive (except for computational time) with respect to the bandwidth w.

Newton direction. This step is the core part of the new algorithm. As already mentioned in the previous section, the coercivity in $H^1(\Gamma)$ of the bilinear form in (6.5) is essential for having a well-defined Newton-type descent direction. Typically, in the

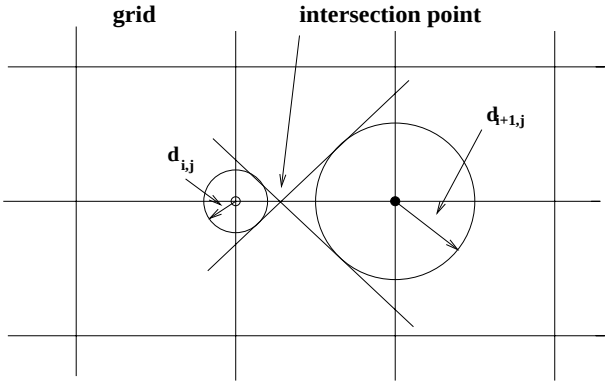


FIG. 2. *Computation of additional intersection points (on the discrete contour).*

course of the iteration it happens that

$$\frac{\partial^2 g_I}{\partial n} + (2\kappa + \nu)\frac{\partial g_I}{\partial n} + \nu\kappa g_I \leq 0 \quad \text{on some parts of } \Gamma_k.$$

Thus, the coercivity of the corresponding bilinear form is lost. To circumvent these difficulties, we incorporate the following modification of (6.5). Find $F : \Gamma \to \mathbb{R}$ such that

$$\int_\Gamma \Big[\Big(\frac{\partial^2 g_I}{\partial n^2} + (2\kappa+\nu)\,\frac{\partial g_I}{\partial n} + \nu\,\kappa\,g_I\Big)_+ F\,G + g_I\langle\nabla_\Gamma F, \nabla_\Gamma G\rangle\Big]\,dS = -\int_\Gamma \Big(\frac{\partial g_I}{\partial n} + (\kappa+\nu)\,g_I\Big)\,G\,dS \tag{7.1}$$

for all test functions $G : \Gamma \to \mathbb{R}$. Above we use $(\cdot)_+ = \max(\cdot, \epsilon)$ for $0 < \epsilon \ll 1$. In our numerical tests it turns out that frequently $\epsilon = 0$ can be set, i.e., we basically cut off the nonconvex part of the shape Hessian. For $\epsilon > 0$, equation (7.1) realizes a small correction of the Newton direction towards the steepest descent direction, i.e., the negative shape gradient.

The discretization of (7.1) is a rather delicate issue. This is due to the fact that $\langle\nabla_\Gamma F, \nabla_\Gamma G\rangle$ corresponds to the Laplace–Beltrami operator on $\Gamma = \Gamma_k$. We first need a discrete model Γ_k^h of Γ_k. For this purpose, we recall that one of the advantages of the level set method is the fact that it operates on a fixed (Cartesian) grid. In our case, the nodes are given by the pixels of the image. The information on the contour Γ_k is included in u^k, i.e., it is the zero level set of u^k. In order to get Γ_k^h we compute additional points on the grid lines (which are the lines joining the pixels of the given image) representing the discrete contour. Let us assume that $x_{i,j}$, $i = 1, \ldots, M$ and $j = 1, \ldots, N$, denote the grid points (nodes) and u_h^k is the signed distance function defined on the nodes. Whenever it is observed that $u_h^k(x_{i,j})$ and $u_h^k(x_{i+1,j})$ change sign, then the interface obviously passes through the grid line connecting $x_{i,j}$ and $x_{i+1,j}$. An analogous observation is true for $x_{i,j}$ and $x_{i,j+1}$. Since $d_{i,j} = |u_h^k(x_{i,j})|$ and $d_{i+1,j} = |u_h^k(x_{i+1,j})|$ give the distances to the contour, we compute an additional intersection point z_l^k as outlined in the graphic in Figure 2. The black node and the white node indicated different signs of u_h^k at these points. The discrete contour is given by the piecewise linear approximation joining the intersection points. For

simplicity, we temporarily assume that $\partial\Gamma_k^h = 0$ for all k and that $N_{\Gamma_k^h}$ represents the number of intersection points. Thus, Γ_k^h is a polygon. Let n_h^k denote the normal to Γ_h^k, and let $[z_l^k, z_{l+1}^k]$ represent a linear piece of Γ_k^h. For the discretization of F^k we use the ansatz

$$F_h^k(z) = \sum_{l=1}^{N_{\Gamma_k^h}} F_l^k \phi_{h,l}(z)$$

with $\phi_{h,l}$ the linear functions on Γ_k^h, which are globally continuous and satisfy $\phi_{h,l}(z_j^k) = \delta_{lj}$ for $l, j = 1, \dots, N_{\Gamma_k^h}$. We define the discretized tangential gradient $\nabla_{\Gamma_k^h}$ by

$$\nabla_{\Gamma_k^h} F_h^k(z) = \sum_{l=1}^{N_{\Gamma_k^h}} F_l^k \nabla_{\Gamma_k^h} \phi_{h,l}(z)$$

with

$$\nabla_{\Gamma_k^h} \phi_{h,l}(z) = \begin{cases} h_l^{-1} & \text{if} \quad z \in [z_{l-1}^k, z_l^k], \\ h_{l+1}^{-1} & \text{if} \quad z \in [z_l^k, z_{l+1}^k], \\ 0 & \text{else.} \end{cases}$$

Here h_l is the length of $[z_{l-1}^k, z_l^k]$; this is analogous for h_{l+1}. Note that $\nabla_{\Gamma_k^h} F_h^k$ is constant on each linear piece of Γ_k^h.

Let a_h^k denote the piecewise constant approximation of

$$a(z) = \left(\left(\frac{\partial^2 g_I}{\partial n^2} \right) + (2\kappa + \nu) \left(\frac{\partial g_I}{\partial n} \right) + \nu\, \kappa\, g_I \right)_+ (z) \quad \text{for } z \in \Gamma_k$$

with

$$a_h^k(z_l^k) = \left(\left(\frac{\partial^2 g_I}{\partial n^2} \right)_h + (2\kappa_h + \nu) \left(\frac{\partial g_I}{\partial n} \right)_h + \nu\, \kappa_h\, g_I \right)_+ (z_l^k) \quad \text{for } z_l^k \in \Gamma_k^h.$$

The approximation of the normal derivatives and the mean curvature in z_l^h, $l = 1, \dots, N_{\Gamma_k^h}$, are discussed below. Let b_h^k denote the piecewise constant approximations of g_I, which are defined as

$$b_h^k(z) = \frac{1}{2}(g_{I,h}(z_l^k) + g_{I,h}(z_{l+1}^k)) \quad \text{for } z \in [z_l^k, z_{l+1}^k].$$

For the discretization of the first term under the integral in (7.1), we use a mass lumping technique which yields a positive definite diagonal matrix. The right-hand side is approximated by utilizing the trapezoidal rule on each linear piece of Γ_k^h. The discretization of (7.1) is then given by

$$\sum_{l=1}^{N_{\Gamma_k^h}} F_l \left(a_h^k(z_l^k) h_l + \int_{\Gamma_k^h} b_h^k \langle \nabla_{\Gamma_k^h} \phi_{h,l} \nabla_{\Gamma_k^h} \phi_{h,j} \rangle \right) = \sum_{l=1}^{N_{\Gamma_k^h}} c_h^k(z_l^k) \hat{h}_l \tag{7.2}$$

for $j = 1, \dots, N_{\Gamma_k^h}$. Above, $\hat{h}_l$ is given by $\hat{h}_l = \frac{1}{2}(h_l + h_{l+1})$. In the case where Γ_k^h contains nonclosed components, $\hat{h}_l$ has to be modified on terminal linear pieces of these components.

When assembling the system matrix in (7.2) one has to be careful in order to produce a tridiagonal band matrix. In order to obtain this structure we employ the following technique. We compute a list containing all intersection points. First, we check whether one of the intersection points in the list is located on the grid lines joining the boundary pixels of the image. If this is the case, then we start with an intersection point on the boundary. In any case, we take z_l with the minimal l and follow the corresponding piece of the contour, compute the respective entries in the stiffness matrix corresponding to the actual intersection point, and delete this point from the list. If we have finished the piece of the discrete contour and the list is not empty, we repeat this procedure by checking intersection points on the boundary. If it turns out that there are no intersection points located on a boundary grid line, then we take z_l with minimal l. The final stiffness matrix is tridiagonal allowing efficient solutions of the discretization of (7.2).

For details concerning asymptotic error estimates for the finite element discretization of the elliptic equation (7.1) as described above, we refer to [17, 18].

The discretization of κ on the (fixed) grid points is based on finite differences like those in [35]. To obtain an approximation of $\frac{\partial g_I}{\partial n}$, we evaluate g_I in the grid points, compute $\nabla_h g_I$ by central differences, and compute $n_h = \frac{\nabla_h u_h^k}{|\nabla_h u_h^k|}$ as an approximation to the normal derivative in all grid points. Then

$$\left(\frac{\partial g_I}{\partial n}\right)_h (x_{i,j}) = \nabla_h g_I(x_{i,j})^T n_h(x_{i,j}).$$

Values for κ_h and $(\frac{\partial g_I}{\partial n})_h$ at intersection points are obtained as weighted averages of the respective quantity at neighboring grid points.

Extension. Since $dJ(\Gamma_k; V_G)$, $d^2J(\Gamma_k; V_F; V_G)$, and, thus, the Newton-type direction F^k are defined only on Γ_k, but the level set equation is defined on Ω (or at least on a band around Γ_k), an extension of F^k to Ω (or the band) must be computed. There exist many possible ways to extend F^k. According to (4.4), F^k_{ext} in step 3 of Algorithm 1 must satisfy

$$\langle \nabla F^k_{ext}, \nabla u^k \rangle = 0, \quad F^k_{ext}|_{\Gamma_k} = F^k. \tag{7.3}$$

On the discrete level, we realize (7.3) by employing the technique of [2]. Again, the fast marching method is used on the narrow band only. For more details we refer to [2].

Update. The discretization of the level set equation follows the standard suggestions in, e.g., [35]; i.e., the time stepping is done by using an explicit Euler scheme combined with an ENO-scheme for the term involving the spatial derivatives. Usually, the CFL-condition gives a link between the step size of the time and the spatial discretization such that the difference scheme is stable [23]. In our situation, the CFL-condition yields

$$\|F^k_{ext,h}\|_\infty \Delta t^k \leq \Delta x,$$

where Δt^k denotes the time step size in iteration k and Δx is the mesh size of the spatial discretization. Obviously, this might lead to very small time step sizes. This is especially true at early stages of the iteration process where the shape gradient is still large. Close to the discrete solution the CFL-condition becomes less stringent since $F^k_{ext,h}$ becomes "smaller."

In contrast to the requirement induced by the CFL-condition, we determine Δt^k based on considerations coming from optimization concepts. First, we relax Δt^k_{CFL}, the time step size required by the CFL-condition, by choosing a threshold $T^k := \ell \Delta t^k_{CFL}$ with $\ell > 1$. Due to our modification of the shape Hessian (its discretization induces a positive definite matrix), we expect that F^k is a local descent direction; i.e., for sufficiently small time step sizes the cost functional J is reduced by propagating Γ_k through the level set equation. A so-called sufficient decrease condition, well known from nonlinear programming [28], is given by the Armijo-condition. In our context this condition becomes

$$J(\Gamma_{k+1}) - J(\Gamma_k) \leq \mu \Delta t^k \langle F^k, dJ(\Gamma_k; F^k)\rangle < 0$$

with a fixed parameter $\mu \in (0,1)$. Numerically we realize the Armijo-condition in the following way: Let $\Gamma^h_k(\Delta t)$ denote the zero level of $u^k_h(\Delta t)$, the result of a time step with Δt in the discretized level set equation with speed function given by $F^k_{ext,h}$. At every iteration level k, we utilize the following algorithm.

ALGORITHM 2.

(1) *Set* $a_0 = \Delta t^k_{CFL}$, $b_0 = T^k$, $r_0 = b_0 - a_0$, $0 < \xi \ll \frac{1}{2}$. *Choose* $\Delta t_0 \in (a_0 + \xi r_0, b_0 - \xi r_0)$ *and* $0 < \mu_1 < \mu_2 < 1$*; set* $l = 0$. *Choose the maximal number of cycles* $L \in \mathbb{N}$.

(2) *Perform a time step in the level set equation with time step size* Δt_l *and speed function* $F^k_{ext,h}$, *and compute* $u^k_h(\Delta t_l)$.

(3) *Compute the zero level set* $\Gamma^h_k(\Delta t_l)$ *of* $u^k_h(\Delta t_l)$. *If* $l = L$, *then* $\Delta t^k = \Delta t_l$, $\hat{u}^{k+1}_h := u^k_h(\Delta t^k)$ *and RETURN to Algorithm* 1. *If* $\Gamma^h_k(\Delta t_l)$ *satisfies*

$$J_h(\Gamma^h_k(\Delta t_l)) - J_h(\Gamma^h_k) \leq \mu_2 \Delta t_l \langle F^k_h, dJ(\Gamma_k; F^k)_h\rangle < 0, \tag{7.4}$$

then $a_{l+1} = \Delta t_l$, $b_{l+1} = b_l$, $r_{l+1} = b_{l+1} - a_{l+1}$, *and compute* $\Delta t_{l+1} \in (a_{l+1} + \xi r_{l+1}, b_{l+1} - \xi r_{l+1})$. *If* (7.4) *is satisfied with* μ_2 *replaced by* μ_1 *and*

$$J_h(\Gamma^h_k(\Delta t_l)) - J_h(\Gamma^h_k) > \mu_2 \Delta t_l \langle F^k_h, dJ(\Gamma_k; F^k)_h\rangle,$$

then $\Delta t^k = \Delta t_l$, $\hat{u}^{k+1}_h := u^k_h(\Delta t^k)$ *and RETURN to Algorithm* 1. *If*

$$J_h(\Gamma^h_k(\Delta t_l)) - J_h(\Gamma^h_k) > \mu_1 \Delta t_l \langle F^k_h, dJ(\Gamma_k; F^k)_h\rangle, \tag{7.5}$$

then $b_{l+1} = \Delta t_l$, $a_{l+1} = a_l$, $r_{l+1} = b_{l+1} - a_{l+1}$, *and compute* $\Delta t_{l+1} \in (a_{l+1} + \xi r_{l+1}, b_{l+1} - \xi r_{l+1})$. *Set* $l = l + 1$ *and go to step* (2).

This step size strategy takes place in step 4 of the discrete analogue of Algorithm 1. The output (of Algorithm 2) is $\hat{u}^{k+1}_h = u^k_h(\Delta t^k)$.

Relaxing the CFL-condition significantly and computing time steps by Algorithm 2 is substantiated by the fact that we are not as interested in tracking an interface as we are in finding—as quickly as possible—a contour which locally minimizes the cost functional.

Reinitialization. Since we allow larger time steps compared to the typical choices for level set–based front propagation, the reinitialization is of importance. Like in the initialization phase, we use a fast marching technique for solving the Eikonal equation

$$|\nabla u| = 1 \quad \text{with} \quad u = 0 \quad \text{on} \quad \Gamma^{k+1}$$

numerically. Here, Γ^{k+1} is the zero level set of $\hat{u}^{k+1}$.

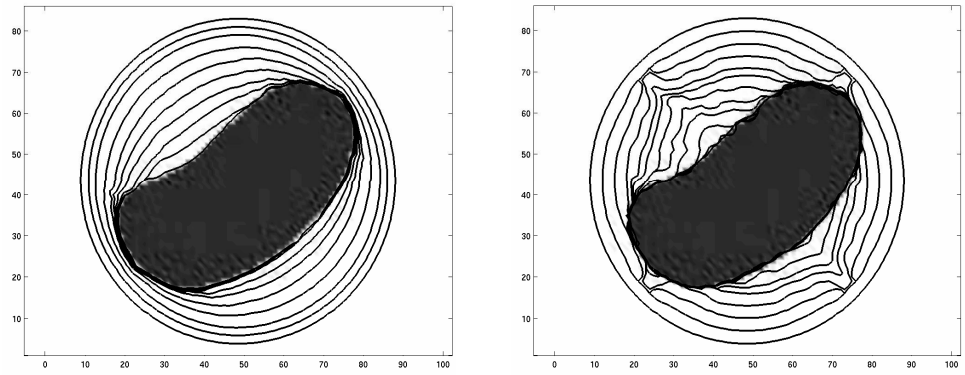


FIG. 3. *Image with zero level sets of u_h^k based on the parameter free ($\nu = 0$) Newton-type direction (left) and on the gradient-based direction with parameter $\nu = 1$ (right).*

TABLE 1
Time step sizes and cost functional values for the left graph of Figure 3.

k	Δt^k	Δt^k_{CFL}	J^k_h	$J^k_{h,r}$
1	0.00027	0.00014	67.71894	67.73983
2	0.00916	0.00458	63.62859	63.58714
3	0.05119	0.01462	55.69355	55.30486
4	0.07655	0.02187	45.59301	45.34222
5	0.11608	0.03317	37.06772	36.81020
6	0.16018	0.04577	28.19008	27.54977
7	0.20494	0.05856	16.41064	15.95286
8	0.31020	0.08862	9.73240	9.92598
9	0.34469	0.09848	4.01012	3.83231

8. Numerical results. In this section we report on numerical tests attained by Algorithm 1 for the discretization described in the previous section. With respect to the grayscales contained in the image data, the first two examples represent the ideal situation. We use these examples to demonstrate the advantages of the Newton-type direction compared to the gradient direction with deflation or inflation. Also comparisons with a method based on the negative gradient are given. The third example is related to the task of segmenting a contrast agent–based image of a kidney. Here we show that the new algorithm can handle inflation, i.e., $\nu < 0$, efficiently.

Let us start by reporting on the results for the image in Figure 3. Table 1 displays the time step sizes Δt^k accepted by Algorithm 2, the corresponding CFL-based time step Δt^k_{CFL}, the cost value J^k_h prior to the reinitialization, and $J^k_{h,r}$ after the reinitialization for Algorithm 1. Moreover, $\nu = 0$ is chosen. From Table 1 we can see that the Newton-type method stops after 9 iterations. The time step sizes Δt^k and Δt^k_{CFL} are increasing, which is expected since F^k_h should ideally vanish at a local solution. Also, our step size rule (Algorithm 2) yields significantly larger time steps than obtained from the CFL-condition. The cost functional is monotonically decreasing, and the reinitialization has only a slight influence on the cost functional value. If the speed function is changed from the Newton-type direction to the gradient direction with $\nu = 0$, then the algorithm (with step size strategy) needs 327 iterations (instead of 9 iterations for the Newton-type direction) to reduce the objective value from $J^1_h = 66.8179$ to $J^{327}_h = 3.6318$. If we allow a constant deflation by choosing

TABLE 2
Comparison of algorithms.

	Newton $\nu = 1$ with Alg. 2	Newton $\nu = 0$ with Alg. 2	Gradient $\nu = 1$ with Alg. 2	Gradient $\nu = 1$ no Alg. 2	Gradient $\nu = 0$ with Alg. 2
# it.	8	9	13	31	327

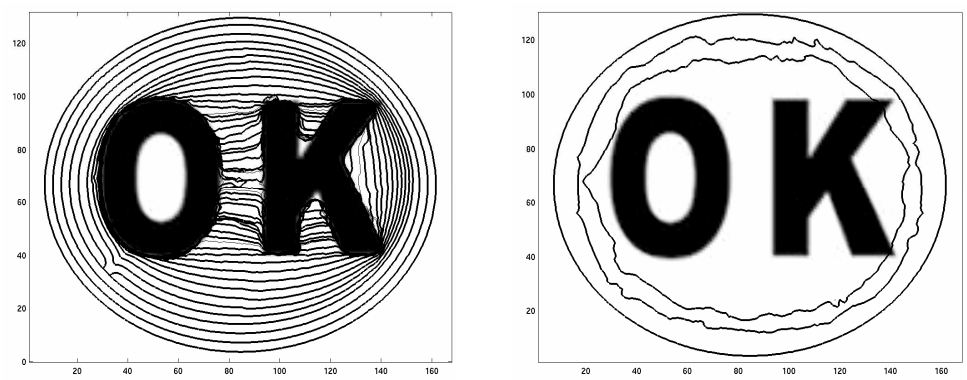


FIG. 4. *Image with zero level sets of u_h^k based on the Newton-type direction with $\nu = 0.01$ (left) and on the gradient-based direction with parameter $\nu = 1$ (right).*

$\nu = 1$ in the gradient-based method, then 13 iterations are needed. The fact that ν has to be chosen appropriately in order to avoid overshooting the desired contour or a slowly converging algorithm is a clear disadvantage. We also ran the algorithm with the gradient-based speed function with $\nu = 1$ and no step size strategy; i.e., $\Delta t^k = \Delta t^k_{CFL}$ was chosen. Then 31 iterations are needed for finding the local minimum numerically. In our test runs we also observe that the Newton-type direction acts more globally than the gradient direction. In fact, in the right graph of Figure 3 we can observe that the gradient direction detects certain parts of the contour rather quickly, while it takes some time to correctly detect the nonconvex part of the contour. The Newton-type direction yields a rather global propagation of the zero level set towards the desired contour; i.e., in the detection process (evolution of the zero level set of u_h^k) the zero level sets approach the desired contour more uniformly; see the left graph of Figure 3. Also, the contours based on the gradient-type propagation are less regular than the contours obtained from the Newton-type propagation. This behavior does not depend on our preference for employing Algorithm 2. It merely reflects our theoretical findings, i.e., computing F^k as the solution of the elliptic equation (7.1) induces additional smoothness properties of F^k. In Table 2 we summarize the convergence behavior.

Our second example is concerned with the segmentation of the letters "O" and "K" as displayed in Figure 4. Besides the aspect that our initial contour has to split into two disjoint contours, it is interesting to investigate how the Newton direction copes with the contours of "K" which involve, e.g., rather acute angles and specific nonconvexities. We shall also see that the appropriate choice of ν is a delicate issue for the gradient method. This is due to the fact that we have to balance the two objectives of fast progress and accurate segmentation. For the Newton method, on the other hand, a rather small value for ν already gives good progress without degrading the

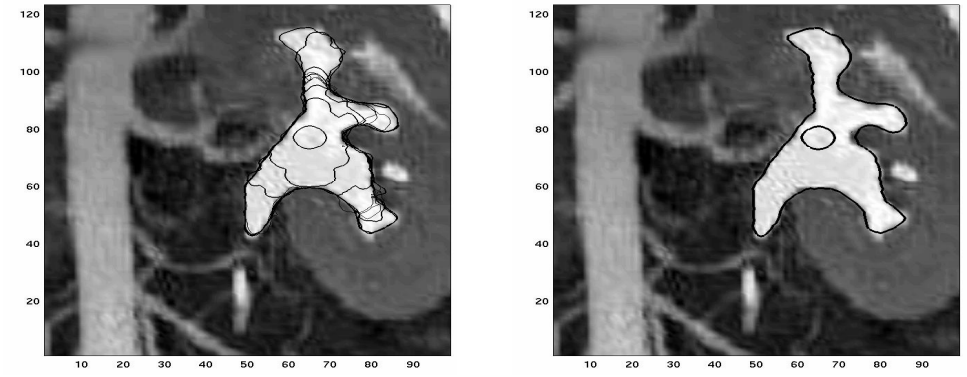


FIG. 5. *Image with zero level sets of u_h^k based on the Newton-type direction with $\nu = -0.3$ (left) and the corresponding initial and final contours (right).*

segmentation behavior.

The gradient-based direction with constant deflation ($\nu = 1$) with or without Algorithm 2 typically overshoots the right uppermost and lowermost corners. As a consequence, the segmentation misses the "K." For smaller values for ν the convergence speed of the gradient-based algorithm is significantly reduced, and too small ν eventually prevents the algorithm from convergence. From the right graph in Figure 4 we can observe that, like in the previous example, the contours for the gradient-based speed function are quite irregular. This prevents the algorithm from taking larger time step sizes.

For the Newton-type speed function, we choose $\nu = 0.01$ and still get reasonable progress in every iteration (successful termination after 33 iterations) but avoid the overshooting of the corners of "K." The evolution of the contours is displayed in the left graph of Figure 4. We initialize the algorithm with the outermost ellipse which shrinks towards the convex hull of the two letters and finally collapses onto two separate contours. With the same value for ν, the gradient-based algorithm with a step size strategy needs more then 100 iterations.

The final example is concerned with the task of segmenting a contrast agent–based image of a kidney. We initialize the algorithm by choosing a small circle inside the part of the image which we aim to segment. Thus, ν has to be assigned a negative value in order to allow inflation of the initial contour. In the left graph of Figure 5 we display some of the iterates of the Newton-type method with $\nu = -0.3$. The algorithm detects the correct contour after 47 iterations. The right graph shows the initial and the final contours for the Newton-type method.

REFERENCES

[1] D. ADALSTEINSSON AND J. A. SETHIAN, *A fast level set method for propagating interfaces*, J. Comput. Phys., 118 (1995), pp. 269–277.

[2] D. ADALSTEINSSON AND J. A. SETHIAN, *The fast construction of extension velocities in level set methods*, J. Comput. Phys., 148 (1999), pp. 2–22.

[3] L. ALVAREZ, P.-L. LIONS, AND J.-M. MOREL, *Image selective smoothing and edge detection by nonlinear diffusion*. II, SIAM J. Numer. Anal., 29 (1992), pp. 845–866.

[4] G. AUBERT AND L. BLANC-FÉRAUD, *Some remarks on the equivalence between 2d and 3d classical snakes and geodesic active contours*, Int. J. Comput. Vision, 34 (1999), pp. 19–28.

[5] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing*, Springer-Verlag, New York, 2002.
[6] V. Caselles, F. Catté, T. Coll, and F. Dibos, *A geometric model for active contours in image processing*, Numer. Math., 66 (1993), pp. 1–31.
[7] V. Caselles and B. Coll, *Snakes in movement*, SIAM J. Numer. Anal., 33 (1996), pp. 2445–2456.
[8] V. Caselles, R. Kimmel, and G. Sapiro, *Geodesic active contours*, Int. J. Comput. Vision, 22 (1997), pp. 61–79.
[9] V. Caselles, R. Kimmel, G. Sapiro, and C. Sbert, *Minimal surfaces based object segmentation*, IEEE Trans. Pattern Anal. Machine Intell., 19 (1997), pp. 394–398.
[10] T. F. Chan, B. Y. Sandberg, and L. A. Vese, *Active contours without edges for vector-valued images*, J. Visual Commun. Image Representation, 11 (2000), pp. 130–141.
[11] T. F. Chan and L. A. Vese, *Image Segmentation Using Level Sets and the Piecewise Constant Mumford-Shah Model*, UCLA CAM report 00-14, University of California, Los Angeles, 2000.
[12] T. F. Chan and L. A. Vese, *A level set algorithm for minimizing the Mumford-Shah functional in image processing*, UCLA CAM Report 00-13, University of California, Los Angeles, 2000.
[13] T. F. Chan and L. A. Vese, *Active contours without edges*, IEEE Trans. Image Process., 10 (2001), pp. 266–277.
[14] D. L. Chopp, *Computing minimal surfaces via level set curvature flow*, J. Comput. Phys., 106 (1993), pp. 77–91.
[15] M. C. Delfour and J.-P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, Adv. Des. Control 4, SIAM, Philadelphia, 2001.
[16] M. C. Delfour and J.-P. Zolésio, *Tangential calculus and shape derivatives*, in Shape Optimization and Optimal Design (Cambridge, 1999), Dekker, New York, 2001, pp. 37–60.
[17] G. Dziuk, *Finite elements for the Beltrami operator on arbitrary surfaces*, in Partial Differential Equations and Calculus of Variations, Lecture Notes in Math. 1357, S. Hildebrandt and R. Leis, eds., Springer-Verlag, Berlin, 1988, pp. 142–155.
[18] G. Dziuk, *An algorithm for evolutionary surfaces*, Numer. Math., 58 (1991), pp. 603–611.
[19] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature*, I, J. Differential Geom., 33 (1991), pp. 635–681.
[20] S. Jehan-Besson, M. Barlaud, and G. Aubert, *DREAM^{2}S: Deformable regions driven by an Eulerian accurate minimization method for image and video segmentation*, Int. J. Comput. Vision, 53 (2003), pp. 45–70.
[21] S. Jehan-Besson, M. Barlaud, and G. Aubert, *Video object segmentation using Eulerian region-based active contours*, in Proceedings of the Eighth IEEE International Conference on Computer Vision, IEEE Press, Piscataway, NJ, pp. 353–361.
[22] M. Kass, A. Witkin, and D. Terzopoulos, *Snakes; active contour models*, Int. J. Comput. Vision, 1 (1987), pp. 321–331.
[23] R. J. LeVeque, *Numerical Methods for Conservation Laws*, 2nd ed., Birkhäuser-Verlag, Basel, 1992.
[24] A. Litman, D. Lesselier, and F. Santosa, *Reconstruction of a two-dimensional binary obstacle by controlled evolution of a level-set*, Inverse Problems, 14 (1998), pp. 685–706.
[25] R. Malladi, R. Kimmel, D. Adalsteinsson, G. Sapiro, V. Caselles, and J.A. Sethian, *A geometric approach to segmentation and analysis of* 3*d medical images*, in Proceedings of the Mathematical Methods in Biomedical Image Analysis Workshop, San Francisco, CA, IEEE Press, Piscataway, NJ, 1996.
[26] R. Malladi, J. Sethian, and B. C. Vemuri, *Evolutionary fronts for topology independent shape modeling and recovery*, in ECCV—'94, Vol. 1, Lecture Notes in Comput. Sci. 800, Springer-Verlag, Berlin, 1994, pp. 3–13.
[27] R. Malladi, J. Sethian, and B. C. Vemuri, *Shape modeling with front propagation: A level set approach*, IEEE Trans. Pattern Anal. Machine Intell., 13 (1995), pp. 158–175.
[28] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, New York, 1999.
[29] P. J. Olver, G. Sapiro, and A. Tannenbaum, *Invariant geometric evolutions of surfaces and volumetric smoothing*, SIAM J. Appl. Math., 57 (1997), pp. 176–194.
[30] P. J. Olver, G. Sapiro, and A. Tannenbaum, *Affine invariant detection: Edge maps, anisotropic diffusion, and active contours*, Acta Appl. Math., 59 (1999), pp. 45–77.
[31] N. Paragios and R. Deriche, *Geodesic active regions: A new paradigm to deal with frame partition problems in computer vision*, Int. J. Visual Commun. Image Representation, 13 (2002), pp. 249–268.
[32] G. Sapiro, *Geometric Partial Differential Equations and Image Analysis*, Cambridge University Press, Cambridge, UK, 2001.

[33] J. A. Sethian, *A fast marching level set method for monotonically advancing fronts*, Proc. Nat. Acad. Sci. U.S.A., 93 (1996), pp. 1591–1595.
[34] J. A. Sethian, *Fast marching methods*, SIAM Rev., 41 (1999), pp. 199–235.
[35] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge University Press, Cambridge, UK, 1999.
[36] J. Sokołowski and J.-P. Zolésio, *Introduction to Shape Optimization*, Springer-Verlag, Berlin, 1992.
[37] A. Yezzi, S. Kichenassamy, A. Kumar, P. Olver, and A. Tannenbaum, *A geometric snake model for segmentation of medical imagery*, IEEE Trans. Med. Imaging, 16 (1997), pp. 199–209.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 468–483

MULTIVALUED FUNDAMENTAL DIAGRAMS AND STOP AND GO WAVES FOR CONTINUUM TRAFFIC FLOW EQUATIONS*

MARCO GÜNTHER†, AXEL KLAR‡, THORSTEN MATERNE‡, AND RAIMUND WEGENER†

Abstract. In the present paper a kinetic model for vehicular traffic leading to multivalued fundamental diagrams is developed and investigated in detail. For this model phase transitions can appear depending on the local density and velocity of the flow. A derivation of associated macroscopic traffic flow equations from the kinetic equation is given. Moreover, numerical experiments show the appearance of stop and go waves for highway traffic with a bottleneck.

Key words. traffic flow, macroscopic equations, kinetic derivation, multivalued fundamental diagram, stop and go waves, phase transitions

AMS subject classifications. 76P05, 90B20, 60K15

DOI. 10.1137/S0036139902404700

1. Introduction.

Classical models for vehicular traffic consider the continuity equation for the density ρ closing the equation by an equilibrium assumption on the mean velocity u, which means approximating u by a uniquely determined equilibrium value $U^e(\rho)$ (see [21]). The function $Q^e(\rho) = \rho U^e(\rho)$ is the so-called fundamental diagram. An additional momentum equation for u has been introduced by Payne [19] and Whitham [21] in analogy to fluid dynamics. Daganzo [3] has pointed out inconsistencies, like wrong-way traffic, of models such as the Payne–Whitham model in certain situations. The inconsistencies are resolved by the introduction of a new macroscopic model by Aw and Rascle [2]. For a mathematical discussion, see [1] or Greenberg [4]. Another basic problem of macroscopic traffic-flow equations has been described by Kerner [10, 11, 12]. The observations there suggest a more complicated dependence of the homogeneous steady speed states on density: the states are not given by a uniquely defined function $u = U^e(\rho)$, as used in the above models, but cover a whole range in the density-flow diagram.

Kinetic equations for vehicular traffic can be found, for example, in [20, 18, 16, 13]. Procedures for deriving macroscopic traffic equations, including the Aw–Rascle model, from underlying kinetic models have been performed in different ways by several authors; see, for example, [6] and [15]. These procedures are developed by analogy to the transition from the kinetic theory of gases to continuum gas dynamics. In the present paper a kinetic model is developed allowing for multiple stationary solutions. This leads to multivalued fundamental diagrams. Different steady speed states can appear for fixed density. They may be interpreted as traffic jams or free flow/synchronized traffic; compare [12]. An overview on these issues is given in [7]. Moreover, we refer to Nelson and Sopasakis [17] for investigations on this point for the Prigogine model, and to [5] for investigations on a simplified macroscopic model. Finally, the macroscopic

*Received by the editors March 28, 2002; accepted for publication (in revised form) June 14, 2003; published electronically December 31, 2003. This research was supported by Deutsche Forschungsgemeinschaft (DFG), KL 1105/5, and the EC network "HYKE."
http://www.siam.org/journals/siap/64-2/40470.html
†Fraunhofer-Institut für Techno- und Wirtschaftsmathematik (ITWM), 67663 Kaiserslautern, Germany (guenther@itwm.fhg.de, wegener@itwm.fhg.de).
‡Fachbereich Mathematik, Technische Universität Darmstadt, 64289 Darmstadt, Germany (klar@mathematik.tu-darmstadt.de, materne@mathematik.tu-darmstadt.de).

equations derived from the kinetic model exhibit the desired features such as stop and go behavior.

The paper is arranged in the following way. In section 2 the kinetic model is presented, reduced to a cumulative description of the highway. In section 3 the stationary distributions of the kinetic model are investigated, and the multivalued fundamental diagrams are determined. Section 4 contains the derivation of macroscopic models. Each section concludes with the discussion of an example. Finally, in section 5, numerical results are given showing the density-velocity relation and the different homogeneous stationary states which appear in the model. Moreover, a nonhomogeneous traffic flow situation with a bottleneck is investigated, showing the appearance of stop and go waves.

2. The kinetic model. The kinetic model developed here is based on the work in [14] and describes highway traffic in a cumulative way, averaging over all lanes.

2.1. Preliminaries. The basic quantity in a kinetic approach is the single-car distribution $f(x, v)$ describing the density of cars at x with velocity v. Here and in the following we do not write the time dependence explicitly. The total density ρ on the highway is defined by

$$\rho(x) = \int_0^w f(x, v)dv,$$

where w denotes the maximal velocity. Let $F(x, v)$ denote the probability distribution in v of cars at x, i.e., $f(x, v) = \rho(x)F(x, v)$. The mean velocity is

$$u(x) = \int_0^w vF(x, v)dv.$$

An important role is played by the distribution $f^{(2)}(x, v, h, v_+)$ of pairs of cars at the spatial point x with velocity v and leading cars at $x + h$ with velocity v_+. This distribution function has to be approximated by the one-vehicle distribution function $f(x, v)$. We use the chaos assumption

$$f^{(2)}(x, v, h, v_+) = q(h, v; \rho, u)\, f(x, v)\, F(x + h, v_+);$$

compare Nelson [16]. For a vehicle with velocity v the function $q(h, v; \rho, u)$ denotes the distribution of leading vehicles with distance h under the assumption that the velocities of the vehicles are distributed according to the distribution function f.

Moreover, we introduce the following thresholds for braking (H_B) and acceleration (H_A):

$$H_X(v) = H_0 + vT_X, \quad X = B, A.$$

$T_B = T_B(\rho, u)$ and $T_A = T_A(\rho, u)$ with $T_B < T_A$ are reaction times which may depend on ρ and u. H_0 denotes the minimal distance between the vehicles. We write $H_X = H_X(v) = H_X(v; \rho, u)$. From a microscopic point of view, drivers will brake once the distance between the driver and its leading car becomes smaller than a threshold H_B, and will accelerate once this distance becomes larger than H_A. Otherwise the cars will not change velocities. Velocities are changed instantaneously once acceleration or braking lines are reached. Models including acceleration of the cars can be developed as well; see [8] for an example.

The basic additional feature here compared to previous models (see [15]) is the fact that the reaction times may depend on ρ and u. This will be sufficient to obtain multivalued fundamental diagrams. Using different values for the reaction times, we assume that, according to their local situation, drivers react with different behavior. Having different T_B, they allow for different distances to their leading cars. Free flow of cars will be related to smaller distances and thus smaller T_B between cars driving at a given speed. So-called synchronized traffic is associated with slightly larger T_B and thus slightly larger distances between cars for a given speed. Traffic which is not synchronized is associated with even larger T_B and thus with lower speeds at the above fixed density. A special choice of $T_B = T_B(u)$ is given and discussed in detail in section 5.

The distribution of leading vehicles $q(h, v; \rho, u)$ is prescribed a priori. The main properties that $q(h, v; \rho, u)$ has to fulfill are positivity,

$$\int_0^\infty q(h, v; \rho, u)dh = 1,$$

and

$$\int_0^w \int_0^\infty hq(h, v; \rho, u)dh\, F(v)dv = \frac{1}{\rho}. \tag{1}$$

Equation (1) means that the average headway of the cars is $1/\rho$. The leading vehicles are assumed to be distributed in an uncorrelated way with a minimal distance H_B from the car under consideration (see Nelson [16]):

$$q(h, v; \rho, u) = \tilde{\rho}\, e^{-\tilde{\rho}(h - H_B(v))}\, \chi_{[H_B(v),\infty)}(h).$$

The reduced density $\tilde{\rho}$ has to be defined in such a way that (1) is fulfilled. One obtains

$$\tilde{\rho} = \frac{\rho}{1 - \rho \int_0^w H_B(v)F(v)dv} = \frac{\rho}{1 - \rho\, H_B(u; \rho, u)}. \tag{2}$$

REMARK 2.1. *The reduced density $\tilde{\rho}$ must be positive, i.e.,*

$$1 - \rho\, H_B(u; \rho, u) = 1 - \rho\,(H_0 + uT_B(\rho, u)) > 0.$$

This defines a range of admissible values in the (ρ, u)-plane.

The probability P_{ov} for overtaking or lane changing and the corresponding probability $P_B = 1 - P_{ov}$ for braking are determined from microscopic considerations. A car overtakes if there is sufficent space in the new lane, i.e., if the cars in the new lane have at least a distance $H_B(v)$ from the changing car. Averaging over the distribution function yields

$$P_{ov}(v; \rho, u) = \rho \int_0^w \int_0^w \int_{H_B(v)+H_B(v')}^\infty \int_h^\infty q(h', \tilde{v}; \rho, u)dh'dhF(\tilde{v})d\tilde{v}F(v')dv'. \tag{3}$$

REMARK 2.2. *The main mechanism for obtaining features like multivalued fundamental diagrams and stop and go behavior is the choice of T_B, i.e., the assumption that there are different states of drivers. We note that the exact choice of T_B is not important. Such an assumption is directly related to the assumption that drivers have different behavior concerning the accepted headway to their leading cars H_B. This, in*

turn, is related to the probability of overtaking. We note that the probability of overtaking is most important in the theory of Kerner [12] *for explaining multivalued fundamental diagrams. Thus, the present approach can be compared—and justified—by comparing the results for the probability of overtaking obtained here to the qualitative results of Kerner* [12]. *For such a comparison, see the figures in* [12] *and Figure* 7 *below.*

REMARK 2.3. *In the following we present a kinetic model. Note that the results like multivalued fundamental diagrams and stop and go behavior of the derived macroscopic equations do not depend on the exact choice of the microscopic interactions that we have chosen here. The model developed in the next section is only chosen due to the fact that explicit stationary solutions are available.*

2.2. The evolution equation. The kinetic model is given by the following evolution equation for the distribution function f:

$$\begin{aligned}\partial_t f + v\partial_x f &= \hat{C}^+(f)\\ &= (\hat{G}_B^+ - \hat{L}_B^+)(f) + (\hat{G}_A^+ - \hat{L}_A^+)(f) + (\hat{G}_S - \hat{L}_S)(f).\end{aligned} \tag{4}$$

$\hat{G}_B^+, \hat{L}_B^+$ denote the gain and loss terms due to braking, and $\hat{G}_A^+, \hat{L}_A^+$ those due to acceleration interactions. $\hat{G}_S$ and $\hat{L}_S$ are terms describing a random behavior of the drivers. They are explained in the following.

For the *braking interaction* one obtains the gain term

$$\hat{G}_B^+(f) = \int\int_{\hat{v}>\hat{v}_+} |\hat{v}-\hat{v}_+| q(H_B(\hat{v}),\hat{v};\rho,u) P_B(\hat{v};\rho,u) \sigma_B(v;\hat{v},\hat{v}_+) f(x,\hat{v}) F(x+H_B(\hat{v}),\hat{v}_+) d\hat{v} d\hat{v}_+$$

with the distribution σ_B of new velocities v after the interaction. The loss term is given by

$$\hat{L}_B^+(f) = \int_{\hat{v}_+<v} |v-\hat{v}_+| q(H_B(v),v;\rho,u) P_B(v;\rho,u) f(x,v) F(x+H_B(v),\hat{v}_+) d\hat{v}_+.$$

In other words, the driver is braking if he is not changing to the left lane for overtaking. Reaching the braking line, the vehicle brakes such that the new velocity v is distributed with a distribution function σ_B, depending on the old velocities $\hat{v}, \hat{v}_+$.

For the *acceleration interaction* the gain term is given by

$$\hat{G}_A^+(f) = \int\int_{\hat{v}<\hat{v}_+} |\hat{v}-\hat{v}_+| q(H_A(\hat{v}),\hat{v};\rho,u) \sigma_A(v;\hat{v},\hat{v}_+) f(x,\hat{v}) F(x+H_A(\hat{v}),\hat{v}_+) d\hat{v} d\hat{v}_+.$$

The loss term is

$$\hat{L}_A^+(f) = \int_{\hat{v}_+>v} |v-\hat{v}_+| q(H_A(v),\hat{v}_+;\rho,u) f(x,v) F(x+H_A(v),\hat{v}_+) d\hat{v}_+.$$

Thus, the new velocity is again distributed according to σ_A depending on the old velocities. Finally, a relaxation term is introduced, describing a random behavior of the drivers. It is given by

$$\hat{G}_S(f) = \nu(\rho,u) \int_0^w \sigma_S(v,\hat{v}) f(x,\hat{v}) d\hat{v},$$

where ν denotes an interaction frequency and $\int_0^\infty \sigma_S(v,\hat{v})dv = 1$. The loss term is

$$\hat{L}_S(f) = \nu(\rho,u)\int_0^w \sigma_S(\hat{v},v)f(x,v)d\hat{v} = \nu(\rho,u)\, f(v).$$

REMARK 2.4. *For remarks on this Boltzmann–Enskog approach to traffic flow modelling, see* [13].

In the following, (4) is simplified by using appropriate averages. We consider the functions $q(H_X(v),v;\rho,u)$, $X = A,B$ appearing in the above integrals and replace the velocity v in these expressions with the mean velocity u. This means that we approximate

$$q(H_A(v),v;\rho,u) \;\sim\; q(H_A(u),u;\rho,u) := q_A(\rho,u) \;=\; \tilde{\rho}\, e^{-\tilde{\rho}(T_A-T_B)u}$$

and

$$q(H_B(v),v;\rho,u) \;\sim\; q(H_B(u),u;\rho,u) := q_B(\rho,u) = \tilde{\rho}.$$

The probability for braking is approximated using formula (3) and substituting $H_B(v)$ and $q(h,v)$ by $H_B(u)$ and $q(h,u)$, respectively. We obtain

$$P_{ov}(\rho,u) = (1-\rho H_B(u))e^{-\frac{\rho H_B(u)}{1-\rho H_B(u)}}. \tag{5}$$

With $P_B = 1 - P_{ov}$ this gives

$$P_B(\rho,u) = 1-(1-\rho H_B(u))e^{-\frac{\rho H_B(u)}{1-\rho H_B(u)}}. \tag{6}$$

To rewrite the equations in a simpler form, we use

$$k = k(\rho,u) = \frac{P_B\, q_B}{q_A + P_B q_B}$$

and

$$\gamma = \gamma(\rho,u) = \frac{q_A}{1-k} = q_A + P_B q_B.$$

Finally, we define c by

$$\gamma\, c \;=\; \nu$$

and assume for simplicity that c depends on ρ, u through k, which means

$$c \;=\; c(k).$$

Using these approximations, (4) is rewritten as

$$\begin{aligned}\partial_t f + v\partial_x f &= C^+(f)\\ &= \gamma\left[k(G_B^+ - L_B^+)(f) + (1-k)(G_A^+ - L_A^+)(f) + c(G_S - L_S)(f)\right],\end{aligned} \tag{7}$$

with

$$G_B^+(f) = \int\int_{\hat{v}>\hat{v}_+} |\hat{v}-\hat{v}_+|\sigma_B(v;\hat{v},\hat{v}_+)f(x,\hat{v})F(x+H_B(\hat{v}),\hat{v}_+)d\hat{v}d\hat{v}_+,$$

$$L_B^+(f) = \int_{\hat{v}_+<v} |v-\hat{v}_+|f(x,v)F(x+H_B(v),\hat{v}_+)d\hat{v}_+,$$

$$G_A^+(f) = \int\int_{\hat{v}<\hat{v}_+} |\hat{v}-\hat{v}_+|\sigma_A(v;\hat{v},\hat{v}_+)f(x,\hat{v})F(x+H_A(\hat{v}),\hat{v}_+)d\hat{v}d\hat{v}_+,$$

$$L_A^+(f) = \int_{\hat{v}_+>v} |v-\hat{v}_+|f(x,v)F(x+H_A(v),\hat{v}_+)d\hat{v}_+,$$

$$G_S(f) = \int_0^w \sigma_S(v,\hat{v})f(x,\hat{v})d\hat{v},$$

$$L_S(f) = f(v).$$

2.3. An example. For the probability distributions σ_A, σ_B we choose the following simple expressions:

$$\sigma_B(v,\hat{v},\hat{v}_+) = \frac{1}{\hat{v}-\hat{v}_+}\,\chi_{[\hat{v}_+,\hat{v}]}(v) \tag{8}$$

and

$$\sigma_A(v,\hat{v},\hat{v}_+) = \frac{1}{\hat{v}_+-\hat{v}}\,\chi_{[\hat{v},\hat{v}_+]}(v). \tag{9}$$

This means that we have an equidistribution of the new velocities between the velocity of the car and the velocity of its leading car. Finally,

$$\sigma_S(v,\hat{v}) = \frac{1}{w}. \tag{10}$$

A special choice of c is given in the last section.

3. Stationary distributions and multivalued fundamental diagrams. In this section we investigate the stationary homogeneous equations and determine the multivalued fundamental diagrams.

3.1. The general case. We consider the local interaction operator

$$C(f) = \gamma\,[k(G_B - L_B)(f) + (1-k)(G_A - L_A)(f) + c(G_S - L_S)(f)]$$

with $f = \rho F$. The gain and loss terms G_B, L_B, etc., are defined in the same way as G_B^+, L_B^+, etc.; however, $x + H_X(v)$, $X = A, B$ is substituted by x wherever it appears. The homogeneous stationary equation is

$$C(f) = 0.$$

We assume that for fixed ρ and k there is a unique solution

$$f = f^e = \rho F^e(k,v)$$

of this equation. This is true for the example stated above. The exact form of the distribution functions is given in detail below.

Thus, for fixed k the mean value of F^e is then

$$u^e(k) := \int_0^w v F^e(k,v) dv.$$

The function u^e is uniquely determined due to the above assumption as a function of k. However, this does not immediately yield the fundamental diagram, i.e., an equilibrium relation between flux and density.

Instead, the fundamental diagram is determined from the following considerations. Let u be the (possibly multivalued) solution of the equation

$$u \;=\; u^e(k(\rho,u))$$

for fixed ρ. We denote this (possibly multivalued) solution by $u = U^e(\rho)$. If there is a unique solution, we obtain a well-defined relation for equilibrium velocity and density and the usual fundamental diagram $Q^e(\rho) = \rho U^e(\rho)$. However, in general this equation will have a multitude of different solutions u, even infinitely many. Plotting a dependence of this solution on the density, one obtains in the general case a two-dimensional region in the density-velocity plane, where the solutions are located. The fundamental diagram is then a multivalued function $Q^e(\rho) = \rho U^e(\rho)$.

3.2. The example. First we determine the stationary solutions $F^e(k)$ for a fixed value of the parameter k. Substituting the explicit expressions (8)–(10) for σ_X, $X = A, B, S$, one obtains

$$\begin{aligned}
G_B(f) &= G_A(f) = \rho \int_0^v F(\hat v)\, d\hat v \int_v^w F(\hat v)\, d\hat v,\\
L_B(f) &= \rho F(v) \left[v \int_0^v F(\hat v)\, d\hat v - \int_0^v \hat v F(\hat v)\, d\hat v \right],\\
L_A(f) &= \rho F(v) \left[\int_v^w \hat v F(\hat v)\, d\hat v - v \int_v^w F(\hat v)\, d\hat v \right],\\
G_S(f) &= \frac{\rho}{w},\\
L_S(f) &= f.
\end{aligned}$$

Defining $\mathcal{F} : [0,w] \to [0,1]$ by $\mathcal{F}(v) = \int_0^v F(\hat v)\, d\hat v$ and denoting the inverse function with $v(p)$, a straightforward computation shows that the integral equation $C(f) = 0$ is equivalent to the following boundary value problem for $v(p)$:

$$v'' = v' \, \frac{3p + k - 2}{p(1-p) + \frac{c}{w}}, \qquad v(0) = 0, \quad v(1) = w. \tag{11}$$

Using

$$h(p) = \frac{k-p}{\left(q - \left(p - \frac12\right)\right)^{\frac12 + r} \left(q + \left(p - \frac12\right)\right)^{\frac12 - r}}$$

with

$$q = \sqrt{\frac{c}{w} + \frac14}\,, \qquad r = \frac{2k-1}{4q},$$

the solution of (11) is explicitly given as

$$v(p) = w\frac{h(p) - h(0)}{h(1) - h(0)}.$$

A parameter representation of the unique stationary solutions $F = F^e(k, v)$ for fixed k is given as

$$F^e(k, v(p)) = \frac{1}{v'(p)}.$$

The mean velocity $u^e(k)$ of $F^e(k)$ is given by

$$u^e(k) = \int_0^w vF^e(k, v)\, dv = \int_0^1 v(\tilde{p})\, d\tilde{p} = w\frac{H(1) - H(0) - h(0)}{h(1) - h(0)}$$

with

$$H(p) = \int_0^p h(\tilde{p})d\tilde{p} = \left(q - \left(p - \frac{1}{2}\right)\right)^{\frac{1}{2}-r} \left(q + \left(p - \frac{1}{2}\right)\right)^{\frac{1}{2}+r}.$$

The multivalued fundamental diagrams are then obtained as the solutions of the equation $u = u^e(k(\rho, u))$ for fixed ρ. A numerical investigation of this nonlinear equation is given in the last section. (Plots of $u^e(k)$ and of the fundamental diagram are shown in Figures 4 and 5, 6, respectively.)

4. Derivation of macroscopic models. In this section macroscopic equations for density and mean velocity are derived following the procedure in [15].

4.1. Balance equations. Multiplying the inhomogeneous kinetic equation (7) with 1 and v and integrating it with respect to v, one obtains the following set of balance equations:

$$\begin{aligned} \partial_t\rho + \partial_x(\rho u) &= 0, \\ \partial_t(\rho u) + \partial_x(P + \rho u^2) + E &= S, \end{aligned} \tag{12}$$

with the "traffic pressure"

$$P = \int_0^w (v - u)^2 f dv, \tag{13}$$

the Enskog flux term

$$E = \int_0^w v[C(f)(x, v, t) - C^+(f)(x, v, t)]dv, \tag{14}$$

and the source term

$$S = \int_0^w vC(f)(x, v, t)dv. \tag{15}$$

To obtain closed equations for ρ and u one has to specify the dependence of P, E, and S on ρ and u.

4.2. Closure relations. We use the ansatz $f^{ex} = f^{ex}(\rho, u) = f^{ex}(\rho, u; v)$ for the distribution function to approximate the true distribution f and to close the equations. Let the function F^{ex} be defined by

$$f^{ex} = \rho F^{ex}.$$

We require that $F^{ex}(u)$ fulfill two properties, namely, having density

$$1 = \int_0^w F^{ex}(u, v)dv \tag{16}$$

and mean value

$$u = \int_0^w vF^{ex}(u, v)dv. \tag{17}$$

Note that the equilibrium distribution $F^e(k(\rho, u), v)$ has a mean value $u^e(k(\rho, u))$ and does not necessarily fulfill the above requirement. We construct F^{ex} using the one parameter family $F^e(k, v)$ and choosing a suitable value for k. The following simple ansatz for F^{ex} fulfills conditions (16), (17):

$$F^{ex}(u, v) = F^e(k^e(u), v),$$

where $F^e = F^e(k, v)$ is the uniquely defined function from section 3 and $k^e = k^e(u)$ is the inverse function to u^e, i.e.,

$$u^e(k^e(u)) = u.$$

Here k^e is well defined if we assume that $u^e(k)$ is strictly monotone decreasing in k. This is true for our example; see Figure 4. We note that for this definition of $f^{ex}(\rho, u)$ one obtains a positive function f^{ex} for all values of v.

Equation (12) is now closed by approximating the traffic pressure P in (13) by

$$P = \int_0^w (v - u)^2 f dv \ \sim \ \int_0^w (v - u)^2 f^{ex}(\rho, u; v)dv =: P^{ex}(\rho, u) = \rho\Theta^{ex}(u),$$

where the definition of the variance Θ^{ex} is given by the last equality sign. Further approximation of the variance gives

$$P \sim \rho\Theta^{ex}(u) \sim \rho\Theta^{ex}(u^e(k)) =: \rho\theta^e(k).$$

Moreover, the Enskog term E is approximated by linearizing expression (14) for E in H and substituting $f^{ex}(\rho, u)$ for f. One obtains

$$E \ \sim \ \rho A^{ex}(u)\partial_x u,$$

where $A^{ex}(u)$ is defined by

$$A^{ex} = -I(F^{ex}, \partial_u F^{ex}),$$

with

$$I(f, g) = I_B(f, g) + I_A(f, g),$$

$$I_B(f, g) = \gamma k \int \int_{\hat{v} > \hat{v}_+} |\hat{v} - \hat{v}_+| H_B(\hat{v}) f(\hat{v}) g(\hat{v}_+) \left[\int_0^w v\sigma_B(v, \hat{v}, \hat{v}_+)dv - \hat{v} \right] d\hat{v}_+ d\hat{v},$$

and

$$I_A(f,g) = \gamma(1-k)\int\int_{\hat{v}<\hat{v}_+} |\hat{v}-\hat{v}_+| H_A(\hat{v}) f(\hat{v}) g(\hat{v}_+) \left[\int_0^w v\sigma_A(v,\hat{v},\hat{v}_+)dv - \hat{v}\right] d\hat{v}_+ d\hat{v}.$$

To compute A^{ex} we use

$$\partial_u F^{ex}(u) = \frac{\partial_k F^e(k^e(u))}{\partial_k u^e(k^e(u))}.$$

Further approximation gives

$$A^{ex}(u) \sim A^{ex}(u^e(k)) =: a^e(k)$$

and

$$E \ \sim \ \rho a^e(k(\rho,u))\partial_x u.$$

Finally, the source term S has to be approximated:

$$S \ \sim \ S^{ex}(\rho,u) = \int_0^w vC(f^{ex})dv.$$

Further approximations are discussed in the next subsection.

One obtains macroscopic equations of the form

$$\begin{aligned} \partial_t \rho + \partial_x(\rho u) &= 0, \\ \partial_t(\rho u) + \partial_x(\rho\theta^e(k) + \rho u^2) + \rho a^e(k)\partial_x u &= S^{ex}(\rho,u), \end{aligned} \tag{18}$$

with $k = k(\rho,u)$ given above.

4.3. The example. In the case of our example the above expressions can be simplified. A short computation shows that for any distribution function f we have the following relation between $P = P(f)$ and $S = S(f)$:

$$S = \gamma\left(\left(\frac{1}{2}-k\right)P + c\rho\left(\frac{w}{2}-u\right)\right). \tag{19}$$

Computation of P^{ex} and θ^e. Using the distribution function $f = F^e(k)$, the above expression is equal to 0 for arbitrary k, since $S(F^e(k))$ is equal to zero by definition of F^e. Choosing $k = k^e = k^e(u)$, we get

$$P^{ex}(\rho,u) \ = \ \rho c(k^e(u))\frac{\frac{w}{2}-u}{k^e(u)-\frac{1}{2}}.$$

Further approximation as before gives

$$P^{ex}(\rho,u) \ \sim \ \rho\theta^e(k(\rho,u))$$

with

$$\theta^e(k) = c(k)\frac{\frac{w}{2}-u}{k-\frac{1}{2}}.$$

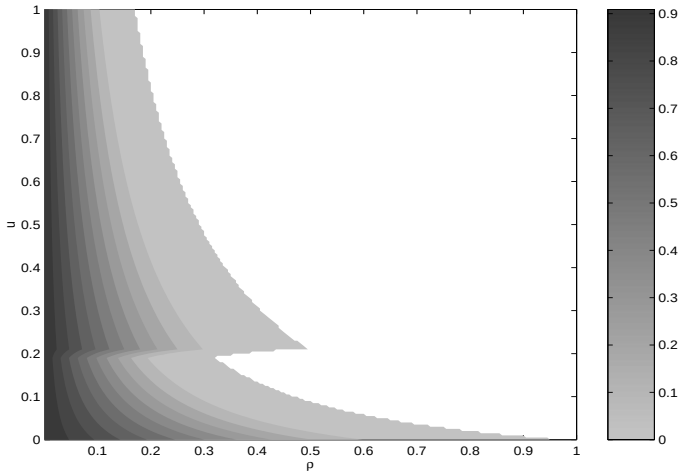


FIG. 1. *Probability* $P_{ov}(\rho, u)$.

Computation and approximation of S^{ex}. To compute S^{ex} we use (19) and substitute P^{ex} found above. We obtain

$$S^{ex}(\rho,u) \;=\; \gamma\rho\left(\frac{w}{2}-u\right)\left(c(k)-c(k^e(u))\frac{\frac{1}{2}-k}{\frac{1}{2}-k^e(u)}\right).$$

A simplification is given by substituting $\rho\theta^e(k)$ instead of P^{ex} into (19):

$$S^{ex}(\rho,u) \;\sim\; \rho\gamma c(k)\,(u^e(k)-u)\,;$$

i.e., the right-hand side has relaxation form.

Computation of A^{ex}. A^{ex} and a^e are computed as described above using the special form of F^e.

REMARK 4.1. *The macroscopic equations are then given by* (18) *with the explicit formulas derived above. As can be observed numerically (see* [15]*), the term containing* $\rho\theta^e(k)$ *in* (18) *is small compared to the other terms. Eventually, we obtain*

$$\partial_t\rho+\partial_x(\rho u)=0, \tag{20}$$
$$\partial_t(\rho u)+\partial_x(\rho u^2)+\rho a^e(k)\partial_x u=\rho\nu\,(u^e(k)-u)\,.$$

Thus, one obtains a multimodal variant of the Aw–Rascle equations with a multimodal relaxation term on the right-hand side. This is exactly what has been used and further investigated in [5].

5. Numerical investigations. In this section we investigate the example described in the text. The stationary homogeneous kinetic equation is discussed together with the resulting macroscopic equations.

5.1. Choice and further discussion of free parameters. For the numerical simulations, we normalize and choose $w=1$ and $H_0=1$. As described above, we use different reaction times to describe the drivers' behavior in different flow situations. Smaller reaction times are associated with higher velocities. To put things as simply as possible, we choose

$$T_B=\begin{cases}10, & u<0.2,\\ 5, & u\geq 0.2.\end{cases}$$

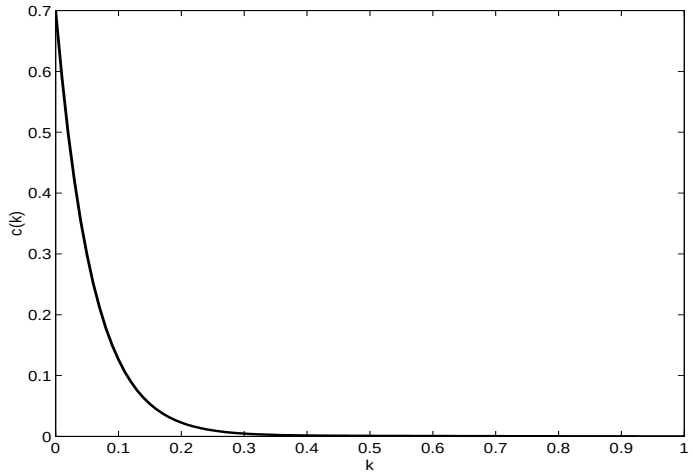


FIG. 2. *Frequency of random events* $c(k)$.

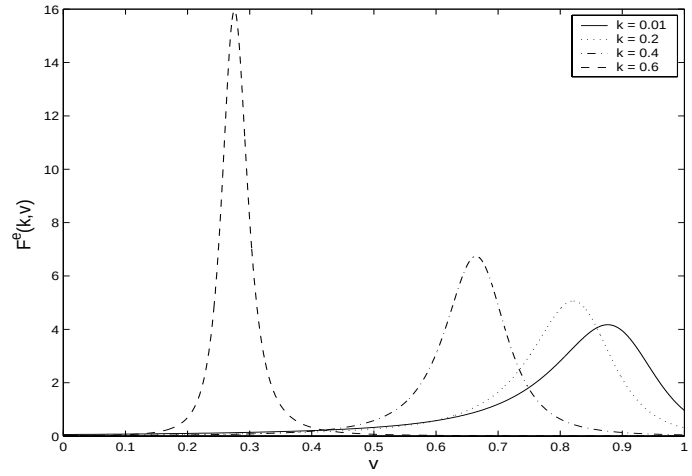


FIG. 3. *Stationary distribution functions* $F^e(k, v)$ *for different* k.

Thus, in terms of the accepted distance H_B to the leading car, drivers behave in the following way: once their velocity falls below a certain speed, relatively larger distances are required as before. Here, we do not distinguish between three traffic states like free flow, synchronized traffic, and jam, but rather between two states. For numerical reasons the step function has been smoothed.

As mentioned before, different choices of T_B would yield similar results as long as T_B was not constant. Choosing a constant reaction time T_B results in a uniquely determined and not a multivalued fundamental diagram.

The resulting probability of overtaking $P_{ov}(\rho, u)$, according to (5), is plotted in Figure 1 in the range of admissible values (ρ, u), according to (2).

Either T_B or P_{ov} should be viewed as the basic quantity of the multivalued approach presented here. In particular, as shown in Figure 7 below, considering P_{ov} for the equilibrium velocity gives good qualitative coincidence with the corresponding

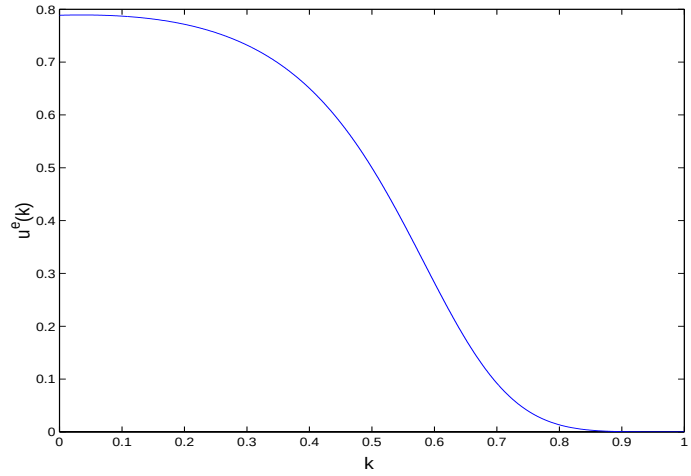


FIG. 4. *Function* $u^e(k)$.

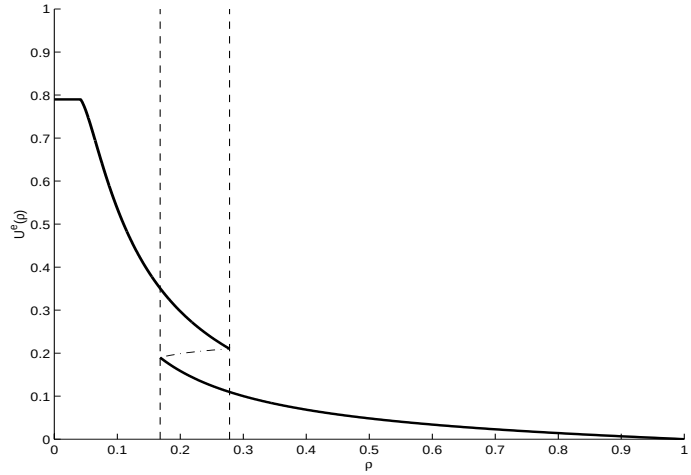


FIG. 5. *Density-velocity dependence* $U^e(\rho)$.

figures in [12], justifying the choice of T_B.

Moreover, we set $T_A = 2T_B$ and choose c as in Figure 2. A reasonable function c should be zero for maximal density ($k = 1$). In this case there is no more random behavior of the drivers; all drivers have velocity 0. For the case $k = 0$ we have chosen c as a finite quantity. If these two features are fulfilled, the qualitative behavior of the model, in particular the multimodal behavior, does not depend on the exact form of c. Moreover, changing the interaction rules for braking and acceleration, results similar to those below can be obtained with models as in [15] with $c = 0$.

5.2. The spatially homogeneous case (stationary, homogeneous kinetic equation). Using the values described above, we compute the stationary solution of the homogeneous kinetic equation following section 3.

The stationary distribution functions are shown in Figure 3.

For maximal density ($k = 1$) the stationary distribution function $f(v)$ is a δ function at $v = 0$. For the low density case the distribution function is a function

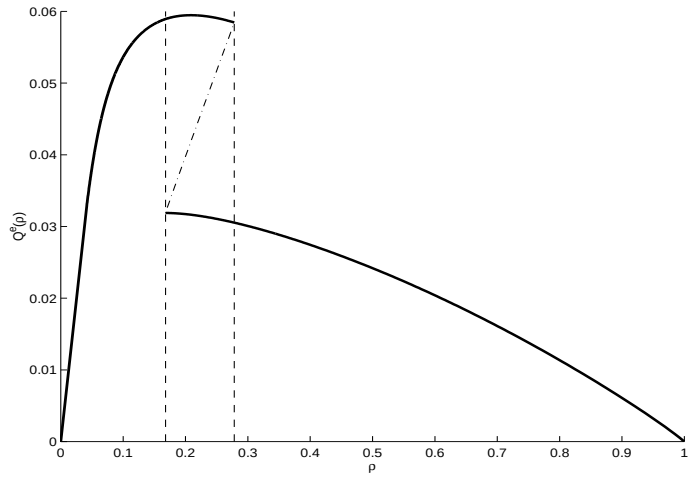


FIG. 6. *Fundamental diagram* $Q^e(\rho)$.

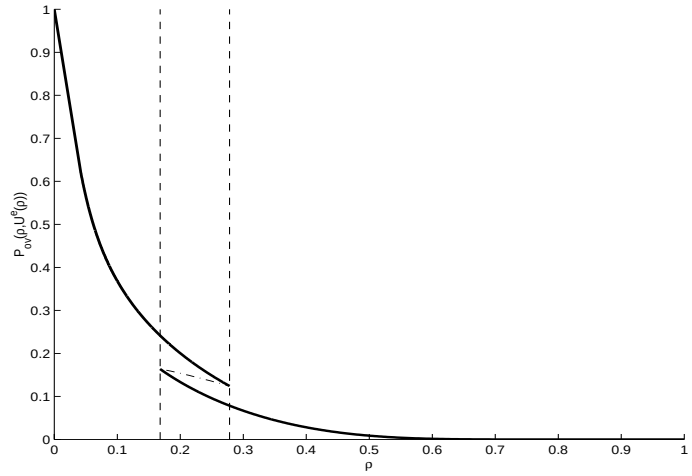


FIG. 7. *Probability* $P_{ov}(\rho, U^e(\rho))$.

with a certain positive variance, depending on the value of c at $k = 1$.

According to section 3, we can compute the function $u^e(k)$ from these stationary distributions in a straightforward way. The dependence of the function u^e on k is plotted in Figure 4.

Moreover, the results found for the nonlinear equation $u = u^e(k(\rho, u))$ are as follows. There are values ρ_1 and ρ_2 such that for $\rho < \rho_1$ we have only one solution, $U_e^1(\rho)$. For $\rho_1 < \rho < \rho_2$ three solutions $U_e^1(\rho)$, $U_e^2(\rho)$, and $\bar{U}(\rho)$ exist. For the region $\rho > \rho_2$ again only one solution $U_e^2(\rho)$ exists. Figure 5 shows the speed-density relation. We note that these values are contained in the admissible range of values (ρ, u) defined by (2). Figure 6 shows the associated multivalued fundamental diagram (flux-density relation).

The resulting multivalued probability of overtaking for equilibrium values of the velocity $P_{ov}(\rho, U^e(\rho))$ is plotted in Figure 7.

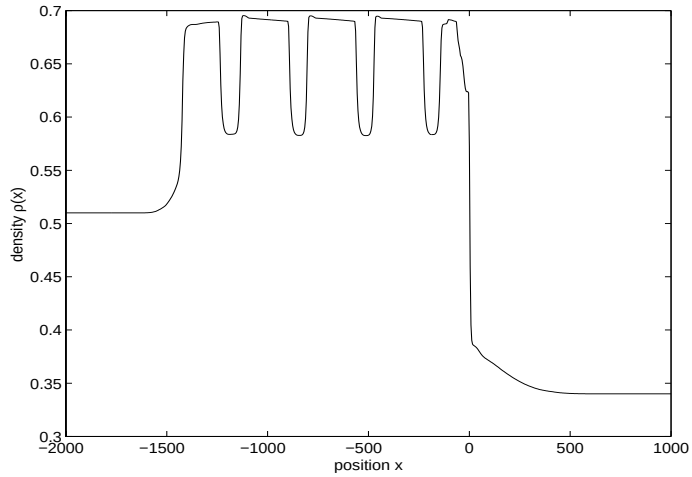


FIG. 8. *Stop and go waves, density* ρ.

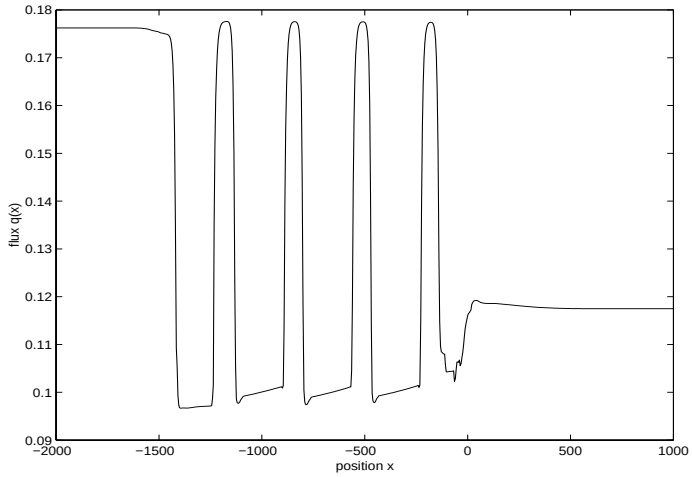


FIG. 9. *Stop and go waves, flux* $q = \rho u$.

As mentioned, $P_{ov}(\rho, U^e(\rho))$ can be compared to the corresponding figures in [12]. Good qualitative coincidence is observed here.

5.3. The spatially inhomogeneous case (macroscopic equations). Finally, the macroscopic equations are investigated for a bottleneck situation. For the computations we choose a second-order shock-capturing method as in [9]. Figures 8 and 9 show the density and flux for a three lane highway with a reduction of lanes from three to two at $x = 0$. One clearly observes large changes in density ρ and flux $q = \rho u$ in the backwards travelling traffic jam, which might be interpreted as stop and go behavior. As one can check numerically, the outgoing flux in Figure 9 is approximately equal to (correctly) averaged flux in the stop and go region.

Summary.

- Multivalued fundamental diagrams are obtained from kinetic equations which match—at least qualitatively—experimental observations in [11, 12]. In par-

ticular, the fundamental quantity, the probability of overtaking P_{ov}, shows good qualitative coincidence with the corresponding results here.

- Macroscopic traffic flow models are derived from the kinetic equation. These models are able to show stop and go patterns for highway traffic with a bottleneck.

REFERENCES

[1] A. Aw, *Modèles hyperboliques de trafic automobile*, Ph.D. thesis, Department of Mathamatics, Nice, 2001.

[2] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[3] C. Daganzo, *Requiem for second order fluid approximations of traffic flow*, Transportation Res. B, 29B (1995), pp. 277–286.

[4] J. M. Greenberg, *Extensions and amplifications of a traffic model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.

[5] J. M. Greenberg, A. Klar, and M. Rascle, *Congestion on multilane highways*, SIAM J. Appl. Math., 63 (2003), pp. 818–833.

[6] D. Helbing, *Gas-kinetic derivation of Navier–Stokes-like traffic equation*, Phys. Rev. E, 53 (1996), pp. 2366–2381.

[7] D. Helbing, *Traffic and related self-driven many-particle systems*, Rev. Modern Phys., 73 (2001), pp. 1067–1141.

[8] R. Illner, A. Klar, and T. Materne, *Vlasov–Fokker–Planck models for multilane traffic flow*, Comm. Math. Sci., 1 (2003), pp. 1–12.

[9] S. Jin and Z. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.

[10] B. Kerner, *Experimental features of self-organization in traffic flow*, Phys. Rev. Lett., 81 (1998), pp. 3797–3800.

[11] B. Kerner, *Congested traffic flow*, Transp. Res. Rec., 1678 (1999), pp. 160–165.

[12] B. Kerner, *Experimental features of the emergence of moving jams in free traffic flow*, J. Phys. A, 33 (2000), pp. 221–228.

[13] A. Klar and R. Wegener, *Enskog-like kinetic models for vehicular traffic*, J. Statist. Phys., 87 (1997), pp. 91–114.

[14] A. Klar and R. Wegener, *A hierarchy of models for multilane vehicular traffic* I: *Modeling*, SIAM J. Appl. Math., 59 (1998), pp. 983–1001.

[15] A. Klar and R. Wegener, *Kinetic derivation of macroscopic anticipation models for vehicular traffic*, SIAM J. Appl. Math., 60 (2000), pp. 1749–1766.

[16] P. Nelson, *A kinetic model of vehicular traffic and its associated bimodal equilibrium solutions*, Transport Theory Statist. Phys., 24 (1995), pp. 383–408.

[17] P. Nelson, and A. Sopasakis, *The Prigogine–Herman kinetic model predicts widely scattered traffic flow data at high concentrations*, Transportation Res. B, 32A (1998), pp. 589–604.

[18] S. Paveri-Fontana, *On Boltzmann like treatments for traffic flow*, Transportation Res., 9 (1975), pp. 225–235.

[19] H. Payne, *FREFLO: A macroscopic simulation model of freeway traffic*, Transportation Res. Record, 722 (1979), pp. 68–75.

[20] I. Prigogine and R. Herman, *Kinetic Theory of Vehicular Traffic*, Elsevier, New York, 1971.

[21] G. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 484–524

RADIATION INDUCED INSTABILITY*

PATRICK HAGERTY†, ANTHONY M. BLOCH†, AND MICHAEL I. WEINSTEIN‡

Abstract. In this paper we discuss the stability and instability properties of two classes of conservative dynamical systems which are comprised of a finite dimensional and an infinite dimensional subsystem. The finite dimensional component is a linear mechanical system with gyroscopic terms. The mechanical system is coupled to a wave equation defined on an infinite spatial domain via two different types of coupling which we denote by Lamb coupling and wave field coupling. We also investigate the effect of dispersion on the coupled system. In particular, we analyze the conditions under which coupling to a wave system induces instability in the finite dimensional system. Analytic results are compared to computer simulations.

Key words. conservative systems, dissipation, stability

AMS subject classifications. 35L10, 34D05

DOI. 10.1137/S0036139902418717

1. Introduction. Energy transfer within interconnected mechanical systems is important in many real world settings. When a finite dimensional system is coupled to an infinite dimensional system, energy can be radiated from the finite dimensional system and absorbed by the infinite dimensional system. We call this process *radiation damping*. For example, satellites and space stations have a central rigid body which can radiate energy through flexible components such as solar panels and antennae. Radiation damping can also describe dissipation (e.g., friction, viscosity) in a conservative context, where energy dissipates from one form (such as motion of a mechanical system) into another form (such as heat) of a larger conservative system. In general relativity, energy may be radiated from a system of masses via gravitational waves (see [23] for details). The Lagrangian or Hamiltonian structure of such systems is also of interest (see [21] or [4], for example).

An early physical model of radiation damping was introduced by Lamb [18]. In the Lamb model, an oscillator coupled to a string describes the free vibrations of a nucleus in an extended medium. The oscillator transfers energy to the string by generating waves as it moves; see also [12]. A model of a particle coupled to a wave field is studied in [15].

In many linear and nonlinear partial differential equations, it is fruitful to view the dynamics in terms of "particle-like" and "field-like" components. A decomposition into these types of modes leads to an equivalent description in terms of two coupled subsystems: the first is finite dimensional and governs the "particle-like" or bound state part of the solution, while the second is infinite dimensional and dispersive. Coupling terms are responsible for how the dynamics of "particles" influence the field and how the dispersive wave field influences the particle dynamics. An approxi-

*Received by the editors November 27, 2002; accepted for publication (in revised form) May 15, 2003; published electronically December 31, 2003.
http://www.siam.org/journals/siap/64-2/41871.html

†Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (hagerty@umich.edu, abloch@math.lsa.umich.edu). The research of the first author was partially supported by the National Science Foundation and the AFOSR. The research of the second author was partially supported by National Science Foundation grants DMS 981283, 0103895, and 305837 and by the AFOSR.

‡Fundamental Mathematics Research Department, Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974 (miw@research.bell-labs.com). The research of this author was partially supported by the National Science Foundation.

mate closed equation for the particle dynamics can, in some cases, be derived, where the effect of the dispersive radiation field appears in the form of radiation damping corrections to the finite dimensional dynamics. This approach has been used in the context of the theory of quantum resonances [28], [22], ionization and parametric resonance problems [14], [13], and in the context of the decay of "breather-like" states of nonlinear wave equations [29]. In this paper we explore how, in manner analogous to the mechanism by which dissipation can induce instability [5], radiation damping can induce instability in mechanical systems.

In section 2, we present the stability properties and Hamiltonian formalism of gyroscopic systems, also referred to as Chetaev systems. We also describe the Chetaev system as a normal form of the linearized equations of motion about a relative equilibrium of a simple mechanical system acted on by an abelian group. As we increase the gyroscopic forces in a Chetaev system, it is possible to stabilize (or specifically *gyroscopically* stabilize) an equilibrium of the system. For example, a charged, inverted spherical pendulum can be gyroscopically stabilized by increasing the strength of the ambient magnetic field.

In section 3, we describe a gyroscopic version of the Lamb model coupled to a standard nondispersive wave equation and to a dispersive wave equation, expanding on the results in [10]. We show that instabilities will arise in certain mechanical systems. However, dispersive wave coupling restricts the oscillator access to low frequency wave modes and allows for a band of stability to exist.

In the dispersionless case, the system is of the form

$$\begin{aligned} \frac{\partial^2 \mathbf{w}}{\partial t^2}(z,t) &= c^2 \frac{\partial^2 \mathbf{w}}{\partial z^2}(z,t), \quad z \in \mathbb{R} - \{0\}, t \in \mathbb{R}, \\ M\ddot{\mathbf{q}}(t) + S\dot{\mathbf{q}}(t) + V\mathbf{q}(t) &= T\Big[\frac{\partial \mathbf{w}}{\partial z}\Big]_{z=0}, \\ \mathbf{w}(0,t) &= \mathbf{q}(t), \end{aligned} \tag{1.1}$$

where c is the speed of transverse waves in the string, T is the tension of the string, $\mathbf{w} = \big[w_1(z,t) \dots w_n(z,t)\big]^T$ is the displacement of the string in the first n dimensions, and $[\frac{\partial \mathbf{w}}{\partial z}]_{z=0}$ is the jump discontinuity in the slope of the string.

In section 4, we introduce a nonlocal field coupling of a gyroscopic system to a dispersive or a nondispersive infinite dimensional system. Dispersive waves still allow for a band of stability, but the instabilities produced by the nondispersive wave coupling have a broader destabilizing effect on mechanical systems. In contrast to the analogous dissipation induced instability, field radiation induces instability in stable, as well as gyroscopically stable, mechanical systems. In this setting, the coupled wave equation and the gyroscopic system take the form

$$\begin{aligned} M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} &= \kappa \int_{\mathbb{R}} \chi(z) w(z,t) dz \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \\ \ddot{w} - c^2 \frac{\partial^2 w}{\partial z^2} &= \kappa \chi(z) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}^T \mathbf{q}, \end{aligned} \tag{1.2}$$

where κ is a coupling parameter and $\chi(\xi)$ is a suitable distribution.

In section 5 we consider an example of a more complicated finite dimensional mechanical system: a rigid body with rotors coupled to a wave field. In this case the configuration space of the system includes a nonabelian group, $SO(3)$.

Finally, we include an appendix which discusses some of details of the physical models used in this paper and the numerical techniques used in simulations.

2. Gyroscopic systems. We recall some general properties of linear systems with gyroscopic forces. The general form of a gyroscopic system is

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} = 0, \tag{2.1}$$

where $\mathbf{q} \in \mathbb{R}^n$, M is a positive definite symmetric $n \times n$ matrix, S is skew, and V is symmetric. As in [5], we shall call this the *Chetaev system* (see [8]). An important property of this system is that it is the normal form for a simple mechanical system about a *relative equilibrium* which is given modulo an abelian group; see [20] and section 5.

We say the system is *gyroscopically stable* if for $S = 0$ the origin is an unstable equilibrium, but for $S \neq 0$, the origin is a spectrally stable equilibrium (i.e., the eigenvalues of the linearized system have nonpositive real part). The matrix S is sometimes referred to as a magnetic term as it can arise from charged oscillators in a magnetic field.

If we include the magnetic terms in the symplectic form, then the Hamiltonian, $H : T^*Q \to \mathbb{R}$, of a gyroscopic system is the sum of the kinetic energy and potential energy,

$$H = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^T V\mathbf{q}. \tag{2.2}$$

The symplectic form, Ω, is defined by the Poisson bracket on the cotangent bundle T^*Q,

$$\Omega(X_F, X_K) = \{F, K\}_{magnetic} = \frac{\partial F}{\partial q^i}\frac{\partial K}{\partial p_i} - \frac{\partial K}{\partial q^i}\frac{\partial F}{\partial p_i} - S^{ij}\frac{\partial F}{\partial p_i}\frac{\partial K}{\partial p_j}, \tag{2.3}$$

where X_G denotes the Hamiltonian vector field with Hamiltonian G.

PROPOSITION 2.1. *The Hamiltonian vector field X_H on the symplectic vector space $(\mathbb{R}^{2n}, \Omega)$ has an equivalent representation $X_{\tilde{H}}$ on the symplectic vector space $(\mathbb{R}^{2n}, \tilde{\Omega})$, where*

$$\begin{aligned} \tilde{H} &= \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^T (V + W)\mathbf{q} + \mathbf{p}^T X\mathbf{q}, \\ \tilde{\Omega}(X_F, X_K) = \{F, G\}_{canonical} &= \frac{\partial F}{\partial q^i}\frac{\partial K}{\partial p_i} - \frac{\partial K}{\partial q^i}\frac{\partial F}{\partial p_i} \end{aligned} \tag{2.4}$$

for a skew matrix X and a symmetric matrix W defined implicitly by

$$XM + MX = -S, \qquad W = -XMX. \tag{2.5}$$

Furthermore, the Lyapunov equation (2.5) *can be solved using the Fredholm alternative.*

Proof. We obtain the implicit definitions of X and W by matching equations of motion on each manifold. Using the Fredholm alternative, we show that (2.5) has a

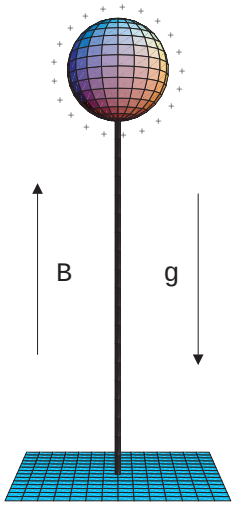


FIG. 1. *Inverted spherical pendulum.*

solution for X. Let us define the linear operator on skew matrices $L(X) = XM + MX$ and show that S is orthogonal to the kernel of L^T with the inner product on skew matrices being the trace $\langle A, B\rangle = \text{Trace}(AB)$. Computing L^T, we have $\langle A, L(B)\rangle = \langle L(A), B\rangle = \text{Trace}((MA + AM)B)$. Hence L is self-adjoint with respect to this inner product.

Suppose that Y is in the kernel of L^T; then we have

$$MY + YM = 0, \qquad Y = -M^{-1}YM\,. \tag{2.6}$$

We also have that $MY + (MY)^T = -2YM$, and hence YM is symmetric. Using the commuting property of the trace, we compute that S is orthogonal to Y, $\langle S, Y\rangle = 0$. The Fredholm alternative is satisfied and we can solve (2.5) for X. This completes the proof. □

Classically, the two representations of the Chetaev systems are equally useful. However, the classical symplectic form, $\tilde{\Omega}$, is preferable when quantizing the mechanical system. (See [5], [2], [8] for further physical discussions.)

2.1. Examples of Chetaev systems. Two physical models of Chetaev systems which can be gyroscopically stabilized are (i) an oscillator in a magnetic field and (ii) an oscillator on a rotating disk. We derive the dynamics from a Lagrangian and a Hamiltonian perspective.

2.1.1. Planar oscillator in a constant magnetic field. The linearized equations of motion of a charged spherical pendulum in a magnetic field are those of a charged planar oscillator in a magnetic field (see Figure 1). The equations of motion form a Chetaev mechanical system. (For derivation of the full equations see Appendix A.) We describe here the motion of a planar charged oscillator in a magnetic field. Let $\mathbf{B}$ be a divergence-free vector field. Let $\mathbf{A}$ be the vector potential $\mathbf{B} = \nabla \times \mathbf{A}$. Note that if we choose $\mathbf{B}$ to be the constant magnetic field in the direction normal to the plane of oscillation, the vector potential can be chosen as $\mathbf{A} = \frac{1}{2}\mathbf{B} \times \mathbf{q}$, where $\mathbf{q} = (x, y, 0)^T$ is the position of the oscillator.

Assume the oscillator has unit mass and unit charge and that the speed of light is unity. The Lagrangian, $L : T\mathbb{R}^2 \to \mathbb{R}$, is defined mechanically by the kinetic energy minus the potential energy:

$$\begin{aligned} L(\mathbf{q}, \dot{\mathbf{q}}) &= \frac{1}{2}\|\dot{\mathbf{q}}\|^2 + \mathbf{A} \cdot \dot{\mathbf{q}} - U(\mathbf{q}) \\ &= \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + \mathbf{A} \cdot (x, y, 0)^T - \frac{1}{2}(\alpha x^2 + \beta y^2). \end{aligned} \tag{2.7}$$

Choosing $\mathbf{B}$ to be of constant strength B normal to the plane of oscillation, we have

$$L = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) - \frac{1}{2}(\alpha x^2 + \beta y^2) + \frac{1}{2}B(x\dot{y} - y\dot{x}), \tag{2.8}$$

where the last term is the velocity dependent magnetic term.

The associated Hamiltonian, $H : T^*\mathbb{R}^2 \to \mathbb{R}$, is

$$H = p_x\dot{x} + p_y\dot{y} - L = \frac{1}{2}(p_x^2 + p_y^2) + \frac{1}{2}(\alpha x^2 + \beta y^2) + H_B, \tag{2.9}$$

where

$$H_B = \frac{1}{2}B(p_x y - p_y x) + \frac{1}{8}B^2(x^2 + y^2), \tag{2.10}$$

and the associated momenta are given by

$$p_x = \frac{\partial L}{\partial \dot{x}} = \dot{x} - \frac{1}{2}By, \qquad p_y = \frac{\partial L}{\partial \dot{y}} = \dot{y} + \frac{1}{2}Bx\,. \tag{2.11}$$

In the above notation, H_B is the contribution of a magnetic field. Since we include the magnetic terms in the Hamiltonian, we use the canonical symplectic form, $\tilde{\Omega}$, to obtain the equations of motion:

$$\begin{aligned} -\dot{p}_x = \frac{\partial H}{\partial x} = \left(\alpha + \frac{1}{4}B^2\right)x - \frac{1}{2}Bp_y, \qquad -\dot{p}_y = \frac{\partial H}{\partial y} = \left(\beta + \frac{1}{4}B^2\right)y + \frac{1}{2}Bp_x, \\ \dot{x} = \frac{\partial H}{\partial p_x} = p_x + \frac{1}{2}By, \qquad \dot{y} = \frac{\partial H}{\partial p_y} = p_y - \frac{1}{2}Bx\,. \end{aligned} \tag{2.12}$$

We thus have the dynamics

$$\begin{aligned} \ddot{x} - B\dot{y} + \alpha x = 0, \\ \ddot{y} + B\dot{x} + \beta y = 0\,. \end{aligned} \tag{2.13}$$

If α and β are both negative, the oscillator is gyroscopically stabilized (i.e., the eigenvalues are purely imaginary) if $B^2 + \alpha + \beta > 2\sqrt{\alpha\beta}$. Hence we can increase the strength of the magnetic field to stabilize the oscillator.

2.1.2. Planar oscillator on a rotating plate. Another physical model of a Chetaev system is a planar oscillator on a plate rotating with angular velocity ω (see Figure 2). We write the Lagrangian as the kinetic energy minus the potential energy,

$$L = \frac{1}{2}((\dot{x} - \omega y)^2 + (\dot{y} + \omega x)^2) - \frac{1}{2}(\alpha x^2 + \beta y^2)\,. \tag{2.14}$$

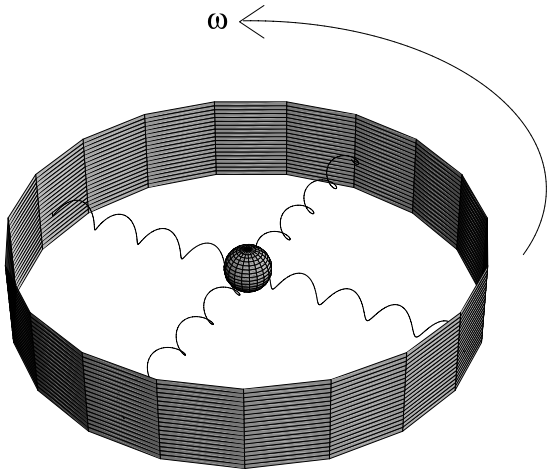


FIG. 2. *Rotating plate with springs.*

In the Hamiltonian setting we have

$$H = \frac{1}{2}(p_x^2 + p_y^2) + \frac{1}{2}(\alpha x^2 + \beta y^2) + H_\omega, \tag{2.15}$$

$$H_\omega = \omega(p_x y - p_y x), \tag{2.16}$$

and the associated momenta are given by

$$p_x = \frac{\partial L}{\partial \dot{x}} = \dot{x} - \omega y, \qquad p_y = \frac{\partial L}{\partial \dot{y}} = \dot{y} + \omega x. \tag{2.17}$$

We obtain the equations of motion:

$$\begin{aligned} -\dot{p}_x = \frac{\partial H}{\partial x} = \alpha x - \omega p_y, \qquad -\dot{p}_y = \frac{\partial H}{\partial y} = \beta y + \omega p_x, \\ \dot{x} = \frac{\partial H}{\partial p_x} = p_x + \omega y, \qquad \dot{y} = \frac{\partial H}{\partial p_y} = p_y - \omega x. \end{aligned} \tag{2.18}$$

This gives the dynamics

$$\begin{aligned} \ddot{x} - 2\omega\dot{y} + (\alpha - \omega^2)x = 0, \\ \ddot{y} + 2\omega\dot{x} + (\beta - \omega^2)y = 0. \end{aligned} \tag{2.19}$$

Remark 1. The rotating disc affects the oscillator differently from a magnetic field—compare (2.19) and (2.13). While the magnetic field only adds $\dot{\mathbf{q}}$ terms, the rotating disc also adds additional $\mathbf{q}$ terms to the dynamics. For the equations (2.13), we can see that in the case of a physically stable oscillator with $\alpha, \beta > 0$ for rotation rate ω sufficiently large, the system becomes only gyroscopically stable, i.e., the coefficients $\alpha - \omega^2$ and $\beta - \omega^2$ are negative but the eigenvalues are on the imaginary axis due to the presence of gyroscopic terms.

Remark 2. In either case above, if the system is gyroscopically stable, it can be shown that adding a small amount of dissipation to the system renders it unstable (i.e., there are unstable eigenvalues). For a more precise statement and generalization, see section 2.2.

2.2. Stability of Chetaev systems. In this section we first summarize the stability properties of Chetaev systems and then discuss their dissipative perturbations. The stability of the Chetaev system depends both on the signature of the bilinear form associated with a quadratic Hamiltonian and also on the magnetic terms in the symplectic form. In particular, the magnetic terms can stabilize gyroscopic systems with negative eigenvalues of V.

As for gyroscopic stability, the number of negative eigenvalues of the quadratic form plays a crucial part as discussed in [8]. We summarize this in the following proposition.

PROPOSITION 2.2. *Consider the canonical gyroscopic system $M\ddot{\mathbf{q}}+S\dot{\mathbf{q}}+V\mathbf{q}=0$, where M is a symmetric positive definite matrix, S is a skew-symmetric matrix, and V is a symmetric matrix:*

- *If V has an odd number of negative eigenvalues (counting multiplicity), then the origin is an unstable equilibrium.*
- *If V has an even number of negative eigenvalues (counting multiplicity), we can choose S so that the origin is a spectrally stable equilibrium.*

Proof. By standard reduction to a first order system for $(\mathbf{q}, \dot{\mathbf{q}})$, it suffices to consider the first order linear operator L,

$$L = \begin{bmatrix} 0 & I \\ -M^{-1}V & -M^{-1}S \end{bmatrix}. \tag{2.20}$$

Let $p(\lambda)$ be the characteristic polynomial of the matrix L, and let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of the matrix V.

$$p(0) = \det L = (-1)^n \det(-M^{-1}V) = \det(M^{-1}) \prod_{i=1}^{n} \lambda_i \,. \tag{2.21}$$

Suppose that V has an odd number of negative eigenvalues. Due to the positive-definiteness of M, we have $p(0) < 0$. Since

$$\lim_{\lambda\to\infty} p(\lambda) = \lim_{\lambda\to\infty} \det\Big(L - \lambda I\Big) = \lim_{\lambda\to\infty} (-1)^{2n}\lambda^{2n} + O(\lambda^{2n-1}) = \infty, \tag{2.22}$$

$p(\lambda)$ must change sign and by continuity has a positive root, corresponding to a real and positive eigenvalue of L, an instability.

Hence, if V has an odd number of negative eigenvalues (counting multiplicity), then the characteristic polynomial of L has a positive real root and the origin cannot be gyroscopically stabilized.

Now to prove the second item, suppose that V has an even number of negative eigenvalues (counting multiplicity). By choosing a basis in which V is diagonal, we can introduce a skew matrix S to stabilize each pair of negative eigendirections, as in the gyroscopic stabilization for $n = 2$ shown in introductory examples. This completes the proof. ∎

Gyroscopically stable systems exhibit interesting instability when perturbed by dissipative forces. Suppose now that V has at least one negative eigenvalue. A key result of [5] is that adding small dissipation always yields instability. More precisely, we show the following.

THEOREM 2.3. *Under the above conditions, if we modify the general Chetaev system by adding a small Rayleigh dissipation term,*

$$M\ddot{\mathbf{q}} + (S + \epsilon R)\dot{\mathbf{q}} + V\mathbf{q} = 0 \tag{2.23}$$

for small $\epsilon > 0$, where R is symmetric and positive definite, then the perturbed linearized equations

$$\dot{z} = L_\epsilon z,$$

where $z = (q, p)$, are spectrally unstable, i.e., at least one pair of eigenvalues of L_ϵ is in the right half plane.

This result builds on basic work of [30], [8], and [11]. We refer to this as *dissipation induced instability*. Some of the radiation induced instabilities that arise in the gyroscopic Lamb model are analogous to the dissipation induced instability. Using the Hamiltonian representation with canonical symplectic form, $\tilde{\Omega}$, we see that the Hamiltonian of a gyroscopically stabilized Chetaev system is indefinite. In this case, with the addition of Rayleigh dissipation, the Hamiltonian does not bound the motion of the Chetaev system. In particular, the displacement and velocities may grow exponentially.

As in section 5.1.2, we present the Chetaev system as a normal form of the linearized equations of motion about the relative equilibrium of an abelian group action. [5] also proves a similar stability result for the nonabelian case, but the abelian result is sufficient for our purposes.

3. Lamb coupling. In the original Lamb model [18], an oscillator is physically coupled to a string (see Figure 3). The vibrations of the oscillator transmit waves into the string and are carried off to infinity. Hence, the oscillator loses energy and is effectively damped by the string.

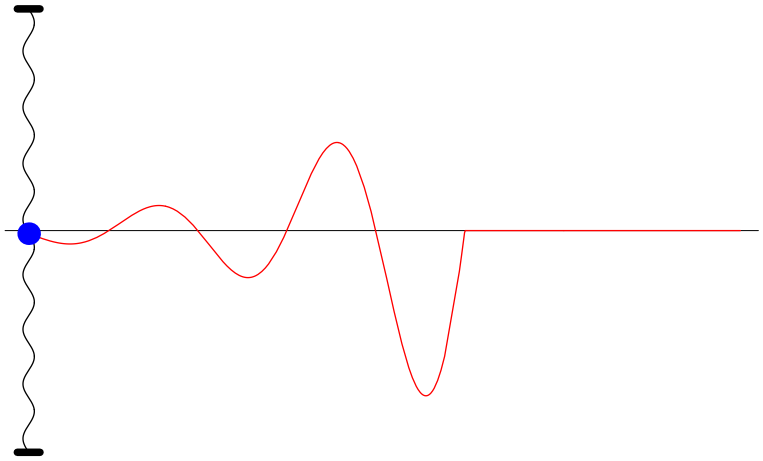


FIG. 3. *Lamb model of an oscillator coupled to a string.*

Let $w(x, t)$ be the displacement of the string at position $x \in \mathbb{R}$ at time t. Let ρ be the mass density of the string, and let T be the tension. We can compute the wave speed of transverse oscillations of the free string to be $c = \frac{T}{\rho}$. Assuming a singular mass density at $x = 0$, we can couple the dynamics of an oscillator, q, of mass M, to the otherwise free wave to resolve the singularity:

$$\begin{aligned} \frac{\partial^2 w}{\partial t^2} &= c^2 \frac{\partial^2 w}{\partial x^2}, \\ M\ddot{q} + Vq &= T[w_x]_{x=0}, \\ q(t) &= w(0, t), \end{aligned} \tag{3.1}$$

where $\sqrt{V/M}$ is the frequency of the uncoupled oscillating mass M and $[w_x]_{x=0} = w_x(0+,t) - w_x(0-,t)$ is the jump discontinuity of the slope of the string. Note that (3.1) is a Hamiltonian system, as demonstrated for a more complex system in section 3.1.

Perturbing the oscillator at time $t = 0$ from its equilibrium position, we can use the d'Alembert solution to the wave equation to solve for w,

$$w = \begin{cases} Ce^{(ct-|x|)\omega} & \text{for } |x| < ct, \\ 0 & \text{for } |x| > ct, \end{cases} \tag{3.2}$$

where $\omega = -\frac{T}{Mc} + i\sqrt{\frac{V^2}{M^2} - (\frac{T}{Mc})^2}$ and C is the size of the initial displacement. For small tension we have damping and oscillator motion, while for large tension, we have pure damping. From the solution of the wave equation, we can compute the jump condition of the oscillator,

$$[w_x]_{x=0} = -2C\omega e^{\omega ct} = -\frac{2}{c}\dot{q}. \tag{3.3}$$

We obtain a reduced form of the dynamics describing the explicit motion of the oscillator subsystem,

$$M\ddot{q} + \frac{2T}{c}\dot{q} + Vq = 0. \tag{3.4}$$

The coupling term arises explicitly as a Rayleigh dissipation term $\frac{2T}{c}\dot{q}$ in the dynamics of the oscillator. For the gyroscopic Lamb coupling, the dissipation term will induce instabilities in gyroscopically stabilized Chetaev systems.

3.1. Gyroscopic Lamb model. In this section, we investigate a variant of the Lamb model, a Chetaev system which includes gyroscopic terms coupled to the standard wave equation (see, e.g., Figure 4). We have the boundary constraint that the displacement of the gyroscopic oscillator fixes a point of the string. We show that this local coupling perturbation destabilizes gyroscopically stabilized Chetaev systems.

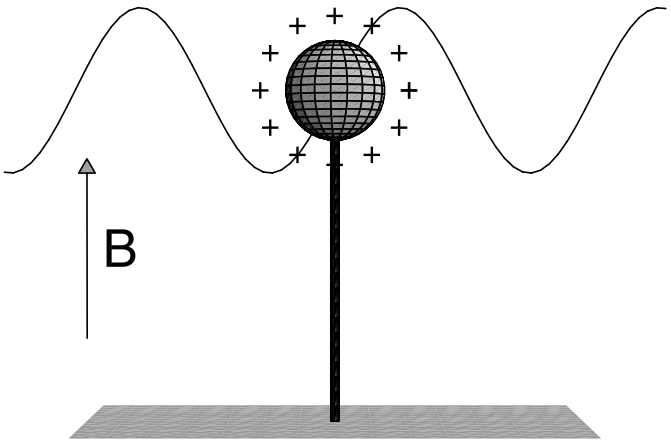


FIG. 4. *Gyroscopic Lamb coupling to a spherical pendulum.*

We begin with a model of a string in $\mathbb{R}^{n+1}$, whose transverse vibrations are independent. Suppose the string lies initially along the x_{n+1}-axis. For simplicity, we denote the x_{n+1} dimension as the z dimension. Coupling the transverse motion to an n dimensional gyroscopic system, we use the method described in [16] to solve the following:

$$\begin{aligned} \frac{\partial^2 \mathbf{w}}{\partial t^2}(z,t) &= c^2 \frac{\partial^2 \mathbf{w}}{\partial z^2}(z,t), \quad z \in \mathbb{R} - \{0\}, t \in \mathbb{R}, \\ M\ddot{\mathbf{q}}(t) + S\dot{\mathbf{q}}(t) + V\mathbf{q}(t) &= T\Big[\frac{\partial \mathbf{w}}{\partial z}\Big]_{z=0}, \\ \mathbf{w}(0,t) &= \mathbf{q}(t), \end{aligned}$$

where c is the speed of transverse waves in the string, T is the tension of the string, $\mathbf{w} = \big[w_1(z,t) \dots w_n(z,t)\big]^T$ is the displacement of the string in the first n dimensions, and $[\frac{\partial \mathbf{w}}{\partial z}]_{z=0}$ is the jump discontinuity in the slope of the string. Let us define the initial conditions on the string by $\mathbf{w}_1(z) = \frac{\partial \mathbf{w}}{\partial t}(z,0)$ and $\mathbf{w}_0(z) = \mathbf{w}(z,0)$.

By direct computation we obtain an exact reduced dynamical system for $\mathbf{q}$. This reduction contains explicit dissipation terms reflecting the oscillator/field coupling. The results on reduction and stability are summarized in the following two results. The proofs are given in section 3.1.2.

PROPOSITION 3.1. *If the initial data* $\mathbf{w}_0$ *and* $\mathbf{w}_1$ *have compact support, then for* $\pm ct \notin \operatorname{supp}(\mathbf{w}_0) \cup \operatorname{supp}(\mathbf{w}_1)$ *the oscillator dynamics of the gyroscopic Lamb model reduces to*

$$M\ddot{\mathbf{q}}(t) + \left(S + \frac{2T}{c}\right)\dot{\mathbf{q}}(t) + V\mathbf{q}(t) = 0. \tag{3.5}$$

In [5], small dissipation is shown to induce instability in gyroscopically stable systems (see Theorem 2.3). In the gyroscopic Lamb model, radiation (energy transfer into an infinite dimensional system) produces a Rayleigh dissipation term which depends on the tension of the string. Since dissipation induces instability in these systems for small dissipation, small tension coupling yields radiation induced instability via an analogous mechanism. We summarize the main result of *radiation induced* instability in gyroscopically stabilized Chetaev systems with a theorem.

THEOREM 3.2. *If a gyroscopic mechanical system is gyroscopically stable (i.e., V has a negative eigendirection), then local coupling via the Lamb model induces instability for small coupling parameter T.*

The ability to describe the coupled system in a conservative manner is one of the key differences between radiation induced instability and dissipation induced instability. We delay the proofs of the above theorems to demonstrate the conservative nature of the gyroscopic Lamb model. While we can obtain the equations of motion from a force calculation, we emphasize the conservative nature of the gyroscopic Lamb model by representing the system as the flow of a Hamiltonian vector field.

3.1.1. Hamilton's equations of motion. For an uncoupled wave equation, the kinetic energy, T_{string}, depends on the velocity of the string, while the potential energy, U_{string}, depends on the shape of the string,

$$T_{string} = \sum_{i=1}^{n} \int_{\mathbb{R}} \left\| \frac{\partial w_i}{\partial t} \right\|^2 dz, \quad U_{string} = \sum_{i=1}^{n} \int_{\mathbb{R}} c^2 \left\| \frac{\partial w_i}{\partial z} \right\|^2 dz. \tag{3.6}$$

Let $Z = \mathcal{H}^1(\mathbb{R}) \times L^2(\mathbb{R})$, where $\mathcal{H}^1(\mathbb{R})$ is the space of square integrable functions whose first derivatives are also square integrable and $L^2(\mathbb{R})$ is the space of square integrable densities. Hence $(w_i, \pi_i) \in Z$ if w_i and $\frac{\partial w_i}{\partial z}$ are each square integrable and $\pi_i = \pi_i' dz$, where π' is a square integrable function on $\mathbb{R}$. We define the canonical symplectic form, Ω_Z, on the vector space Z as

$$\Omega_Z((w_1, \pi_1), (w_2, \pi_2)) = \int_{\mathbb{R}} w_1 \pi_2 - w_2 \pi_1. \tag{3.7}$$

PROPOSITION 3.3. *The gyroscopic Lamb model* (3.1) *has a Hamiltonian representation with Hamiltonian* $(H, (Z, \Omega_Z)^n)$, *with*

$$\begin{aligned} H(\mathbf{w}, \pi) &= \int \sum_{i=1}^{n} \frac{1}{2}(\pi_i')^2 + \frac{1}{2}c^2 \left\| \frac{\partial w_i}{\partial z} \right\|^2 + \delta(z) H_{osc.}(\mathbf{w}, \pi') dz, \\ H_{osc.}(\mathbf{w}, \pi') &= \frac{1}{2}\pi'^T M^{-1} \pi' + \frac{1}{2}\mathbf{w}^T (V + W)\mathbf{w} + \pi'^T X \mathbf{w} - T\mathbf{w}^T [\mathbf{w}_z]_{z=0}. \end{aligned} \tag{3.8}$$

Proof. The symplectic form induces Hamilton's equations for field theory,

$$\Omega(X_H, \cdot) = dH, \tag{3.9}$$

$$\frac{\partial w_i}{\partial t} = \frac{\delta H}{\delta \pi_i}, \quad \frac{\partial \pi_i}{\partial t} = -\frac{\delta H}{\delta w_i}. \tag{3.10}$$

Computing the variations, we have

$$\begin{aligned} \int \frac{\delta H}{\delta \pi_i} \cdot \delta \pi_i dz &= \lim_{\epsilon \to 0} \epsilon^{-1} (H(\mathbf{w}, \pi + \epsilon \delta \pi_i) - H(\mathbf{w}, \pi)) \\ &= \int \left(\pi_i' + \delta(z) \frac{\partial H_{osc.}}{\partial \pi_i} \right) \cdot \delta \pi_i dz \end{aligned} \tag{3.11}$$

and

$$\begin{aligned} \int \frac{\delta H}{\delta w_i} \cdot \delta w_i dz &= \lim_{\epsilon \to 0} \epsilon^{-1} (H(\mathbf{w} + \epsilon \delta w_i, \pi) - H(\mathbf{w}, \pi)) \\ &= \int \left(-c^2 \frac{\partial^2 w_i}{dz^2} + \delta(z) \frac{\partial H_{osc.}}{\partial w_i} \right) \cdot \delta w_i dz. \end{aligned} \tag{3.12}$$

For $z \neq 0$, we have

$$\frac{\partial w_i}{\partial t} = \pi_i', \quad \frac{\partial \pi_i'}{\partial t} = c^2 \frac{\partial^2 w_i}{\partial z^2}, \tag{3.13}$$

or equivalently

$$\frac{\partial^2 \mathbf{w}}{\partial t^2}(z, t) = c^2 \frac{\partial^2 \mathbf{w}}{\partial z^2}(z, t), \quad z \in \mathbb{R} - \{0\}, t \in \mathbb{R}. \tag{3.14}$$

For the singular part of the Hamiltonian at $z = 0$, we define $\mathbf{q}(t) = \mathbf{w}(0,t)$ and $\mathbf{p}(t) = \pi'(0,t)$. Notice that $[\mathbf{w}_z]_{z=0}$ is independent of variations $\delta\mathbf{w}(0)$. We use the induced canonical symplectic form, $\tilde{\Omega}$, on $T^*\mathbb{R}^n$ to obtain the Hamilton's equations of motion:

$$\begin{aligned}\dot{\mathbf{q}} &= \frac{\partial H_{osc.}}{\partial \mathbf{p}} = M^{-1}\mathbf{p} + X\mathbf{q},\\ \dot{\mathbf{p}} &= \frac{\partial H_{osc.}}{\partial \mathbf{q}} = (V+W)\mathbf{q} - X\mathbf{p} - T[\mathbf{w}_z]_{z=0}.\end{aligned} \tag{3.15}$$

Differentiating and writing the above first order system as a second order system, we obtain

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} = T[\mathbf{w}_z]_{z=0}. \tag{3.16}$$

This completes the proof. □

3.1.2. Radiation induced instability in the Lamb model. To prove radiation induced instability, we obtain a reduced form of the Chetaev subsystem. We can use the d'Alembert method of solving the wave equation or find the solution via Laplace transforms. We save the latter technique for the more general dispersive case.

Proof of Proposition 3.1. [16] explicitly develops the d'Alembert solution of the wave equation coupled to a simple harmonic oscillator. We use the same method applied to a Chetaev system coupled to a higher dimensional wave equation. The method is to use the d'Alembert method of solving the wave equation on the domain unaffected by the coupling. Inside the domain affected by the coupling, we use continuity and coupling dynamics to solve for the wave equation.

Our model decomposes into n independent one dimensional wave equations; we can use the d'Alembert decomposition to two traveling waves:

$$\mathbf{u}(z,t) = \mathbf{f}_\pm(z - ct) + \mathbf{g}_\pm(z + ct), \qquad \pm z > 0, \tag{3.17}$$

where $\mathbf{f}_\pm, \mathbf{g}_\pm$ are functions on $\mathbb{R}^n$. In components, we have

$$\mathbf{f}_\pm = \left[f_{\pm 1} \dots f_{\pm 2n}\right]^T, \quad \mathbf{g}_\pm = \left[g_{\pm 1} \dots g_{\pm 2n}\right]^T, \tag{3.18}$$

where $f_i, g_i, i = \pm 1, \dots, \pm n$ are real-valued functions on $\mathbb{R}$. By the d'Alembert method of solving the wave equation, we have the formulas

$$\mathbf{f}_\pm(z) = \frac{1}{2}\mathbf{w}_0(z) - \frac{1}{2c}\int_0^z \mathbf{w}_1(v)dv, \qquad \pm z > 0, \tag{3.19}$$

$$\mathbf{g}_\pm(z) = \frac{1}{2}\mathbf{w}_0(z) + \frac{1}{2c}\int_0^z \mathbf{w}_1(v)dv, \qquad \pm z > 0. \tag{3.20}$$

Since the d'Alembert formula for $|z| \geq c|t|$ is defined in terms of known functions, the method remains valid and we have

$$\mathbf{w}(z,t) = \frac{\mathbf{w}_0(z-ct) + \mathbf{w}_0(z+ct)}{2} + \frac{1}{2c}\int_{z-ct}^{z+ct} \mathbf{w}_1(v)dv. \tag{3.21}$$

To use the d'Alembert formula for $|z| < c|t|$, we need to define $\mathbf{f}_+(z)$ for $z < 0$ and to define $\mathbf{g}_-(z)$ for $z > 0$. From the boundary condition and from continuity of the string, we have

$$\mathbf{f}_\pm(-ct) + \mathbf{g}_\pm(ct) = \mathbf{q}(t), \qquad t > 0, \tag{3.22}$$

$$M\ddot{\mathbf{q}}(t) = -S\dot{\mathbf{q}}(t) - V\mathbf{q}(t) + T\left[\frac{\partial \mathbf{f}_+}{\partial z}(-ct) + \frac{\partial \mathbf{g}_+}{\partial z}(ct) - \frac{\partial \mathbf{f}_-}{\partial z}(-ct) - \frac{\partial \mathbf{g}_-}{\partial z}(ct)\right]. \tag{3.23}$$

Differentiating and solving for $\mathbf{f}'_+$ and $\mathbf{g}'_-$, where the $'$ denotes $\frac{\partial}{\partial z}$, we have

$$\begin{aligned} -c\mathbf{f}'_+(-ct) &= \dot{\mathbf{q}}(t) - c\mathbf{g}'_+(ct), \\ c\mathbf{g}'_-(ct) &= \dot{\mathbf{q}}(t) + c\mathbf{f}'_-(-ct), \qquad t > 0. \end{aligned} \tag{3.24}$$

Substituting, we arrive at a differential equation for $\mathbf{q}$ in terms of known functions

$$\begin{aligned} M\ddot{\mathbf{q}}(t) + S\dot{\mathbf{q}}(t) + V\mathbf{q}(t) &= 2T\left[\mathbf{g}'_+(ct) - \mathbf{f}'_-(-ct) - \frac{1}{c}\dot{\mathbf{q}}(t)\right], \quad t > 0. \\ &= -\frac{2T}{c}\dot{\mathbf{q}}(t) + T\left[\mathbf{w}'_0(ct) - \mathbf{w}'_0(-ct) + \frac{1}{c}(\mathbf{w}_1(ct) + \mathbf{w}_1(-ct))\right]. \end{aligned} \tag{3.25}$$

The effect of coupling to a wave equation (3.25) is manifested by the presence of an explicit dissipative term. If we assume that the initial conditions of the string are such that $\mathbf{w_1}$ and $\mathbf{w_0}$ both have compact support, then for large t we have $\pm ct$ outside the support of $|\mathbf{w_0}| + |\mathbf{w_1}|$, reducing (3.25) to

$$\begin{aligned} M\ddot{\mathbf{q}}(t) + S\dot{\mathbf{q}}(t) + V\mathbf{q}(t) &= -\frac{2T}{c}\dot{\mathbf{q}}(t), \\ \pm ct &\notin \operatorname{supp}(|\mathbf{w_0}| + |\mathbf{w_1}|). \end{aligned} \tag{3.26}$$

This completes the proof the of proposition. ∎

Proof of Theorem 3.2. Assuming that the initial string has a finite energy, we have that $\mathbf{w}$ and $\mathbf{w_1}$ decay as $|z| \to \infty$. Hence, there are initial conditions of the string which have compact support in an H^1 Sobolev neighborhood of $\mathbf{w}$. Hence we can use the result from Proposition 3.1. We construct a Lyapunov function W as the sum of the energy of the system and a magnetic term,

$$W(\mathbf{q}) = \frac{1}{2}\dot{\mathbf{q}}^T M\dot{\mathbf{q}} + \frac{1}{2}\mathbf{q}^T V\mathbf{q} + \delta\mathbf{q}^T VM\dot{\mathbf{q}}. \tag{3.27}$$

In [5], W is shown to have a negative eigendirection for δ sufficiently small if V has a negative eigendirection, and that $\dot{W}$ is negative definite. Invoking the Lyapunov instability theorem, the gyroscopic Lamb model is spectrally unstable if V has a negative eigendirection. ∎

The results from Theorem 3.2 are consistent with Kreĭn signature results on instability (see [1] and [5]), in contrast to the broader instability one sees in the nonlocal coupling case (see section 4).

3.2. Dispersive gyroscopic Lamb model. In this section, we develop an explicit solution to a dispersive gyroscopic Lamb model, i.e., a gyroscopically stable Chetaev system coupled to a dispersive wave equation. As the dispersive wave equation we choose the Klein–Gordon equation with "mass" m:

$$\begin{aligned} \frac{\partial^2 \mathbf{w}}{\partial t^2}(z,t) &= c^2 \frac{\partial^2 \mathbf{w}}{\partial z^2}(z,t) - m^2 \mathbf{w}(z,t), \quad z \in \mathbb{R} - \{0\}, t \in \mathbb{R}, \\ M\ddot{\mathbf{q}}(t) + S\dot{\mathbf{q}}(t) + V\mathbf{q}(t) &= T\Big[\frac{\partial \mathbf{w}}{\partial z}\Big]_{z=0}, \\ \mathbf{w}(0,t) &= \mathbf{q}(t). \end{aligned} \tag{3.28}$$

The coupled system has a Hamiltonian, H, given by

$$H(\mathbf{w},\pi) = \int \sum_{i=1}^{n} \frac{1}{2}(\pi_i')^2 + \frac{1}{2}c^2 \left\| \frac{\partial w_i}{\partial z} \right\|^2 + \frac{1}{2}m^2 \|w\|^2 + \delta(z) H_{osc.}(\mathbf{w},\pi') dz, \tag{3.29}$$

$$H_{osc.}(\mathbf{w},\pi') = \frac{1}{2}\pi'^T M^{-1} \pi' + \frac{1}{2}\mathbf{w}^T (V+W)\mathbf{w} + \pi'^T X \mathbf{w} - T\mathbf{w}\Big[\mathbf{w}_z\Big]_{z=0}$$

on the same symplectic vector space $((Z,\Omega_Z)^n)$ as the gyroscopic Lamb model.

Equation (3.28) is linear translation invariant in time, so the initial value problem can be solved by a Laplace transform, which we denote by

$$\mathcal{L}[f](s) = \int_0^\infty f(t) e^{st} dt.$$

In the transformed space, we have

$$\begin{aligned} (s^2 M + sS + V)\mathcal{L}[\mathbf{q}] &= M(\dot{\mathbf{q}}(0) + s\mathbf{q}(0)) + s\mathbf{q}(0) + T\Big[\mathcal{L}[\mathbf{w}_z]\Big]_{z=0}, \\ (s^2 + m^2)\mathcal{L}[\mathbf{w}] - c^2 \mathcal{L}[\mathbf{w}_{zz}] &= \mathbf{w}_t(z,0) + s\mathbf{w}(z,0), \quad z \neq 0, \\ \mathcal{L}[\mathbf{w}](0,s) &= \mathcal{L}[\mathbf{q}](s). \end{aligned} \tag{3.30}$$

For simplicity of the calculation, we assume that $\mathbf{q}(0) = \mathbf{w}_t(z,0) = \mathbf{w}(z,0) = 0$ and that the perturbation comes from a nontrivial velocity of the oscillator at the origin. We can integrate the homogeneous dispersive wave equation on each interval to obtain

$$\mathcal{L}[\mathbf{w}](z,s) = \begin{cases} \mathbf{a} e^{\frac{\sqrt{s^2+m^2}\, z}{c}} & \text{for } z < 0, \\ \mathbf{a} e^{-\frac{\sqrt{s^2+m^2}\, z}{c}} & \text{for } z > 0. \end{cases} \tag{3.31}$$

From the boundary constraints, we have $\mathcal{L}[\mathbf{q}] = \mathbf{a}$.

Solving for $\mathcal{L}[\mathbf{w}]$ as a function of $\mathcal{L}[\mathbf{q}]$ and computing the jump condition of the string at $z = 0$, we have

$$\begin{aligned} \mathcal{L}[\mathbf{w}_z](z,s) &= -\operatorname{sign}(z) \frac{\sqrt{s^2+m^2}}{c} e^{-\operatorname{sign}(z)\frac{\sqrt{s^2+m^2}\, z}{c}} \mathcal{L}[\mathbf{q}], \\ \Big[\mathcal{L}[\mathbf{w}_z]\Big]_{z=0} &= -2\frac{\sqrt{s^2+m^2}}{c} \mathcal{L}[\mathbf{q}]. \end{aligned} \tag{3.32}$$

The equations of motion for the coupled Chetaev system decouple from the dispersive wave equation,

$$\left(s^2 M + sS + V + 2T\frac{\sqrt{s^2+m^2}}{c} I \right) \mathcal{L}[\mathbf{q}] = M\dot{\mathbf{q}}(0). \tag{3.33}$$

Taking the inverse Laplace transform and using the identity

$$\mathcal{L}\left[\frac{J_1(t)}{t}\right] = \sqrt{s^2+1} - s, \tag{3.34}$$

we obtain

$$M\ddot{\mathbf{q}} + \left(S + \frac{2T}{c}I\right)\dot{\mathbf{q}} + V\mathbf{q} = -\frac{2Tm}{c}\int_0^t \frac{1}{t-s}J_1(m(t-s))\mathbf{q}(s)ds, \tag{3.35}$$

where J_1 is the Bessel function of the first kind. We can define the approximate identity $J^\epsilon = \frac{J_1(\frac{t}{\epsilon})}{t}$. Since J_1 is an $L^1(\mathbb{R})$ function such that $\int_0^\infty J_1(t)dt = 1$, we can use the approximate identity nature of the convolution to have a necessary condition on stability summarized in a lemma.

LEMMA 3.4. *If* $\mathbf{q}$ *is continuous and bounded for all time, then*

$$\lim_{\epsilon\to 0} J^\epsilon * \mathbf{q}(t) = \mathbf{q}(t). \tag{3.36}$$

By the above lemma, we obtain an asymptotic expansion for the perturbed dynamics of the coupled oscillator.

PROPOSITION 3.5. *Consider the dynamical system*

$$M\ddot{\mathbf{q}} + \left(S + \frac{2T}{c}I\right)\dot{\mathbf{q}} + \left(V + \frac{Tm}{c}I\right)\mathbf{q} = 0. \tag{3.37}$$

As m *tends to infinity, the zero equilibrium of the dispersive Lamb model becomes unstable if the zero equilibrium of system* (3.37) *is unstable.*

Proof. In order to produce a contradiction, assume that the dispersive Lamb model is stable as $m \to \infty$. By applying Lemma 3.4 to the reduced dynamics (3.35), we have

$$M\ddot{\mathbf{q}} + \left(S + \frac{2T}{c}I\right)\dot{\mathbf{q}} + \left(V + \frac{Tm}{c}I\right)\mathbf{q} = O\left(\frac{1}{\sqrt{m}}\right). \tag{3.38}$$

Hence, the associated dynamical system (3.37) must also have a stable zero equilibrium. This contradicts the assumption that the system (3.37) is unstable. Our assumption that the dispersive Lamb model is stable as $m \to \infty$ must be false. This completes the proof. ∎

For large values of m, the asymptotic analysis suggests that a gyroscopically stabilized Chetaev system will maintain its stability after dispersive Lamb coupling. The dispersive nature of the wave equation restricts the access of the oscillator from interacting with low frequencies of the wave. This restriction effectively shifts the potential of the oscillator. The potential shift can even stabilize previously unstable oscillators.

The evolution of a Chetaev subsystem mirrors that of stable Chetaev system with dissipation. In the limit as m tends to zero, the asymptotic behavior of the system retains the explicit damping contribution as in the nondispersive coupling.

4. Nonlocal field coupling to gyroscopic Chetaev systems. While the Lamb model imposes the boundary constraint that the oscillator is physically attached to the wave field, we now investigate the effect of an ambient field on a Chetaev system without such a constraint. Energy is transmitted between the ambient field and the Chetaev system not through boundary constraints, but rather through coupling in the Hamiltonian. We analyze the contribution of a nonlocal distribution on the wave field acting as a force on the oscillator. Mathematically the field acts through a coupling distribution which for simplicity we take for computations to be a Dirac-δ function.

4.1. Nondispersive, nonlocal wave field coupling to a Chetaev system. Here we investigate a field coupling of the mechanical gyroscopic system to the wave equation. As before, we have the standard nondispersive wave equation

$$\frac{\partial^2 w}{\partial t^2} - c^2 \frac{\partial^2 w}{\partial z^2} = 0. \tag{4.1}$$

We model force of the wave field on the Chetaev system by the magnitude of the wave field. In the spirit of conservation, the coupled system remains Hamiltonian, while each perturbed subsystem may not. Coupling of this type is important in various physical models; see [28], [27], [29] and references therein. The interaction between the wave equation and the Chetaev system is modeled with a coupling parameter κ and with a coupling distribution $\chi(\xi)$; the equations of motion are

$$\begin{aligned} M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} &= \kappa \int_{\mathbb{R}} \chi(z) w(z,t) dz \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \\ \frac{\partial^2 w}{\partial t^2} - c^2 \frac{\partial^2 w}{\partial z^2} &= \kappa \chi(z) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}^T \mathbf{q}. \end{aligned} \tag{4.2}$$

For simplicity, we choose $\chi(z)$ to be the Dirac-δ distribution. (So in fact, even though we will continue to use the terminology "nonlocal coupling" for the type of system discussed in this section, the special case we consider here is also a type of focused, local with respect to the ambient field, coupling without boundary constraints.)

The above coupled system remains Hamiltonian with H defined as

$$H = \frac{1}{2}\left(\int_{\mathbb{R}} w_t^2 + c^2 w_z^2 dz + \dot{\mathbf{q}}^T M \dot{\mathbf{q}} + \mathbf{q}^T V \mathbf{q} \right) - \kappa \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}^T \mathbf{q} \int_{\mathbb{R}} \chi(z) w(z,t) dz. \tag{4.3}$$

Notice that the Hamiltonian is indefinite; hence the energy of the system does not uniformly bound the motion of the Chetaev system nor the motion of the wave field. Furthermore, the magnetic terms are present in the mechanical symplectic form defined by the Chetaev symplectic form, Ω, defined in (2.3).

4.1.1. Reduction of nondispersive system to a finite dimensional system. Even though the Chetaev system and the wave equation seem intimately coupled, we can decouple the wave motion from that of the Chetaev system as follows.

Taking the Fourier transform with respect to z in the wave equation yields

$$\begin{aligned}
&\hat{w}(k,t) = \int_{\mathbb{R}} e^{-ikz} w(z)dz,\\
&\hat{w}_{tt}(k,t) + c^2k^2\hat{w}(k,t) = \kappa \left[1 \dots 1\right] \mathbf{q}(t),\\
&\hat{w}(k,t) = \kappa \int_0^t \frac{\sin(ck(t-s))}{ck} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}^T \mathbf{q}(s)ds + \hat{w}_{free}(k,t).
\end{aligned} \tag{4.4}$$

The above equation is the retarded Green's function for the wave equation, and w_{free} is the homogeneous solution satisfying initial conditions. Computing $w(0,t)$, we have

$$\begin{aligned}
w(0,t) &= \frac{1}{2\pi}\int_{\mathbb{R}} \hat{w}(k,t)dk\\
&= w_{free}(0,t) + \frac{\kappa}{4\pi c}\int_0^t \int_{\mathbb{R}} \frac{2\sin(ck(t-s))}{ck} d(ck) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}^T \mathbf{q}(s)ds\\
&= w_{free}(0,t) + \frac{\kappa}{2c}\int_0^t \operatorname{sgn}(t-s) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}^T \mathbf{q}(s)ds\\
&= w_{free}(0,t) + \frac{\kappa}{2c}\int_0^t \left[1 \dots 1\right] \mathbf{q}(s)ds.
\end{aligned} \tag{4.5}$$

Thus the Chetaev system reduces to

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} = \kappa w_{free}(0,t)\left[1 \dots 1\right]^T + \frac{\kappa^2}{2c} Ones(n) \int_0^t \mathbf{q}(s)ds, \tag{4.6}$$

where $Ones(n)$ is an $n \times n$ matrix with each matrix element equal to one.

4.1.2. Differentiated system. We can interpret (4.6) as an integral feedback on the Chetaev system. To analyze the stability, we investigate the stability of the differentiated system.

Suppose that initially the coupling field is identically zero, and hence $w_{free}(0,t) = 0$ for all time $t > 0$. Since our system decouples, we consider the related system of ordinary differential equations. Let us write the system as a first order system with the variable Q:

$$Q = \begin{bmatrix} \mathbf{q}^T & \dot{\mathbf{q}}^T & \ddot{\mathbf{q}}^T \end{bmatrix}^T. \tag{4.7}$$

Differentiating the reduced system yields

$$\dot{Q} = AQ, \tag{4.8}$$

where

$$A = \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & I \\ \frac{\kappa^2}{2c}M^{-1}Ones(n) & -M^{-1}V & -M^{-1}S \end{bmatrix}. \tag{4.9}$$

PROPOSITION 4.1. *If $\kappa \neq 0$, then the system* (4.8) *is unstable.*

Proof. Since $\mathrm{Trace}(M^{-1}S) = -\mathrm{Trace}(SM^{-1}) = -\mathrm{Trace}(M^{-1}S)$, $M^{-1}S$ is traceless. For nonzero κ, we compute the rank of A as $\mathrm{rank}(A) = 2n+1$. The matrix A has an odd number $(2n+1)$ of nonzero eigenvalues which sum to zero. Since the nonzero eigenvalues satisfy an odd degree polynomial with real coefficients, A has at least one nonzero real eigenvalue. Hence, A has an eigenvalue with positive real part; thus the system is unstable. ∎

Example. Assume $n = 2$, $M = I$, $S = \left[\begin{smallmatrix} 0 & -B \\ B & 0 \end{smallmatrix}\right]$, and $V = \left[\begin{smallmatrix} \alpha & 0 \\ 0 & \beta \end{smallmatrix}\right]$. Computing the characteristic equation of the matrix A, we have

$$p(\mu) = \mu\left(\mu^5 + (B^2+\alpha+\beta)\mu^3 - 2\frac{\kappa^2}{2c}\mu^2 + \alpha\beta\mu - \frac{\kappa^2}{2c}(\alpha+\beta)\right). \tag{4.10}$$

Let $\mu_1, \ldots, \mu_5$ be the nonzero eigenvalues of A. Then we have

$$\prod_{i=1}^{5} \mu_i = \frac{\kappa^2}{2c}(\alpha+\beta). \tag{4.11}$$

Note the system is gyroscopically stabilized when $B^2 + \alpha + \beta \geq 2\sqrt{\alpha\beta}$. Also, $\mathrm{Trace}(A) = 0$. Hence if $\gamma(\alpha+\beta) \neq 0$, then A has an eigenvalue with positive real part. In particular, if $\alpha, \beta < 0$ and $\kappa \neq 0$, then there are real negative eigenvalues. Thus, there exists a conjugate pair of eigenvalues with positive real part. Hence the differentiated system is unstable.

We can quantify the rate at which the coupled system becomes unstable. If we additionally have $\alpha \neq \beta$, we use first order perturbation theory to compute the speed at which the eigenvalues leave the imaginary axis,

$$\mu' = \frac{d\mu}{d\gamma}\Big|_{\gamma=0}. \tag{4.12}$$

Computing μ', we have

$$\mu' = \frac{2\mu^2 + (\alpha+\beta)}{5\mu^4 + 3(w_0^2+w_1^2)\mu^2 + w_0^2 w_1^2}, \tag{4.13}$$

where the eigenvalues of the unperturbed system are $\pm iw_0, \pm iw_1$. Simplifying in terms of w_0 and w_1, we have

$$(\pm iw_j)' = \frac{-2w_j^2 + (\alpha+\beta)}{2w_j^2(w_j^2 - w_{1-j}^2)}. \tag{4.14}$$

We compare μ' with the μ' from a purely dissipative perturbation (see the analogous calculation in [19], [5] and references therein); we have

$$(\pm iw_j)' = \frac{(\pm iw_j)'_{dissipative}}{w_j^2}. \tag{4.15}$$

Since we use the differentiated systems, we require the existence of higher derivatives. In particular, a C^3 solution to the reduced system (4.6) is a solution to the differentiated system (4.8). Conversely, a solution to the differentiated system (4.8) is a solution to the reduced system (4.6) if initial conditions are satisfied. Satisfying

the initial conditions imposes a mild linear constraint on the system that is easily satisfied. We summarize the result with a theorem.

THEOREM 4.2. *The Chetaev system coupled to the standard, nondispersive wave equation via nonlocal coupling* (4.2)*, where* $\chi(z)$ *is the Dirac-*δ *distribution, yields the reduced finite dimensional system* (4.6)*. All* C^2 *solutions of this system are unstable.*

Proof. Let us write the solution to the differentiated problem as

$$\mathbf{q}(t) = \sum_i \lambda_i \mathbf{A}_i e^{\mu_i t}, \tag{4.16}$$

where $\mu_i, \mathbf{A_i}$ are eigenvalues and position components of the eigenvectors for the linear system and λ_i is a scaling chosen to satisfy initial conditions. Evaluation of (4.6) at $t = 0$ with $w_{free} = 0$ yields

$$M\ddot{\mathbf{q}}(0) + S\dot{\mathbf{q}}(0) + V\mathbf{q}(0) = \mathbf{0}. \tag{4.17}$$

We have

$$\ddot{\mathbf{q}}(0) = \sum_i \mu_i^2 \mathbf{A}_i, \qquad \dot{\mathbf{q}}(0) = \sum_i \mu_i \mathbf{A}_i, \qquad \mathbf{q}(0) = \sum_i \mathbf{A}_i. \tag{4.18}$$

Using the fact that μ_i is an eigenvalue of the differentiated system, we have

$$(\mu_i^3 M + \mu_i^2 S + \mu_i V - \gamma Ones(n))\mathbf{A}_i = \mathbf{0}. \tag{4.19}$$

Substituting, the initial condition is satisfied if

$$\gamma Ones(n) \sum_{i,\mu_i \neq 0} \frac{\lambda_i}{\mu_i} \mathbf{A}_i = \mathbf{0}. \tag{4.20}$$

This is just a linear constraint on the initial conditions, which can be satisfied while exhibiting a negative eigendirection. ∎

This instability is unlike the effect of Rayleigh dissipation of the type in gyroscopic Lamb coupling (3.5), i.e., dissipation arising from a term consisting of a positive definite symmetric matrix multiplying velocities. For small coupling, Rayleigh dissipation induces instability only if the Chetaev system is gyroscopically stable (i.e., it is unstable for $B = 0$ but stable for suitably large B). In the nonlocal field coupling model, instability is always induced. (See Chetaev [8] and [5] for a discussion of Rayleigh dissipation induced instability).

Example 1. For $n = 2$, $\alpha = -1$, $\beta = -2$, $c = 1$, and $B = 3$, the origin of the uncoupled system is gyroscopically stabilized. Increasing the coupling parameter, we show that the weaker eigenvalues destabilize as κ increases from 0 to 1 in Figure 5.

Example 2. As discussed above, we have shown that nonlocal coupling has a broader destabilizing effect than dissipation. In particular, the integral coupling can destabilize a system regardless of whether or not it is gyroscopically stabilized, while Rayleigh dissipative perturbations destabilize a Chetaev system only if it is initially gyroscopically stabilized. For $\alpha = 1$, $\beta = 2$, $c = 1$, and $B = 3$, we show numerically how increasing the coupling parameter destabilizes the system; see Figure 6.

4.2. Dispersive, nonlocal wave field coupling to a Chetaev system. In this section, we demonstrate that the nature of the wave field, to which the Chetaev system is coupled, has a significant impact on the stability of the overall system. We

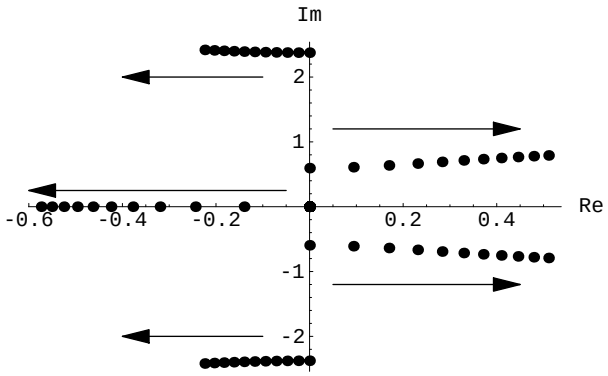


FIG. 5. *Destabilization of a gyroscopically stabilized system.*

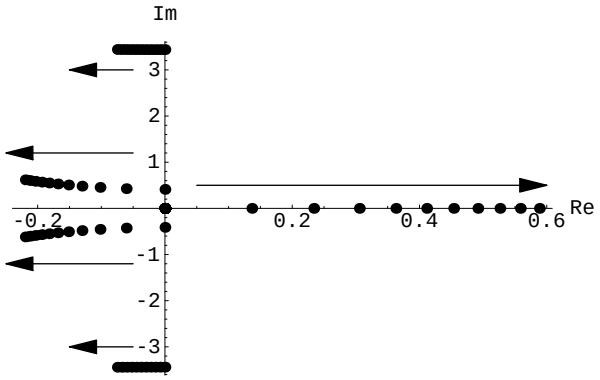


FIG. 6. *Destabilization of stabilized system ($\alpha = 1, \beta = 2, B = 3, c = 1$). κ increases from* 0 *to* 1. *A lone eigenvalue escapes from the origin, while the eigenvalues of the original system move into the left half plane.*

now consider the case of a gyroscopic system coupled to the linear Klein–Gordon equation, a standard dispersive wave equation,

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} = \kappa \int_{\mathbb{R}} \chi(z) w(z,t) d\xi \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \tag{4.21}$$

$$\frac{\partial^2 w}{\partial t^2} - c^2 \frac{\partial^2 w}{\partial z^2} + m^2 w = \kappa \chi(z) \begin{bmatrix} 1 \dots 1 \end{bmatrix} \mathbf{q}.$$

In this case, we show that a region of stability is possible for small coupling.

4.2.1. Energy bounds. For some potentials, we can explicitly show that the Hamiltonian is positive definite and hence can be used to bound the motion of the system. In the case that the Hamiltonian is indefinite, our system may or may not be stable. The main stability results of this section we present in the following theorem.

THEOREM 4.3. *If $|\kappa| < \min(m^2, c^2)$, then the nonlocal dispersive wave coupling to a mechanical gyroscopic system* (4.21)*, with $\chi(z) = \delta(z)$, is stable if $V - |\kappa| Ones(n)$*

is positive definite, where Ones(n) is an $n \times n$ matrix with each entry equal to unity. Explicitly, we have a uniform bound of $\|w\|_{H^1}$, $\|w_t\|$, $|\mathbf{q}|$, and $|\dot{\mathbf{q}}|$ in terms of initial energy.

Proof of Theorem 4.3. We will use energy estimates to give a uniform bound in phase space of the gyroscopic system coupled to the Klein–Gordon equation.

We use the form of the Chetaev system with the magnetic terms in the symplectic form, Ω. In this case, the Hamiltonian of the coupled system becomes

$$H = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^T V\mathbf{q} + \frac{1}{2}\int_{\mathbb{R}} w_t^2 + c^2 w_z^2 + m^2 w^2 dz - \kappa w(0,t)\mathbf{q}_\Sigma, \tag{4.22}$$

where $\mathbf{q}_\Sigma = \begin{bmatrix}1 \dots 1\end{bmatrix}\mathbf{q}$. Using the Cauchy–Schwarz inequality, we have the following energy estimate:

$$\begin{aligned} H &\geq \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^T V\mathbf{q} \\ &\quad + \frac{1}{2}(\|w_t\|_{L_2}^2 + \min(c^2, m^2)\|w\|_{H^1}^2) - \frac{1}{2}|\kappa|(w(0,t)^2 + (\mathbf{q}_\Sigma)^2) \\ &= \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^T (V - |\kappa| Ones(n))\mathbf{q} \\ &\quad + \frac{1}{2}(\|w_t\|_{L_2}^2 + \min(c^2, m^2)\|w\|_{H^1}^2) - \frac{1}{2}|\kappa| w(0,t)^2, \end{aligned} \tag{4.23}$$

where $Ones(n)$ is an $n \times n$ matrix with all entries equal to unity. From a standard Sobolev inequality, we have the result that if $w \in C_0^1(\mathbb{R}^2)$, then $w(0,t)^2 \leq \|w\|_{H^1}^2$. Using the Sobolev inequality, we have another bound on the energy:

$$\begin{aligned} H &\geq \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^T \Big(V - |\kappa| Ones(n)\Big)\mathbf{q} \\ &\quad + \frac{1}{2}\Big(\|w_t\|_{L_2}^2 + (\min(c^2, m^2) - |\kappa|)\|w\|_{H^1}^2\Big). \end{aligned} \tag{4.24}$$

Our Hamiltonian is positive definite if $|\kappa| < \min(c^2, m^2)$ and the matrix $\tilde{H}$ is positive definite, where

$$\tilde{H} = \begin{bmatrix} V - |\kappa| Ones(n) & 0 \\ 0 & M^{-1} \end{bmatrix}. \tag{4.25}$$

Equivalently, H is positive definite if $|\kappa| < \min(c^2, m^2)$ and $V - |\kappa| Ones(n)$ is positive definite. ∎

By the same argument we obtain the following theorem.

THEOREM 4.4. *Suppose $|\kappa| < \frac{1}{n}\min(m^2, c^2)$, and let $\lambda_1, \dots, \lambda_n$ be eigenvalues of V. Let $\alpha = \min_{i=1,\dots,n}(|\lambda_i|)$. If $|\kappa| < \alpha$, then the nonlocal dispersive wave field coupling to the Chetaev system* (4.21)*, with $\chi(z) = \delta(z)$, is stable. Explicitly, we have a uniform bound of $\|w\|_{H^1}$, $\|w_t\|$, $|\mathbf{q}|$, and $|\dot{\mathbf{q}}|$ in terms of initial energy.*

Proof. Using the Cauchy–Schwarz inequality in each component of $\mathbf{q}$ in equation (4.23), we have

$$\begin{aligned} H &\geq \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^T (V - |\kappa| I)\mathbf{q} \\ &\quad + \frac{1}{2}\Big(\|w_t\|_{L_2}^2 + (\min(c^2, m^2) - n|\kappa|)\|w\|_{H^1}^2\Big). \end{aligned} \tag{4.26}$$

∎

4.2.2. Linear algebra and stability of Chetaev systems. By energy estimates, stability is guaranteed if either the matrix $V - |\kappa| Ones(n)$ or the matrix $V - |\kappa| I$ is positive definite. Dispersion enables purely stable Chetaev systems (stable if $S = 0$) to remain stable for small coupling parameters. In this subsection, we prove some spectral results for these perturbed matrices to see how large the perturbation can be and guarantee stability. Even though positive definiteness of the Hamiltonian is a sufficient condition for stability, it is not a necessary condition as seen in gyroscopically stabilized Chetaev systems.

PROPOSITION 4.5. *Let $\lambda_1, \ldots, \lambda_n$ be eigenvalues of a symmetric $n \times n$ matrix V. Let $\alpha = \min_{i=1,\ldots,n}(|\lambda_i|)$. If $|\kappa| < \frac{\alpha}{n}$, then $W = V - \kappa Ones(n)$ has the same signature as V.*

Proof. Since V is symmetric and $Ones(n)$ is symmetric, W is symmetric and hence has real eigenvalues. Since a change of signature requires a zero eigenvalue of the perturbed matrix, we equivalently show

$$\det\Big(V - \kappa Ones(n)\Big) \neq 0 \quad \forall \kappa \in \Big(0, \frac{\alpha}{n}\Big). \tag{4.27}$$

If V and $Ones(n)$ commute, we can find a change of basis which simultaneously diagonalizes both V and $Ones(n)$. Working in this basis, we have that

$$\begin{aligned} \det\Big(V - \kappa Ones(n)\Big) &= \det\Big(\mathrm{diag}(\lambda_1, \ldots, \lambda_n) - \kappa \mathrm{diag}(\sigma(0, \ldots, 0, n))\Big) \\ &\neq 0 \qquad \forall \kappa \in \Big(0, \frac{\alpha}{n}\Big), \end{aligned} \tag{4.28}$$

where σ is a permutation of n elements. The eigenvalues of W are the eigenvalues of V with the exception that one of the eigenvalues is moved to the left by κn units. Hence if $0 < \kappa < \frac{\alpha}{n}$, then V and W have the same signature, as desired.

In the general case, we define a function f of κ as follows:

$$f(\kappa) = \det\Big(V - \kappa Ones(n)\Big). \tag{4.29}$$

Changing bases so that $Ones(n)$ is diagonal, we have

$$f(\kappa) = \det(\tilde{V} - \kappa \mathrm{diag}(n, 0, 0, \ldots, 0)), \tag{4.30}$$

where $\tilde{V}$ does not depend on κ. Hence f is linear in κ.

We now find a bound on the slope of $f(\kappa)$.

$$\begin{aligned} f'(\kappa) = f'(0) &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}(f(\epsilon) - f(0)) = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\Big(\det\Big(V - \epsilon Ones(n)\Big) - \det(V)\Big) \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}\Big(f(0)\det\Big(I - \epsilon(V^{-1} Ones(n))\Big) - f(0)\Big) \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}\Big(f(0)\left(1 - \epsilon \mathrm{Trace}\Big(V^{-1} Ones(n)\Big) + O(\epsilon^2)\right) - f(0)\Big) \\ &= -\det(V)\mathrm{Trace}\Big(V^{-1} Ones(n)\Big). \end{aligned} \tag{4.31}$$

Let C be an orthogonal matrix that diagonalizes $Ones(n)$ to $D = \mathrm{diag}(n, 0, 0, \ldots, 0)$.

$$\begin{aligned} \mathrm{Trace}\Big(V^{-1} Ones(n)\Big) &= \mathrm{Trace}\Big(C V^{-1} C^{-1} C Ones(n) C^{-1}\Big) \\ &= \mathrm{Trace}\Big(C V^{-1} C^{-1} D\Big). \end{aligned} \tag{4.32}$$

A well-known result of [25] states that the diagonal elements of $CV^{-1}C^{-1}$ lie in the convex hull of the permutations of $\{\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}\}$. Minimizing over permutations yields

$$\left|\text{Trace}\Big(CV^{-1}C^{-1}D\Big)\right| \leq \max_{d=(d_1,\ldots,d_n)} \left(\frac{nd_1}{|\lambda_{\sigma(1)}|} + \sum_{i=2}^{n} \frac{d_2}{|\lambda_{\sigma(i)}|}\right), \tag{4.33}$$

where σ is a permutation of n elements and $d_i \geq 0$ with $\sum_{i=1}^{n} d_i = 1$. The right-hand side of the above equation is a linear function on a convex set and hence attains its maximum at an extreme point,

$$\left|\text{Trace}\Big(CV^{-1}C^{-1}D\Big)\right| \leq \max_{i\in\{1,\ldots,n\}} \Big(\frac{n}{|\lambda_i|}\Big) \leq \frac{n}{\alpha}. \tag{4.34}$$

Let $\beta = \max_{i=1,\ldots,n}(|\lambda_i|)$. Finally, we have a bound on the slope of $f(\kappa)$,

$$\frac{|f(0)|}{\beta/n} \leq |f'(\kappa)| \leq \frac{|f(0)|}{\alpha/n}. \tag{4.35}$$

Thus for $|\kappa| < \frac{\alpha}{n}$ we have $f(\kappa) \neq 0$; hence the signatures of V and of $V - \kappa Ones(n)$ are the same, as desired. ∎

COROLLARY 4.6. *Let $\lambda_1, \ldots, \lambda_n$ be eigenvalues of a symmetric, positive definite $n \times n$ matrix V. Suppose $\kappa > \max_{i=1,\ldots,n}(\lambda_i)$. Then $V - \kappa Ones(n)$ has a different signature from the signature of V. In particular, $V - \kappa Ones(n)$ has at least one negative eigenvalue.*

Proof. For $\kappa > \max_{i=1,\ldots,n}(\lambda_i)$, we have $\text{Trace}(V - \kappa Ones(n)) < 0$. A symmetric matrix with negative trace has at least one negative eigenvalue. ∎

COROLLARY 4.7. *Let $\lambda_1, \ldots, \lambda_n$ be eigenvalues of a symmetric $n \times n$ matrix V. Suppose $|\lambda_i| > 0$ for $i = 1, \ldots, n$. Let $\alpha = \min_{i=1,\ldots,n}(|\lambda_i|)$ and $\beta = \max_{i=1,\ldots,n}(|\lambda_i|)$. We have the following results:*

- *As $|\kappa|$ increases, the matrix $V - |\kappa|Ones(n)$ changes signature at most once.*
- *If $V - |\kappa|Ones(n)$ changes signature, it does so for $|\kappa| \leq \frac{\beta}{n}$.*
- *If $V - |\kappa|Ones(n)$ changes signature, then an odd number of eigenvalues change sign.*

Proof. Linearity of $f(\kappa) = \det(V - \frac{\kappa}{n}Ones(n))$ in κ proves the first and third claims. Bounds in the $f'(\kappa)$ prove the second claim. ∎

4.2.3. Reduction of dispersive system to a finite dimensional system. Analogously to the computation in section 4.1.1, we eliminate the dynamics of the wave field to obtain reduced dynamics of a perturbed Chetaev system. Taking the Fourier transform of the field equation, we have

$$\ddot{\hat{w}} + \omega^2(k)\hat{w} = \kappa\hat{\chi}\mathbf{q}_\Sigma(t), \tag{4.36}$$

where $\omega = \sqrt{c^2k^2 + m^2}$ and

$$\mathbf{q}_\Sigma(s) = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}^T \mathbf{q}(s).$$

Let $\hat{g}$ be the Green's function for the above equation,

$$\ddot{\hat{g}} + \omega^2\hat{g} = \delta(t), \qquad \hat{g} = H(t)\frac{\sin\omega t}{\omega}, \tag{4.37}$$

where $H(t)$ is the Heaviside function. Computing $w(0,t)$ from the Green's function, we have

$$\begin{aligned} w(0,t) &= w_{free}(0,t) + \frac{1}{2\pi}\int_{\mathbb{R}} \hat{w}\ dk \\ &= w_{free}(0,t) + \frac{1}{2\pi}\int_{\mathbb{R}} (\kappa\hat{\chi}\mathbf{q}_\Sigma) * \hat{g}\ dk \\ &= w_{free}(0,t) + \frac{\kappa}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} \hat{\chi}\mathbf{q}_\Sigma(s)\hat{g}(t-s)\ ds\ dk, \end{aligned} \tag{4.38}$$

where w_{free} is a homogeneous solution to the dispersive wave equation. Assuming that the coupling starts at $t = 0$, we have $\mathbf{q}_\Sigma(t) = 0$ for $t < 0$. We can interchange the order of integration,

$$w(0,t) = w_{free}(0,t) + \frac{\kappa}{2\pi}\int_0^t \int_{\mathbb{R}} \hat{\chi}\mathbf{q}_\Sigma(s)\frac{\sin\omega(t-s)}{\omega}\ dk\ ds. \tag{4.39}$$

Choosing χ to be the Dirac-δ distribution and zero initial conditions in the wave field, the reduced dynamics becomes

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} = \frac{\kappa^2}{2\pi}\int_0^t \int_{\mathbb{R}} \mathbf{q}_\Sigma(s)\frac{\sin\omega(t-s)}{\omega}\ dk\ ds \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{4.40}$$

Explicitly using the form of the Klein–Gordon coupling, we can integrate the interior integral,

$$\int_{\mathbb{R}} \frac{\sin\omega s}{\omega} dk = 2\int_m^\infty \frac{\sin\omega s}{c\sqrt{\omega^2 - m^2}} d\omega = \frac{\pi}{c} J_0(ms), \tag{4.41}$$

where J_0 is the Bessel function of the first kind. We obtain an integrodifferential equation for the reduced dynamics,

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} = \frac{\kappa^2}{2c} Ones(n) \int_0^t J_0(m(t-s))\mathbf{q}(s)\ ds. \tag{4.42}$$

From the reduced dynamics (4.42), we can investigate the stability.

4.3. Asymptotic analysis and numerical simulations. For small coupling, we have shown that strongly stable Chetaev systems remain stable for $m \neq 0$ by Theorems 4.3 and 4.4. However, the energy bound results yield no information for gyroscopically stabilized Chetaev systems. Using the compact reduced form (4.42), asymptotic analysis suggests that gyroscopic systems maintain their stability under dispersive wave field perturbations. Moreover, the asymptotic analysis also suggests that an unstable system may be stabilized for a band of coupling parameters. In this section, we develop asymptotic expansions for the reduced dynamics (4.42) and compare the results with numerical simulations.

For large values of m, the wave field's effect on the Chetaev system appears to be a convolution with an approximate identity, $mJ_0(mt)$. Since $J_0(t)$ is not in $L^1(\mathbb{R})$, showing the approximate identity nature of the convolution becomes nontrivial, but nonetheless true, which is summarized in a theorem.

THEOREM 4.8. *Consider the integrodifferential equation* (4.42) *with the nonresonant hypothesis* H1 *and the stability hypothesis* H2*:*

H1. $M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} = O(\kappa^2)$,

H2. $\mathbf{q}$ *and* $\dot{\mathbf{q}}$ *are bounded.*

As the dispersion parameter, m, tends to infinity, we have an asymptotic expansion for the dynamics,

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + \Big(V - \frac{\kappa^2}{2mc} Ones(n)\Big)\mathbf{q} = O(\kappa^2 m^{\frac{-3}{2}}). \tag{4.43}$$

Since the error of the asymptotic expansion is $O(\kappa^2 m^{\frac{-3}{2}})$, the expansion will remain valid even for large coupling parameter κ, provided m is sufficiently large. The hypotheses H1 and H2 guarantee the validity of the expansion, but numerical simulations suggest that the expansion is accurate even when the system becomes unstable. In either case, the stability of the dynamical system (4.43) is a necessary condition for the stability of the coupled Chetaev system. If we relax the H2 hypothesis, the asymptotic expansion holds over any finite time interval.

We also note that the asymptotic expansion is consistent with the results from energy bounds. In particular, we can use the results from section 4.2.2 to investigate the stability of the Chetaev system

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + \Big(V - \frac{\kappa^2}{2mc} Ones(n)\Big)\mathbf{q} = 0. \tag{4.44}$$

For small values of m, we retain the nondispersive wave field dynamics.

THEOREM 4.9. *An asymptotic expansion of the reduced wave field dynamics* (4.42)*, as m tends to zero, is given by*

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} = \frac{\kappa^2}{2c} Ones(n) \int_0^t \mathbf{q}(s)ds + O(m). \tag{4.45}$$

We now begin the proofs of the theorems.

Proof of Theorem 4.8. The strategy used in this proof is to reduce the coupling term to a convolution with an approximate identity though integration by parts and substitution of H1. To simplify notation, we define the following integrals for Bessel function parameters $\nu \in \{0, 1\}$ and $n \in \{0, 1, 2\}$ derivatives of $\mathbf{q}$:

$$I_{n,\nu} = \int_0^t \frac{J_\nu(m(t-s))}{(t-s)^\nu} \frac{d^n\ \mathbf{q}(s)}{ds^n} ds. \tag{4.46}$$

In the new notation, our reduced system becomes

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} = \frac{\kappa^2}{2c} Ones(n) I_{0,0}. \tag{4.47}$$

Integrating $I_{0,0}$ by parts and using the fact that $sJ_0(s) = d/ds(sJ_1(s))$, we have

$$\begin{aligned}
I_{0,0} &= \int_0^t J_0(m(t-s))\mathbf{q}(s)ds \\
&= \int_0^t J_0(m(t-s))(t-s)\frac{\mathbf{q}(s)}{t-s} ds \\
&= -\frac{1}{m} J_1(m(t-s))\mathbf{q}(s)\Big|_0^t + \frac{1}{m}\int_0^t J_1(m(t-s))\left(\dot{\mathbf{q}}(s) + \frac{\mathbf{q}(s)}{t-s}\right) ds \\
&= \frac{J_1(mt)}{m}\mathbf{q}(0) + \frac{1}{m} I_{0,1} + \frac{1}{m^2} J_0(m(t-s))\dot{\mathbf{q}}(s)\Big|_0^t - \frac{1}{m^2} I_{2,0} \\
&= \frac{J_1(mt)}{m}\mathbf{q}(0) + \frac{1}{m^2}(\dot{\mathbf{q}}(t) - J_0(mt)\dot{\mathbf{q}}(0)) + \frac{1}{m} I_{0,1} - \frac{1}{m^2} I_{2,0}.
\end{aligned} \tag{4.48}$$

Similarly to the dispersive Lamb model, we recognize $I_{n,1}$ as an approximate identity convolution with $\frac{d^n\mathbf{q}}{dt^n}$ as m tends to infinity.

With the H1 nonresonant hypothesis, we solve $I_{0,0}$ to order κ^2 in terms of approximate identity convolutions, $I_{n,1}$, and nonintegral terms. In particular, we have

$$\ddot{\mathbf{q}} = -M^{-1}S\dot{\mathbf{q}} - M^{-1}V\mathbf{q} + O(\kappa^2). \tag{4.49}$$

We also integrate by parts to simplify $I_{1,0}$:

$$\begin{aligned} I_{1,0} &= \int_0^t J_0(m(t-s))(t-s)\frac{\dot{\mathbf{q}}(s)}{t-s}ds \\ &= -\frac{1}{m}J_1(m(t-s))\dot{\mathbf{q}}(s)\Big|_0^t + \frac{1}{m}\int_0^t J_1(m(t-s))\Big(\ddot{\mathbf{q}}(s) + \frac{\dot{\mathbf{q}}(s)}{t-s}\Big)ds \\ &= \frac{J_1(mt)}{m}\dot{\mathbf{q}}(0) + \frac{1}{m}I_{1,1} + \frac{1}{m}\int_0^t J_1(m(t-s))\ddot{\mathbf{q}}(s)ds. \end{aligned} \tag{4.50}$$

Employing the H1 hypothesis, we have

$$\begin{aligned} I_{1,0} &= \frac{J_1(mt)}{m}\dot{\mathbf{q}}(0) + \frac{1}{m}I_{1,1} \\ &\quad - \frac{1}{m}\int_0^t J_1(m(t-s))\Big(M^{-1}S\dot{\mathbf{q}}(s) + M^{-1}V\mathbf{q}(s)\Big)ds + O(\kappa^2) \\ &= \frac{J_1(mt)}{m}\dot{\mathbf{q}}(0) + \frac{1}{m}I_{1,1} - \frac{1}{m^2}J_0(m(t-s))M^{-1}S\dot{\mathbf{q}}(s)\Big|_0^t + \frac{1}{m^2}M^{-1}SI_{2,0} \\ &\quad - \frac{1}{m}\int_0^t J_1(m(t-s))M^{-1}V\mathbf{q}(s)ds + O(\kappa^2) \\ &= \Big(\frac{J_1(mt)}{m} + \frac{J_0(mt)}{m^2}M^{-1}S\Big)\dot{\mathbf{q}}(0) - \frac{1}{m^2}M^{-1}S\dot{\mathbf{q}}(t) + \frac{1}{m}I_{1,1} + \frac{1}{m^2}M^{-1}SI_{2,0} \\ &\quad - \frac{1}{m}\int_0^t J_1(m(t-s))M^{-1}V\mathbf{q}(s)ds + O(\kappa^2) \\ &= \Big(\frac{J_1(mt)}{m} + \frac{J_0(mt)}{m^2}M^{-1}S\Big)\dot{\mathbf{q}}(0) - \frac{1}{m^2}M^{-1}S\dot{\mathbf{q}}(t) + \frac{1}{m}I_{1,1} + \frac{1}{m^2}M^{-1}SI_{2,0} \\ &\quad + \frac{1}{m^2}M^{-1}V\Big(J_0(mt)\mathbf{q}(0) - \mathbf{q}(t)\Big) + \frac{1}{m^2}M^{-1}VI_{1,0} + O(\kappa^2). \end{aligned} \tag{4.51}$$

Integrating the nonresonant hypothesis H1, we have

$$I_{2,0} = -M^{-1}SI_{1,0} - M^{-1}VI_{0,0} + O(\kappa^2). \tag{4.52}$$

Taking linear combinations of (4.48), (4.51), and (4.52), we have an approximation for $I_{0,0}$ in terms of approximate identity convolutions of the form $I_{n,1}$ and nonintegral terms. Using H2 and invoking the convolution Lemma 3.4, we have

$$\lim_{m\to\infty} I_{n,1} = \frac{d^n}{dt^n}\mathbf{q}(t). \tag{4.53}$$

In particular, the error in the limit is $O(m^{\frac{-1}{2}})$. Collecting highest order terms in m of the linear combination, we have

$$I_{0,0} = \frac{1}{m}I_{0,1} + O(m^{\frac{-3}{2}}) = \frac{1}{m}\mathbf{q}(t) + O(m^{\frac{-3}{2}}). \tag{4.54}$$

Hence we have an asymptotic expansion of the perturbed dynamics,

$$M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + \Big(V - \frac{\kappa^2}{2mc} Ones(n)\Big)\mathbf{q} = O(\kappa^2 m^{\frac{-3}{2}}), \tag{4.55}$$

as desired. ∎

Proof of Theorem 4.9. Since $\mathbf{q}$ analytically depends on the parameter, m, we have

$$\begin{aligned} M\ddot{\mathbf{q}} + S\dot{\mathbf{q}} + V\mathbf{q} &= \frac{\kappa^2}{2c} Ones(n) \int_0^t J_0(m(t-s))\mathbf{q}(s)ds\Big|_{m=0} + O(m) \\ &= \frac{\kappa^2}{2c} Ones(n) \int_0^t \mathbf{q}(s)ds + O(m) \end{aligned} \tag{4.56}$$

as m tends to zero. ∎

We have already shown the instability of the nondispersive wave field coupling, and the asymptotic analysis suggests that the nondispersive limit has a similar instability.

4.3.1. One dimensional example: Simple harmonic oscillator coupled to the Klein–Gordon equation. We illustrate the asymptotic behavior first in the case of a one dimensional Chetaev system coupled to a dispersive wave equation. Consider a simple harmonic oscillator coupled to the Klein–Gordon equation with the following Hamiltonian system:

$$\begin{aligned} H &= H_0 + \kappa^2 H_1, \\ H_0 &= \frac{1}{2}(\dot{x}^2 + \alpha x^2) + \frac{1}{2}\int_{\mathbb{R}} w_t^2 + c^2 w_z^2 + m^2 q^2 dz, \\ H_1 &= -\frac{1}{\kappa} w(0,t)x. \end{aligned} \tag{4.57}$$

The Hamiltonian equations of motion are

$$\begin{aligned} \ddot{x} + \alpha^2 x &= \kappa w(0,t), \\ \ddot{w} - c^2 \frac{\partial^2 w}{\partial z^2} + m^2 w &= \kappa \delta(z) x. \end{aligned} \tag{4.58}$$

We can reduce the coupled system to an integral equation for a perturbed oscillator by solving for $w(0,t)$:

$$\begin{aligned} w(0,t) &= \frac{1}{2\pi}\int_{\mathbb{R}} \hat{w}(k,t)dk \\ &= \frac{\kappa}{2\pi}\int_0^t x(s)\int_{\mathbb{R}} \frac{\sin(\omega(t-s))}{\omega} dk ds \\ &= \frac{\kappa}{2c}\int_0^t J_0(m(t-s))x(s)ds, \end{aligned} \tag{4.59}$$

where $\omega = \sqrt{c^2k^2 + m^2}$ and J_0 is the Bessel function of the first kind. The perturbed oscillator remains Hamiltonian with the following dynamics:

$$\ddot{x} + \alpha^2 x = \frac{\kappa^2}{2c}\int_0^t J_0(m(t-s))x(s)ds. \tag{4.60}$$

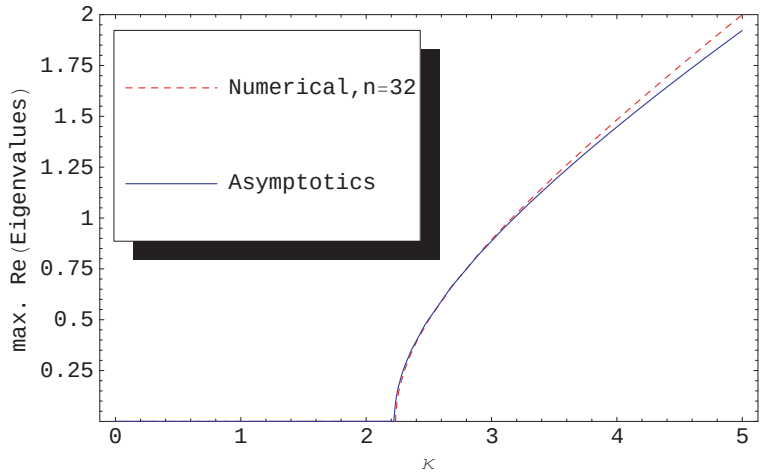


FIG. 7. *Bifurcation of eigenvalues. One dimensional dispersive wave field coupling:* $\Omega = 1$, $n = 32$, $m = 5$, $c = 1$, *and* $\alpha = 1$.

For the large m limit, we have

$$\ddot{x} + \left(\alpha^2 - \frac{\kappa^2}{2mc}\right) x \simeq 0. \tag{4.61}$$

Figure 7 shows a bifurcation diagram of the asymptotic results compared to a numerical solution to the full system. The maximum real part of the spectrum is plotted as a function of κ. The full coupled system is discretized with lattice size Ω and n lattice points for the wave system.

4.3.2. Two dimensional gyroscopic system coupled to a dispersive wave equation. The interesting gyroscopic behavior of a Chetaev system is captured in the two dimensional model. We see from section 4.2.2 the large m coupling can effectively move an eigenvalue of V to the left. The continuous nature of the movement allows for bands of stability for small coupling parameter κ and instability for larger coupling. We demonstrate, with numerical results, that stability is maintained for small coupling parameter. See Figures 8 and 9.

The effective movement of an eigenvalue of V allows for an interesting gyroscopic stabilization of an unstable Chetaev system for an interval of positive coupling parameters κ. We realize this dispersive wave field stabilization numerically and asymptotically in Figure 10.

5. The nonabelian setting. In this section we present an example of wave field coupling to a more complicated mechanical system—a rigid body with rotors. The configuration space is this case is the cross product of a nonabelian group, $SO(3)$, and two copies of the circle. This system models a controlled satellite. To analyze this system we need a little more geometry.

5.1. Normal form of principle bundles. Here we present a geometric development of the Chetaev system as a normal form of a group action on a mechanical system. For a more complete development of the normal forms and reduction theory of nonabelian group actions, see [5] and [26], from which we highlight some results.

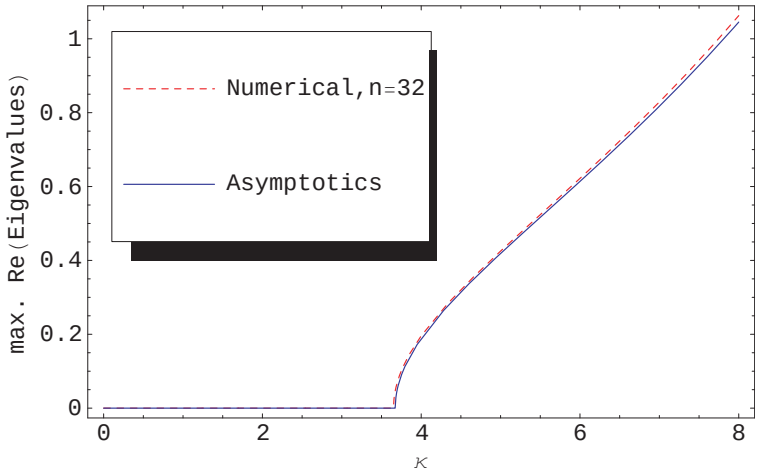


FIG. 8. *Bifurcation of eigenvalues. Two dimensional dispersive wave field coupling to stable Chetaev system:* $\Omega = 1$, $n = 32$, $m = 10$, $c = 1$, $\alpha = 1$, $\beta = 2$, *and* $B = 3$.

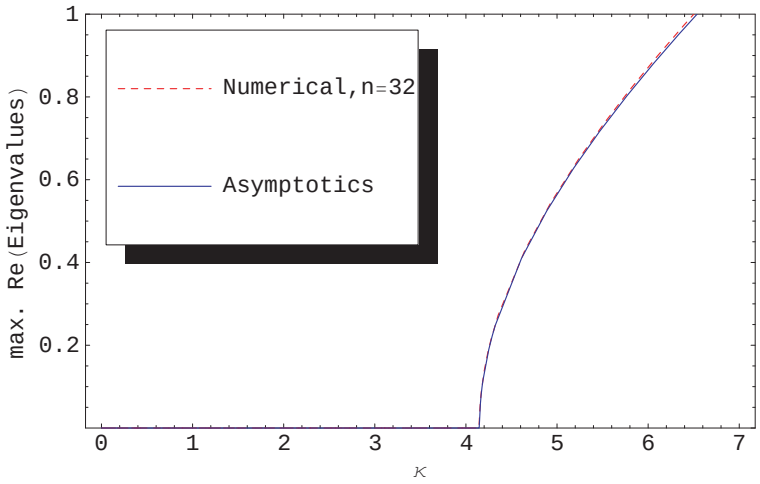


FIG. 9. *Bifurcation of eigenvalues. Two dimensional dispersive wave field coupling to gyroscopically stabilized Chetaev system:* $\Omega = 1$, $n = 32$, $m = 10$, $c = 1$, $\alpha = -1$, $\beta = -2$, *and* $B = 3$.

We show that the Chetaev system is the normal form of the linearized motion about a relative equilibrium of an abelian group action.

We consider the configuration space, Q, to be a Riemannian manifold with metric $\langle\langle\ ,\ \rangle\rangle$. Let G be a Lie group which acts freely on Q by isometries. By lifting the group action to the tangent bundle, TQ, or to the cotangent bundle, T^*Q, we have that G acts symplectically. For a mechanical system, the Lagrangian is of the form

$$L(q,v) = \frac{1}{2}\|v\|_q^2 - V(q), \tag{5.1}$$

and the Hamiltonian is of the form

$$H(q,p) = \frac{1}{2}\|p\|_q^2 + V(q), \tag{5.2}$$

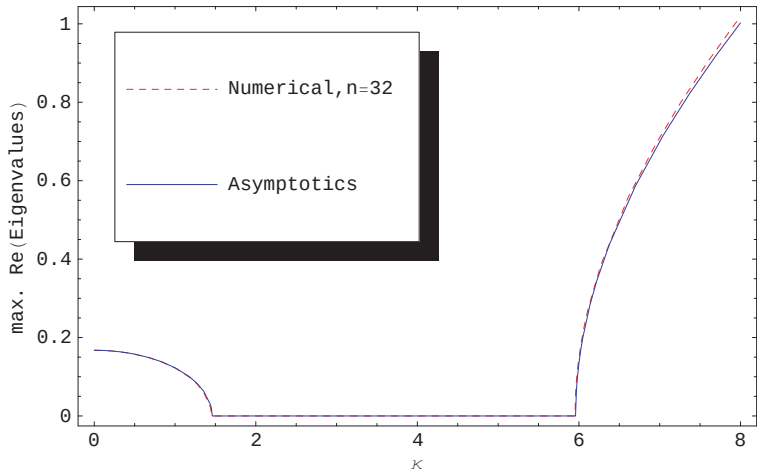


FIG. 10. *Bifurcation of eigenvalues. Two dimensional dispersive wave field coupling to unstable Chetaev system:* $\Omega = 1$, $n = 32$, $m = 10$, $c = 1$, $\alpha = \frac{1}{10}$, $\beta = -2$, *and* $B = 3$.

where $\|\cdot\|_q$ is the norm on T_qQ or the norm induced on T_q^*Q, and where V is a G-invariant potential. Since G acts freely on Q, the projection $\pi : Q \to Q/G$ defines a principle G-bundle.

Recall that the infinitesimal generator of $\xi \in \mathfrak{g}$ on Q is denoted ξ_Q and defined

$$\xi_Q(q) = \frac{d}{dt}\Big|_{t=0} \exp(\xi t) \cdot q. \tag{5.3}$$

For each $q \in Q$, we can define the *locked inertia tensor* to be the map $\mathbb{I} : \mathfrak{g} \to \mathfrak{g}^*$ defined by $\langle \mathbb{I}\eta, \xi \rangle = \langle\langle \eta_Q(q), \xi_Q(q) \rangle\rangle$. For coupled rigid bodies, $\mathbb{I}(q)$ is the classical moment of inertia tensor. We can now define a mechanical connection A.

DEFINITION 5.1. *We define the* mechanical connection *on the principle bundle* $Q \to Q/G$ *to be the map* $A : TQ \to \mathfrak{g}$ *given by* $A(q, v) = \mathbb{I}(q)^{-1}(\mathbf{J}(q, v))$, *where* $\mathbf{J} : TQ \to \mathfrak{g}^*$ *is the* momentum map *defined by* $\langle \mathbf{J}(q, v), \xi \rangle = \langle\langle v, \xi_Q(q) \rangle\rangle$. *In other words,* A *is the map that assigns to each* (q, v) *the corresponding angular velocity of the locked system.*

For $\mu \in \mathfrak{g}^*$, we let G_μ be the isotropy subgroup for the co-adjoint action of G on $\mathfrak{g}^*$,

$$G_\mu = \{g \in G \mid Ad^*_{g^{-1}}\mu = \mu\}. \tag{5.4}$$

The mechanical connection provides a natural decomposition into a horizontal space, hor_q, and a vertical space, ver_q,

$$\text{hor}_q = \{(q, v) \mid \mathbf{J}(q, v) = 0\} \subset T_qQ \tag{5.5}$$

$$\text{ver}_q = \{\xi_Q(q) \mid \xi \in \mathfrak{g}\} = \mathfrak{g} \cdot q \subset T_qQ. \tag{5.6}$$

The horizontal space is the space horizontal to G-orbits. The vertical space is the kernel of the projection map π.

We obtain the Chetaev system from the reduced dynamics of a Hamiltonian system about a *relative equilibrium*, $z_e = (q_e, v_e)$, defined by the Hamiltonian vector field at z_e pointing in the direction of the group orbit through z_e,

$$X_H(z_e) \in T_{z_e}(G \cdot z_e). \tag{5.7}$$

From the relative equilibrium theorem, there exists a $\xi \in \mathfrak{g}$ such that z_e is a critical point of the augmented Hamiltonian, H_ξ,

$$H_\xi(z) = H(z) - \langle \mathbf{J}(z) - \mathbf{J}(z_e), \xi \rangle. \tag{5.8}$$

To investigate the stability of the relative equilibrium modulo $G_{\mathbf{J}(z_e)}$, we analyze the second variation of $\delta^2 H_\xi(z_e)$ (the first variation vanishes at z_e) and the corresponding symplectic form Ω. Assuming that z_e is a regular point and $\mathbf{J}(z_e)$ is a generic point of $\mathfrak{g}^*$, we can define a rigid-internal splitting such that $\delta^2 H_\xi(z_e)$ block diagonalizes. The rigid component has a co-adjoint orbit representation, while the internal component, or shape-space, consists of $T^*(Q/G)$.

The rigid component is a subset of the vertical space, ver_{z_e}, orthogonal to the isotropy Lie algebra, $\mathfrak{g}_\mu$, for the momentum of the relative equilibrium $\mu = \mathbf{J}(z_e)$. Explicitly, we have

$$\mathcal{V}_{RIG} = \{\xi_Q(z_e) \in T_{z_e}Q \mid \langle \mathbb{I}\xi, \eta \rangle = 0 \ \forall \eta \in \mathfrak{g}_{\mathbf{J}(z_e)}\}. \tag{5.9}$$

The internal coordinates or shape-space represents the part of the system which is unaffected by the group action, Q/M. Due to the linearization of the system about the relative equilibrium, the internal variables live in $T_{[z_e]}Q/G$:

$$\mathcal{V}_{INT} = \{\delta q \in T_{z_e}Q \mid [\delta q] \in T_{[z_e]}Q/G\}. \tag{5.10}$$

In this splitting ($(r, q, p) \in \mathcal{V}_{RIG} \times \mathcal{V}_{INT} \times \mathcal{V}^*_{INT}$; see [26], [5] and references therein) $\delta^2 H_\xi$ block diagonalizes

$$\delta^2 H_\xi = \begin{bmatrix} A_\mu & 0 & 0 \\ 0 & V & 0 \\ 0 & 0 & M^{-1} \end{bmatrix}, \tag{5.11}$$

with symplectic form, Ω,

$$\Omega = \begin{bmatrix} L_\mu & C & 0 \\ -C^T & S & I \\ 0 & -I & 0 \end{bmatrix}, \tag{5.12}$$

where A_μ is the energy from the co-adjoint orbit block and V is the amended potential.

The linearized Hamiltonian vector field is defined from $\Omega(X_H, \cdot) = dH\cdot$,

$$X_H(r, q, p) = \begin{bmatrix} -L_\mu^{-1}A_\mu & 0 & -L_\mu^{-1}CM^{-1} \\ 0 & 0 & M^{-1} \\ -C^T L_\mu^{-1} A_\mu & -V & -\tilde{S}M^{-1} \end{bmatrix} \begin{bmatrix} r \\ q \\ p \end{bmatrix}, \tag{5.13}$$

where $S = \tilde{S} - C^T L_\mu C$.

5.1.1. Nonabelian example: Rigid body with two internal rotors. We consider a rigid body with two symmetric internal rotors; see Figure 11. For more detail on the stability of the rigid body, see [6] and [5]. The configuration space is $SO(3) \times S^1 \times S^1$, where each of the rotors has a configuration space of S^1. The Lagrangian is a function on the tangent bundle, $T(SO(3) \times S^1 \times S^1)$. Let $\mathbb{I}_{body}$ be the inertia tensor of the rigid body, $\mathbb{I}_{rotor} = \mathrm{diag}(J_1^1, J_2^2, 0)$ the diagonal matrix of rotor inertia about the principle axes, and $\mathbb{I}'_{rotor}$ be the remaining rotor inertia about the other axes. We define the locked inertia tensor, $\mathbb{I}_{lock}$, of the full system as

$$\mathbb{I}_{lock} = \mathbb{I}_{body} + \mathbb{I}_{rotor} + \mathbb{I}'_{rotor} = \mathrm{diag}(B_1, B_2, B_3). \tag{5.14}$$

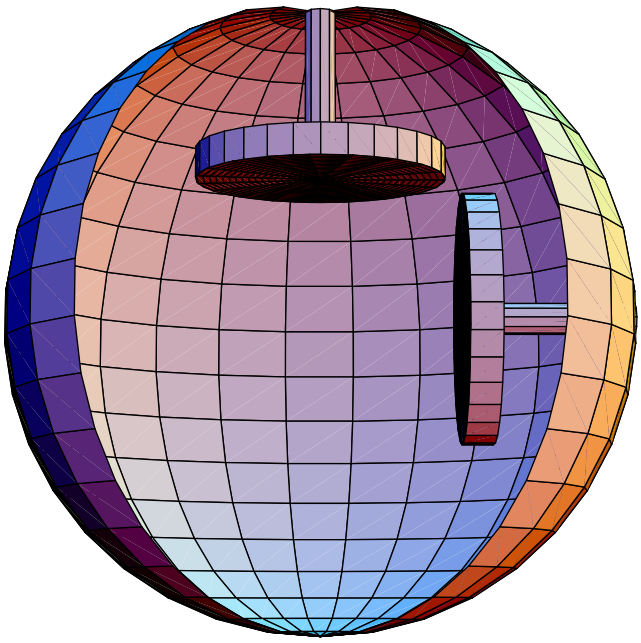


FIG. 11. *Rigid body with two internal rotors.*

Let $\mathbb{I}_{lock} - \mathbb{I}_{rotor} = \mathrm{diag}(A_1, A_2, A_3)$. Assume $B_1 > B_2 > B_3$.

As a simple mechanical system, the Lagrangian is the kinetic energy of the system, i.e., the total kinetic energy of the body plus the total kinetic energy of the rotor.

(5.15)
$$\begin{aligned} L((R,\theta_r),(\Omega,\Omega_r)) &= \frac{1}{2}\Omega\cdot\mathbb{I}_{body}\Omega + \frac{1}{2}\Omega\cdot\mathbb{I}'_{rotor}\Omega + \frac{1}{2}(\Omega+\Omega_r)\cdot\mathbb{I}_{rotor}(\Omega+\Omega_r) \\ &= \frac{1}{2}\Omega\cdot(\mathbb{I}_{lock}-\mathbb{I}_{rotor})\Omega + \frac{1}{2}(\Omega+\Omega_r)\cdot\mathbb{I}_{rotor}(\Omega+\Omega_r), \end{aligned}$$

where $R \in SO(3)$ is the attitude of the rigid body relative to an inertial frame, $\theta_r \in S^1 \times S^1 \times \{0\}$ is the angular configuration of the rotors, $\Omega \in \mathbb{R}^3 \cong \mathfrak{so}(3)$ is the vector of body angular velocities, and $\Omega_r \in \mathbb{R}^2 \times \{0\} \cong T_{\theta_r}(S^1 \times S^1) \times \{0\}$ is the vector of rotor angular velocities about the principle axes with respect to a body fixed frame.

The Euler–Lagrange equations of motion are

(5.16)
$$\begin{aligned} (\mathbb{I}_{lock}-\mathbb{I}_{rotor})\dot{\Omega} &= (\mathbb{I}_{lock}\Omega + \mathbb{I}_{rotor}\Omega_r)\times\Omega, \\ (\mathbb{I}_{lock}-\mathbb{I}_{rotor})\dot{\Omega}_r &= \Omega\times(\mathbb{I}_{lock}\Omega + \mathbb{I}_{rotor}\Omega_r), \\ \dot{R} &= R\Omega, \\ \dot{\theta}_r &= \Omega_r. \end{aligned}$$

We consider the normal form of the linearized equations of motion about the relative equilibrium $z_e = (q_e, v_e)$ with $q_e = (R, 0, 0)$ and $v_e = (\Omega^e, \Omega^e_r)$ with $\Omega^e = (0, 0, \omega)^T$

and $\Omega_r^e = (0,0,0)^T$. The relative equilibrium, z_e, is a member of an equivalence class of TQ/G, where the action of $G = SO(3)$ can be lifted from acting on Q by rotations of the rigid body to acting on the tangent bundle isometrically. Notice that the associated Hamiltonian vector field X_H on T^*Q evaluated at z_e points in the direction of the group orbit through z_e,

$$X_H(z_e) = (R\Omega^e, 0, 0, 0) \in T_{z_e}(G \cdot z_e). \tag{5.17}$$

We will use the mechanical connection to split the configuration variables into rigid and internal components. After the reduction of the phase space, TQ/G, the rigid component of the splitting is the subset of the rigid tangent space, $\mathfrak{so}(3)$, orthogonal to the isotropy subgroup $G_\mu = S^1$. The internal component is the part of phase space which is not affected by the group action.

First, we define the rigid component of the principle bundle:

$$\begin{aligned}
\mathcal{V}_{RIG} &= \{\xi_Q(q_e) \in T_{q_e}Q \mid \langle \mathbb{I}(q)\xi, \eta\rangle = 0 \ \forall \eta \in \mathfrak{s}^1 \subset \mathfrak{so}(3)\} \\
&= \{\xi_Q(q_e) \in T_{q_e}Q \mid d\Omega_3 \cdot \xi_Q(q_e) = 0\} \\
&= \{((\delta\Omega_1, \delta\Omega_2, 0) \times R, 0, 0, 0) \in T_{q_e}Q \mid (\delta\Omega_1, \delta\Omega_2, 0) \in \mathfrak{so}(3) \cong \mathbb{R}^3\} \\
&= \{(\delta\Omega_1, \delta\Omega_2) \in \mathbb{R}^2\}.
\end{aligned} \tag{5.18}$$

We can realize $\mathfrak{s}^1$ as the Lie algebra to the isotropy subgroup for $\mathbf{J}(z_e) \in \mathfrak{so}(3)^*$ for the co-adjoint action of $SO(3)$ on $\mathfrak{so}(3)^*$. Thus, the rigid variables, $r = (\delta\Omega_1, \delta\Omega_2)$, are even dimensional and represent the co-adjoint orbit block.

In this case, the internal component is the $T(\mathbb{T}^2)$. We can choose the natural configuration for the shape-space, $q = M^{-1}\delta\theta_r$, with associated reduced momenta, $p = \delta\Omega_r$. From the linearized equations of motions, we have

$$\frac{d}{dt}\begin{bmatrix} r \\ q \\ p \end{bmatrix} = \begin{bmatrix} -L_\mu^{-1}A_\mu & 0 & -L_\mu^{-1}CM^{-1} \\ 0 & 0 & M^{-1} \\ -C^T L_\mu^{-1}A_\mu & -V & -\tilde{S}M^{-1} \end{bmatrix} \begin{bmatrix} r \\ q \\ p \end{bmatrix}, \tag{5.19}$$

where

$$\begin{aligned}
L_\mu &= \begin{bmatrix} 0 & -\frac{1}{\omega} \\ \frac{1}{\omega} & 0 \end{bmatrix}, \quad A_\mu = \begin{bmatrix} \frac{B_3 - B_1}{A_2} & 0 \\ 0 & \frac{B_3 - B_2}{A_1} \end{bmatrix}, \\
M^{-1} &= \begin{bmatrix} \frac{J_1^1}{A_2} & 0 \\ 0 & \frac{J_2^2}{A_1} \end{bmatrix}, \quad \tilde{S} = \begin{bmatrix} 0 & \omega \\ -\omega & 0 \end{bmatrix}, \\
C &= -I, \quad V = 0.
\end{aligned} \tag{5.20}$$

5.1.2. Abelian group action. If the group G is an abelian group, the co-adjoint orbit block becomes trivial. The isotropy subgroup, G_μ, of the co-adjoint action is all of G. Hence the linearized system about a relative equilibrium contains only the internal components,

$$H_{linear} = \begin{bmatrix} V & 0 \\ 0 & M^{-1} \end{bmatrix}, \tag{5.21}$$

with symplectic form

$$\Omega = \begin{bmatrix} S & I \\ -I & 0 \end{bmatrix}. \tag{5.22}$$

The corresponding second order equations are precisely the Chetaev system

$$M\ddot{q} + S\dot{q} + Vq = 0. \tag{5.23}$$

The abelian case is a classical result dating back to [24].

5.2. Nonabelian example of a wave field coupling: Rigid body with internal rotors. In the previous sections, we have coupled a wave equation to the Chetaev system and investigated the stability of the trivial equilibrium. In this section, we demonstrate a similar destabilizing effect for a wave field coupling to the relative equilibrium of a nonabelian group action. In particular, we couple the standard wave equation to a rigid body with internal rotors. We use the same notation as introduced in section 5.1.1. We model internal radiation, that is, the field coupling to the internal rotors:

$$\begin{aligned}
(\mathbb{I}_{lock} - \mathbb{I}_{rotor})\dot{\Omega} &= (\mathbb{I}_{lock}\Omega + \mathbb{I}_{rotor}\Omega_r) \times \Omega, \\
(\mathbb{I}_{lock} - \mathbb{I}_{rotor})\dot{\Omega}_r &= \Omega \times (\mathbb{I}_{lock}\Omega + \mathbb{I}_{rotor}\Omega_r) + \kappa \int_{\mathbb{R}} \chi(\xi) u(\xi,t) d\xi \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \\
\dot{\theta}_r &= \Omega_r, \\
\ddot{u} - c^2 u_{\xi\xi} &= \kappa \chi(\xi) Ones(3) \theta_r.
\end{aligned} \tag{5.24}$$

Assuming the distribution χ is the Dirac-δ distribution, the wave equation decouples in precisely the same way as in the abelian case, yielding the following equations of motion for the rigid body:

$$\begin{aligned}
(\mathbb{I}_{lock} - \mathbb{I}_{rotor})\dot{\Omega} &= (\mathbb{I}_{lock}\Omega + \mathbb{I}_{rotor}\Omega_r) \times \Omega, \\
(\mathbb{I}_{lock} - \mathbb{I}_{rotor})\dot{\Omega}_r &= \Omega \times (\mathbb{I}_{lock}\Omega + \mathbb{I}_{rotor}\Omega_r) + \frac{\kappa^2}{2c} Ones(3) \int_0^t \Omega_r(s) ds, \\
\dot{\theta}_r &= \Omega_r,
\end{aligned} \tag{5.25}$$

ignoring the free wave homogeneous contribution.

We linearize the system about the relative equilibrium $z_e \in \mathfrak{so}(3) \times T(\mathbb{T}^2)$, where $z_e = ((0,0,\omega),(0,0),(0,0))$. We have the following integral equation in terms of $Q = \begin{bmatrix} r & q & p \end{bmatrix}^T$, with $r = \delta\Omega$, $q = M^{-1}\delta\theta_r$, and $p = \delta\Omega_r$:

$$\dot{Q} = \begin{bmatrix} -L_\mu^{-1} A_\mu & 0 & -L_\mu^{-1} C M^{-1} \\ 0 & 0 & M^{-1} \\ -C^T L_\mu^{-1} A_\mu & -V + \frac{\kappa^2}{2mc} (\mathbb{I}_{lock} - \mathbb{I}_{rotor})^{-1} Ones(3) \int_0^t M \cdot ds & -\tilde{S} M^{-1} \end{bmatrix} Q, \tag{5.26}$$

where

$$\begin{aligned}
L_\mu &= \begin{bmatrix} 0 & -\frac{1}{\omega} \\ \frac{1}{\omega} & 0 \end{bmatrix}, \quad A_\mu = \begin{bmatrix} \frac{B_3 - B_1}{A_2} & 0 \\ 0 & \frac{B_3 - B_2}{A_1} \end{bmatrix}, \\
M^{-1} &= \begin{bmatrix} \frac{J_1^1}{A_2} & 0 \\ 0 & \frac{J_2^2}{A_1} \end{bmatrix}, \quad \tilde{S} = \begin{bmatrix} 0 & \omega \\ -\omega & 0 \end{bmatrix}, \\
C &= -I, \quad V = 0.
\end{aligned} \tag{5.27}$$

Differentiating the system, we reduce the integrodifferential equation into a standard ordinary differential equation,

(5.28)
$$\frac{d}{dt}\begin{bmatrix}\dot{r}\\ q\\ \dot{q}\\ \dot{p}\end{bmatrix} = \begin{bmatrix} -L_\mu^{-1}A_\mu & 0 & 0 & -L_\mu^{-1}CM^{-1}\\ 0 & 0 & I & 0\\ 0 & 0 & 0 & M^{-1}\\ -C^TL_\mu^{-1}A_\mu & \frac{\kappa^2}{2mc}(\mathbb{I}_{lock}-\mathbb{I}_{rotor})^{-1}Ones(3)M & -V & -\tilde{S}M^{-1}\end{bmatrix}\begin{bmatrix}\dot{r}\\ q\\ \dot{q}\\ \dot{p}\end{bmatrix}.$$

Implicitly, we define the matrix W as

(5.29)
$$\frac{d}{dt}\begin{bmatrix}r\\ q\\ \dot{q}\\ \dot{p}\end{bmatrix} = W\begin{bmatrix}r\\ q\\ \dot{q}\\ \dot{p}\end{bmatrix}.$$

We can clearly see that the diagonal in W is zero; thus the sum of the eigenvalues is zero. To show instability we need only show that there exists an eigenvalue with nonzero real part. We can also see that the matrix W is rank 5 under the generic condition

$$\omega^2\kappa^2\Big(A_1(B_2-B_3)(B_3-B_2+J_1^1)+A+2(B_1-B_3)(B_3-B_2+J_2^2)\Big)\neq 0. \tag{5.30}$$

Generically, W has an odd number of nonzero eigenvalues. The matrix W has real entries, so the complex eigenvalues come in conjugate pairs. There must exist a real nonzero eigenvalue. Thus, our system becomes unstable.

We have presented only one example of the nonabelian coupling, but the instability is completely analogous to the abelian or Chetaev case. We plan to address the general instability of the nonabelian case in future research.

6. Conclusion. We have presented several mechanisms through which internal energy transfer may destabilize a system. One such mechanism models radiation damping, or energy transfer from a finite dimensional subsystem to an infinite dimensional subsystem.

We have generalized the Lamb model of an oscillator coupled to a wave equation to include coupling of Chetaev systems. For gyroscopic Lamb coupling, energy is transferred from the oscillator into the wave field, inducing instability. The radiation induced instability is analogous to dissipation induced instability studied in [5]. When we include dispersion in the gyroscopic Lamb model, there is effectively a stabilizing shift in the potential of the oscillator to balance the damping term.

Similarly to the gyroscopic Lamb model, the wave field coupling can exhibit radiation induced instability. In contrast to the Lamb model, the wave field coupling to an oscillator does not rely on boundary constraints as a mechanism for energy transfer. We investigate a local coupling whose interaction with an oscillator depends on the strength of the field at a point. While the gyroscopic Lamb model induces instability in gyroscopically stabilized oscillators, the wave field coupling induces instability in all oscillators. For large dispersion, the mechanical system has restricted access to low frequency wave modes, allowing a band of stable coupling parameters to exist. Furthermore, largely dispersive wave field coupling also shifts the potential to allow even unstable oscillators to be stabilized.

In addition to dispersive and nondispersive wave equations, future research plans include investigating the stability of the beam equation coupled to a rigid body. This is of particular interest in modeling antennae on satellites.

Another avenue of research includes coupling multiple oscillators through a wave equation or a beam equation. It might be possible to tune the system so as to trap energy between the oscillators. We can further extend the model to describe a lattice of nonlinear, coupled oscillators, such as the Toda lattice, via an infinite dimensional system.

Appendix A. Spherical pendulum. In this section, we derive the linearized equations of motion for the whirling spherical pendulum about the inverted equilibrium.

Consider a rigid spherical pendulum moving in a constant gravitational field. Let θ, ϕ, and r be the standard spherical coordinates for the pendulum.

We compute the equations of motion in a rotating frame. Since our pendulum is rigid, r is constant. The Lagrangian is

$$L(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2} m \|\dot{\mathbf{q}}\|^2 - U(\mathbf{q}) = \frac{\mathbf{1}}{\mathbf{2}} \mathbf{m}(\mathbf{r^2}\dot{\theta}^{\mathbf{2}} + \mathbf{r^2}\sin^{\mathbf{2}}\theta\dot{\phi}^{\mathbf{2}}) - \mathbf{mg}\cos\theta. \tag{A.1}$$

The Euler–Lagrange equation of motion are

$$mr^2\ddot{\theta} - mr^2\dot{\phi}^2 \sin\theta\cos\theta - mg\sin\theta = 0, \tag{A.2}$$

$$mr^2(\sin^2\theta\ddot{\phi} + 2\sin\theta\cos\theta\dot{\phi}\dot{\theta}) = 0. \tag{A.3}$$

Expanding the Lagrangian about $\theta = \pi$, the stable equilibrium, and keeping terms to second order in θ yields the following Lagrangian:

$$L(\theta, \phi, \dot{\theta}, \dot{\phi}) = \frac{1}{2} m(r^2\dot{\theta}^2 + r^2\theta^2\dot{\phi}^2) + mg\left(1 - \frac{1}{2}\theta^2\right), \tag{A.4}$$

with Euler–Lagrange equations of motion

$$mr^2\ddot{\theta} - mr^2\dot{\phi}^2\theta + mg\theta = 0, \tag{A.5}$$

$$mr^2(\theta^2\ddot{\phi} + 2\theta\dot{\phi}\dot{\theta}) = 0. \tag{A.6}$$

In a rotating coordinate system, we make the change of variables $x = r\theta\cos(\phi - \psi)$, $y = r\theta\sin(\phi - \psi)$, where ψ is the angle of rotation. Let the angular velocity $\omega = \dot{\psi}$ be constant. In the new coordinates the Lagrangian becomes

$$L(x, y, \dot{x}, \dot{y}) = \frac{1}{2} m((\dot{x} - \omega y)^2 + (\dot{y} + \omega x)^2) + mg\left(1 - \frac{x^2 + y^2}{2}\right). \tag{A.7}$$

Our linearized equations of motion are

$$\ddot{x} - 2\omega\dot{y} + (g - \omega^2)x = 0, \tag{A.8}$$

$$\ddot{y} + 2\omega\dot{x} + (g - \omega^2)y = 0. \tag{A.9}$$

We now physically motivate the rotating frame by modeling a rotor, with moment of inertia I, which rotates the pendulum. The Lagrangian is as follows:

$$L(\theta, \phi, \dot{\theta}, \dot{\phi}) = \frac{1}{2} m(r^2\dot{\theta}^2 + r^2\sin^2\theta\dot{\phi}^2) - mg\cos\theta + \frac{1}{2} I\omega^2. \tag{A.10}$$

The Euler–Lagrange equations of motion remain unchanged with an additional equation:

$$mr^2\ddot{\theta} - mr^2\dot{\phi}^2\sin\theta\cos\theta - mg\sin\theta = 0, \tag{A.11}$$

$$mr^2(\sin^2\theta\ddot{\phi} + 2\sin\theta\cos\theta\dot{\phi}\dot{\theta}) = 0, \tag{A.12}$$

$$\dot{\omega} = 0. \tag{A.13}$$

Without a coupling term, a rotating frame is indistinguishable from the inertial frame. The linearization remains identical.

We now wish to stabilize the inverted equilibrium by creating a restoring potential that breaks the symmetry of the motion of the pendulum. Imagine the pendulum with the pivot at the origin. Assume there is an asymmetric potential within the sphere. In particular, we have a potential V dependent on θ and $\phi-\psi$, where ψ is the angle of equatorial rotation of the spherical shell and ϕ is the standard angle of the pendulum.

Assuming that the rotor that is spinning the spherical shell is rotating with a constant angular velocity ω, we compute the Lagrangian

$$L(\theta,\phi,\dot\theta,\dot\phi)=\frac{1}{2}m(r^2\dot\theta^2+r^2\sin^2\theta\dot\phi^2)-mg\cos\theta-V(\theta,\phi-\omega t). \tag{A.14}$$

The Euler–Lagrange equations of motion are

$$\begin{aligned} mr^2\ddot\theta-mr^2\dot\phi^2\sin\theta\cos\theta-mg\sin\theta+\frac{\partial V(\theta,\phi-\omega t)}{\partial\theta}=0,\\ mr^2(\sin^2\theta\ddot\phi+2\sin\theta\cos\theta\dot\phi\dot\theta)+\frac{\partial V(\theta,\phi-\omega t)}{\partial\phi}=0. \end{aligned} \tag{A.15}$$

We work out the equations of motion for an explicit asymmetric potential V:

$$V(\theta,\phi-\psi)=-\frac{mr^2\sin^2\theta}{2}(\alpha\cos^2(\phi-\psi)+\beta\sin^2(\phi-\psi)). \tag{A.16}$$

Expanding the Lagrangian about the inverted equilibrium $\theta=\pi$, and in the rotating coordinate system, we have $x=r\theta\cos(\phi-\psi)$, $y=r\theta\sin(\phi-\psi)$, and

$$L=\frac{1}{2}m((\dot x-\omega y)^2+(\dot y+\omega x)^2)+mg\left(1-\frac{x^2+y^2}{2}\right)+\frac{m}{2}(\alpha x^2+\beta y^2). \tag{A.17}$$

Our linearized equations of motion are

$$\ddot x-2\omega\dot y+(g-\omega^2-\alpha)x=0, \tag{A.18}$$

$$\ddot y+2\omega\dot x+(g-\omega^2-\beta)y=0. \tag{A.19}$$

We would like to perturb the motion of the spherical pendulum by coupling it to a magnetic field.

Let $\mathbf{B}$ be a divergence-free vector field. Let $\mathbf{A}$ be the vector potential, $\mathbf{B}=\nabla\times\mathbf{A}$. Note if $\mathbf{B}$ is a constant magnetic field, we can choose $\mathbf{A}=\frac{1}{2}\mathbf{B}\times\mathbf{q}$.

Assume the bob on a spherical pendulum has a charge e and mass m. The Lagrangian, as a function from TS^2 to $\mathbb{R}$, is

$$L(\mathbf{q},\dot{\mathbf{q}})=\frac{1}{2}m\|\dot{\mathbf{q}}\|^2+\frac{e}{c}\mathbf{A}\cdot\dot{\mathbf{q}}-U(\mathbf{q}) \tag{A.20}$$

$$=\frac{1}{2}m(r^2\dot\theta^2+r^2\sin^2\theta\dot\phi^2)-mg\cos\theta+\frac{e}{c}\mathbf{A}\cdot\dot{\mathbf{q}}. \tag{A.21}$$

We now assume that the magnetic field $\mathbf{B}$ is constant, parallel to the gravitational field. The Lagrangian becomes

$$L(\theta,\phi,\dot\theta,\dot\phi)=\frac{1}{2}m(r^2\dot\theta^2+r^2\sin^2\theta\dot\phi^2)-mg\cos\theta+\frac{e}{2c}B\dot\phi r^2\sin^2\theta. \tag{A.22}$$

The Euler–Lagrange equations are

$$mr^2\ddot{\theta} - mr^2 \sin\theta\cos\theta\dot{\phi}^2 - mg\sin\theta - \frac{e}{c}B\dot{\phi}r^2\sin\theta\cos\theta = 0, \tag{A.23}$$

$$mr^2\sin^2\theta\ddot{\phi} + mr^2\dot{\theta}\sin\theta\cos\theta\left(\dot{\phi} + \frac{eB}{2mc}\right) = 0. \tag{A.24}$$

We expand the Lagrangian about $\theta = \pi$ and keep terms to second order in θ. In the coordinates $x = r\theta\cos(\phi - \psi)$, $y = r\theta\sin(\phi - \psi)$, we have the Lagrangian

$$L = \frac{1}{2}m((\dot{x} - \omega y)^2 + (\dot{y} + \omega x)^2) + mg - \frac{mg + \frac{eB\omega}{c}}{2}(x^2 + y^2) + \frac{eB}{2c}(x\dot{y} - y\dot{x}). \tag{A.25}$$

The angular velocity of rotation ω is constant; hence the Euler–Lagrange equations follow

$$\ddot{x} - \left(2\omega + \frac{eB}{mc}\right)\dot{y} + \left(g - \frac{eB\omega}{mc} - \omega^2\right)x = 0, \tag{A.26}$$

$$\ddot{y} + \left(2\omega + \frac{eB}{mc}\right)\dot{x} + \left(g - \frac{eB\omega}{mc} - \omega^2\right)y = 0. \tag{A.27}$$

Similarly, linearization about the inverted equilibrium yields the following equations of motion:

$$\ddot{x} - \left(2\omega + \frac{eB}{mc}\right)\dot{y} - \left(g + \frac{eB\omega}{mc} + \omega^2\right)x = 0, \tag{A.28}$$

$$\ddot{y} + \left(2\omega + \frac{eB}{mc}\right)\dot{x} - \left(g + \frac{eB\omega}{mc} + \omega^2\right)y = 0. \tag{A.29}$$

We can increase the magnetic field to *gyroscopically* stabilize the inverted equilibrium. Explicitly, for $B^2 \geq \frac{4gm^2c^2}{e^2}$, the inverted equilibrium is gyroscopically stabilized.

Appendix B. Network theory. It is also of interest to discuss the linked Chetaev system and the wave equation from the point of view of network theory and linear systems theory. As have seen, the wave coupling induces instability. It is of interest (see [3]) to analyze infinite dimensional systems from the point of view of systems theory and, in particular, to understand asymptotic stability or instability. One problem of interest is the Darlington synthesis problem (see, e.g., [7], [17] and references therein), where one shows that one can realize any positive real transfer functions by terminating a lossless two port by a 1-ohm resistor. One can then extend this to the infinite dimensional Hamiltonian setting by terminating by an infinite transmission line.

The situation here is somewhat different but related: In particular, we view the Chetaev system as a two port connected to a wave system with integral coupling. The previous calculations show that this system can be reduced to a two port with integral feedback.

Consider the two-degree-of-freedom Chetaev system with inputs:

$$\begin{aligned} \ddot{x} - B\dot{y} + \alpha x &= u_x, \\ \ddot{y} + B\dot{x} + \beta y &= u_y. \end{aligned} \tag{B.1}$$

Without coupling, we can think of this system as a linear four dimensional first order system with two inputs u_x and u_y (to be determined) and two outputs x and y. This makes the Chetaev system a classical two port.

We can also think of the wave field as a first order system:

$$\frac{\partial}{\partial t}\mathbf{w} = A\mathbf{w}, \tag{B.2}$$

where $\mathbf{w} = [w, \dot{w}]^T$ and

$$A = \begin{bmatrix} 0 & 1 \\ m^2 + c^2 \frac{\partial^2}{\partial \xi^2} & 0 \end{bmatrix}.$$

We think of the input as the second variable $\dot{w}$.

With the integral coupling, we connect the finite and the infinite systems through their inputs. Equivalently, the Chetaev system is given dynamical feedback through the wave equation. In this context, the reduced system in each type of coupling (3.35 and 4.42) becomes a first order system with integral feedback. Moreover, the integral feedback affects the Chetaev system in a not-necessarily-Hamiltonian fashion: as we have seen we may have only one unstable eigenvalue.

We remark also that instead of the mechanical systems discussed here we can physically realize the Chetaev system via coupled LC-circuits. The discussion also extends essentially without change to the n port case.

Appendix C. Bessel functions. In this appendix, we explicitly derive the integral formula

$$2\int_m^\infty \frac{\sin \omega s}{c\sqrt{\omega^2 - m^2}} d\omega = \frac{\pi}{c} J_0(ms), \tag{C.1}$$

where J_0 is the Bessel function of the first kind.

Let us begin by taking the Laplace transform, denoted $\mathcal{L}[\cdot]$, with respect to s of the right-hand side of equation (C.1),

$$\mathcal{L}\Big[2\int_m^\infty \frac{\sin \omega s}{c\sqrt{\omega^2 - m^2}} d\omega\Big] = 2\int_0^\infty \int_m^\infty \frac{\sin \omega s}{c\sqrt{\omega^2 - m^2}} e^{-sz} d\omega ds, \tag{C.2}$$

provided $\mathrm{R}e(z) > 0$. Integrating with respect to s, we have

$$\int e^{-sz} \sin \omega s \; ds = -e^{sz} \frac{\omega \cos \omega z + z \sin \omega s}{\omega^2 + z^2} \tag{C.3}$$

and

$$\mathcal{L}\Big[2\int_m^\infty \frac{\sin \omega s}{c\sqrt{\omega^2 - m^2}} d\omega\Big] = 2\int_m^\infty \frac{w}{c\sqrt{\omega^2 - m^2}(\omega^2 + z^2)} d\omega. \tag{C.4}$$

With an appropriate substitution, we use the integral

$$\int \frac{w}{\sqrt{\omega^2 - m^2}(\omega^2 + z^2)} d\omega = \frac{\arctan\Big(\frac{\sqrt{\omega^2 - m^2}}{\sqrt{m^2 + z^2}}\Big)}{\sqrt{m^2 + z^2}} \tag{C.5}$$

to finish computing the Laplace transform:

$$\mathcal{L}\Big[2\int_m^\infty \frac{\sin \omega s}{c\sqrt{\omega^2 - m^2}} d\omega\Big] = \frac{\pi}{c\sqrt{m^2 + z^2}}. \tag{C.6}$$

To finish deriving (C.1), we need only compute the Laplace transform of $J_0(s)$. This computation can be done explicitly with contour integration (see [9]); however, we present a more dynamic approach. From the Bessel differential equation, of which $J_0(s)$ is a solution, we have

$$sJ_0''(s) + J_0'(s) + sJ_0(s) = 0, \qquad J_0(0) = 1, \qquad J_0'(0) = 0, \qquad \int_0^\infty J_0(s)ds = 1. \tag{C.7}$$

Taking the Laplace transform of the differential equation, we obtain

$$-\frac{d}{dz}\Big(z^2\mathcal{L}\Big[J_0\Big] - J_0'(0) - zJ_0(0)\Big) + z\mathcal{L}\Big[J_0\Big] - J_0(0) - \frac{d}{dz}\mathcal{L}\Big[J_0\Big] = 0. \tag{C.8}$$

We solve the above differential equation, with the initial condition $\mathcal{L}[J_0](0) = \int_0^\infty J_0(s)ds = 1$, to get

$$\mathcal{L}\Big[J_0\Big] = \frac{1}{\sqrt{1 + z^2}}. \tag{C.9}$$

A change of variables produces the desired result,

$$\mathcal{L}\Big[\frac{\pi}{c}J_0(ms)\Big] = \frac{\pi}{c\sqrt{m^2 + z^2}}. \tag{C.10}$$

REFERENCES

[1] V. I. Arnol′d and A. Avez, *Ergodic Problems of Classical Mechanics*, W. A. Benjamin, New York, Amsterdam, 1968.

[2] J. Baillieul and M. Levi, *Constrained relative motions in rotational mechanics*, Arch. Ration. Mech. Anal., 115 (1991), pp. 101–135.

[3] J. S. Baras, R. W. Brockett, and P. A. Fuhrmann, *State-space models for infinite-dimensional systems*, IEEE Trans. Automat. Control, 19 (1974), pp. 693–700.

[4] A. M. Bloch and P. E. Crouch, *Representations of Dirac structures on vector spaces and nonlinear L-C circuits*, in Differential Geometry and Control (Boulder, CO, 1997), AMS, Providence, RI, 1999, pp. 103–117.

[5] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and T. S. Ratiu, *Dissipation induced instabilities*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 11 (1994), pp. 37–90.

[6] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and G. Sánchez de Alvarez, *Stabilization of rigid body dynamics by internal and external torques*, Automatica J. IFAC, 28 (1992), pp. 745–756.

[7] R. W. Brockett, *Path integrals, Lyapunov functions and quadratic minimization*, in Proceedings of the 4th Allerton Conference, University of Illinois, 1966, pp. 665–697.

[8] N. G. Chetaev, *The Stability of Motion*, Pergamon Press, New York, 1961.

[9] G. B. Folland, *Fourier Analysis and Its Applications*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1992.

[10] P. J. Hagerty, A. M. Bloch, and M. I. Weinstein, *Radiation induced instability in interconnected systems*, in Proceedings of the 38th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 1999, pp. 651–656.

[11] W. Hahn, *Stability of Motion*, Springer-Verlag, New York, 1967.

[12] J. B. Keller and L. L. Bonilla, *Irreversibility and nonrecurrence*, J. Statist. Phys., 42 (1986), pp. 1115–1125.

[13] E. Kirr and M. I. Weinstein, *Metastable states in parametrically excited multimode Hamiltonian systems*, Comm. Math. Phys., 236 (2003), pp. 335–372.
[14] E. Kirr and M. I. Weinstein, *Parametrically excited Hamiltonian partial differential equations*, SIAM J. Math. Anal., 33 (2001), pp. 16–52.
[15] A. Komech, M. Kunze, and H. Spohn, *Effective dynamics for a mechanical particle coupled to a wave field*, Comm. Math. Phys., 203 (1999), pp. 1–19.
[16] A. I. Komech, *On stabilization of string-nonlinear oscillator interaction*, J. Math. Anal. Appl., 196 (1995), pp. 384–409.
[17] P. S. Krishnaprasad, *On the geometry of linear passive systems*, in Algebraic and Geometric Methods in Linear Systems Theory (AMS-NASA-NATO Summer Seminar, Harvard University, Cambridge, MA, 1979), AMS, Providence, RI, 1980, pp. 253–275.
[18] H. Lamb, *On the peculiarity of the wave-system due to the free vibrations of a nucleus in an extended medium*, Proc. London Math. Soc., 32 (1900), pp. 208–211.
[19] R. S. MacKay, *Movement of eigenvalues of Hamiltonian equilibria under non-Hamiltonian perturbation*, Phys. Lett. A, 155 (1991), pp. 266–268.
[20] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Springer-Verlag, New York, 1999.
[21] B. M. Maschke and A. J. van der Schaft, *Interconnected mechanical systems*, in Modelling and Control of Mechanical Systems, A. E. A. Astolfi, ed., Imperial College Press, London, 1997.
[22] M. Merkli and I. Sigal, *A time-dependent theory of quantum resonances*, Comm. Math. Phys., 201 (1999), pp. 549–576.
[23] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*, W. H. Freeman, San Francisco, CA, 1973.
[24] E. J. Routh, *The Advanced Part of a Treatise on the Dynamics of a System of Rigid Bodies*, MacMillan, New York, 1905.
[25] I. Schur, *Über eine klasse von mittelbildungen mit anwendungen auf der determinantentheorie*, Sitzungsberichte der Berliner Mathematischen Gesellschaft, 22 (1923), pp. 9–20.
[26] J. C. Simo, D. Lewis, and J. E. Marsden, *Stability of relative equilibria.* I. *The reduced energy-momentum method*, Arch. Ration. Mech. Anal., 115 (1991), pp. 15–59.
[27] A. Soffer and M. I. Weinstein, *Nonautonomous Hamiltonians*, J. Statist. Phys., 93 (1998), pp. 359–391.
[28] A. Soffer and M. I. Weinstein, *Time dependent resonance theory*, Geom. Funct. Anal., 8 (1998), pp. 1086–1128.
[29] A. Soffer and M. I. Weinstein, *Resonances, radiation damping and instability in Hamiltonian nonlinear wave equations*, Invent. Math., 136 (1999), pp. 9–74.
[30] L. Thomson and P. G. Tait, *Principles of Mechanics and Dynamics*, Cambridge University Press, Cambridge, UK, 1897.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 525–545

GRATING PROFILE RECONSTRUCTION BASED ON FINITE ELEMENTS AND OPTIMIZATION TECHNIQUES*

J. ELSCHNER†, G. C. HSIAO‡, AND A. RATHSFELD†

Abstract. We consider the inverse diffraction problem to recover a two-dimensional periodic structure from scattered waves measured above and beneath the structure. The task is reformulated in the form of an optimization problem including special regularization terms. The solvability and the dependence on the parameter of regularization is analyzed. Numerical results for synthetic data demonstrate the practicability of the inversion algorithm.

Key words. diffraction grating, profile reconstruction, optimization method, conjugate gradient algorithm

AMS subject classifications. 35R30, 35J05, 78A46, 78M50

DOI. 10.1137/S0036139902420018

1. Introduction.

The scattering theory in periodic structures has many applications in micro-optics, where periodic structures are often called diffraction gratings (cf. [21] for an introduction to the direct problem). The treatment of the inverse problem, recovering the periodic structure or the shape of the grating profile from the scattered field, is useful, e.g., in quality control and design of diffractive elements with prescribed far field patterns (see [5], [22]).

Various methods for the computation of the grating profile curve of perfectly conducting gratings have been proposed by Ito and Reitich [14], Arens and Kirsch [3], Hettlich [18], Bruckner, Elschner, and Yamamoto [8], and Bruckner and Elschner [7]. Chandezon, Poyedinchuk, and Yashina [9] propose an algorithm for the determination of the interface separating dielectric substrate and superstrate materials. We follow the technique of [8] (cf. [11, section 5.4] for the original algorithms applied to obstacle scattering). However, we consider reflection by and transition through gratings described by general material dependent wave number functions and replace the boundary integral approach of [8] by a finite element algorithm (cf. [1] for a similar finite element optimization of a different functional over a set of transition curves). A related nonperiodic inverse scattering problem has been studied by Angell, Hsiao, and Wen [2] using a similar optimization procedure based upon a domain integral representation.

To be more precise, we start with a short introduction to the direct problem of diffraction by gratings in section 2. The (transverse electric) TE component of the electric field of the time-harmonic light wave is the solution of a two-dimensional Helmholtz equation over the cross section of the grating device. We recall the variational formulation corresponding to the coupling of differential and boundary integral representations and define the Rayleigh coefficients of the Helmholtz solution. These correspond to the portion of light and the phase shift of the reflected and transmitted

*Received by the editors December 19, 2002; accepted for publication (in revised form) June 11, 2003; published electronically December 31, 2003. This work was supported by the German Ministry of Education, Research and Technology under grant 03-ELM3B5.

http://www.siam.org/journals/siap/64-2/42001.html

†Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstr. 39, D-10117 Berlin, Germany (elschner@wias-berlin.de, rathsfeld@wias-berlin.de).

‡Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (hsiao@math.udel.edu).

modes. In section 3 we introduce the inverse problem. From measured Rayleigh coefficients for several incidence directions, we wish to reconstruct the grating, i.e., the wave number distribution over the grating cross section. The solution is obtained as the minimizer of an optimization problem, where the objective function consists of three terms. The first is the residual of the Helmholtz equation, the second the deviation of the computed Rayleigh coefficients from the measured data, and the third is a regularization term to cope with the ill-posedness of the inverse problem. We show the existence of minimizers and prove the convergence of these minimizers to the true solution if the regularization parameter tends to zero. Section 4 is devoted to the finite element discretization of the Helmholtz equation, and section 5 to the discretization of the optimization problem. For the solution of the finite-dimensional optimization problem we propose the conjugate gradient algorithm of Fletcher, Reeves, Polak, and Ribière (cf., e.g., [20]). In the last section we present numerical experiments. In particular, we compare the conjugate gradient algorithm with an SQP method.

Finally, we remark that the proposed treatment of the inverse problem is a first theoretical approach. Due to the severe ill-posedness of the problem, the accuracy of the reconstruction cannot be satisfactory for realistic applications. For better approximations the class of admissible gratings must be restricted in accordance with the technical requirements.

2. The direct problem. Consider an ideal optical grating (cf. the cross section in Figure 2.1). This is an ideal infinite plate in three-dimensional space covering a half space filled by a substrate material. The plate consists of different materials. Moreover, the materials are disposed in such a manner that the material does not change in one of the two directions parallel to the plane of the plate. With respect to the other direction parallel to the plane, the material distribution is supposed to be periodic with period d. The materials are nonmagnetic with the permeability μ_0 and have the dielectric constants ε. The coordinate system is chosen such that the x_2 axis is perpendicular to the plane of the grating, such that the material distribution together with the resulting diffraction solution is invariant in the x_3 direction, and such

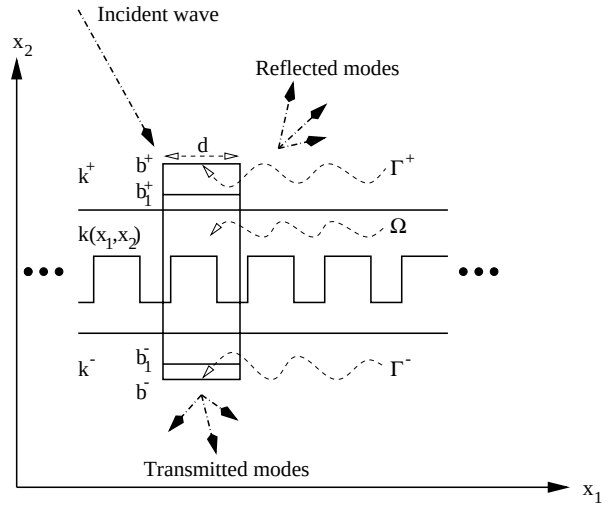


FIG. 2.1. *Cross section of grating.*

that the x_1 axis is parallel to the plane of the grating. Thus the materials of the problem are determined by the function $\varepsilon(x_1, x_2)$, which is d-periodic in x_1. Due to the invariance with respect to x_3, it is sufficient to consider the electromagnetic fields restricted to the plane spanned by the x_1 and x_2 axes. More precisely, we introduce two artificial boundaries $\Gamma^{\pm} := \{(x_1, x_2) : x_2 = b^{\pm}\}$ forming the upper and lower bounds of the cross section of the grating structure, respectively, and denote by Ω the rectangle $(0, d) \times (b^-, b^+)$, which covers one period of the cross section. We assume that the material above Γ^+ and below Γ^- is homogeneous with $\varepsilon = \varepsilon^+ > 0$ and $\varepsilon = \varepsilon^-$, respectively. Between Γ^+ and Γ^- the material may be inhomogeneous, and we assume that the function ε is piecewise continuous. Further, we introduce the wave number function $k = k(x_1, x_2) := \omega\sqrt{\mu_0\varepsilon}$ and $k^{\pm} := \omega\sqrt{\mu_0\varepsilon^{\pm}}$ with ω the angular frequency of the incident light wave. Thus the wave can be described by a time independent factor times $\exp(-\mathbf{i}\omega t)$. We suppose that

$$k^+ > 0, \quad \Re\mathrm{e}\, k^- > 0, \quad \Im\mathrm{m}\, k^- \geq 0, \quad \Re\mathrm{e}\, k(x_1, x_2) > 0, \quad \Im\mathrm{m}\, k(x_1, x_2) \geq 0. \tag{2.1}$$

Moreover, we suppose that there exists $b_1^{\pm}$ with $b^- < b_1^- < b_1^+ < b^+$ such that $k|_{\Omega^{\pm}} \equiv k^{\pm}$ for $\Omega^- := (0, d) \times (b^-, b_1^-)$ and $\Omega^+ := (0, d) \times (b_1^+, b^+)$.

Assume that an incoming plane wave is incident in the (x_1, x_2)-plane upon the grating from the top with the angle of incidence $\theta \in (-\pi/2, \pi/2)$. Then the electromagnetic field does not depend on x_3. For simplicity, we restrict ourselves to the case of TE polarization, i.e., the electric field $\mathbf{E}$ is supposed to remain parallel to the x_3 axis (to the grooves) and is therefore determined by a single scalar quantity $v = v(x_1, x_2)$ (the transverse component of $\mathbf{E}$). Due to our special geometry, Maxwell's equations for the electric and magnetic field reduce to a single equation for v. The function v satisfies the two-dimensional Helmholtz equation

$$\Delta v + k^2 v = 0 \tag{2.2}$$

in the regions with continuous permittivity ε. In the infinite regions the usual outgoing wave conditions are required. At the material interfaces the solutions are subjected to the transmission conditions, i.e., the solution v and its normal derivative $\partial_n v$ have to cross the interface continuously.

The diffraction problems admit variational formulations in the bounded periodic cell Ω which were introduced in [23, 6, 5, 15]. The incoming wave has the form $v^i(x_1, x_2) = \exp(\mathbf{i}\alpha x_1 - \mathbf{i}\beta x_2)$, where $\alpha = k^+ \sin\theta$, $\beta = k^+ \cos\theta$. If we define the function $u(x_1, x_2) := v(x_1, x_2)\exp(-\mathbf{i}\alpha x_1)$, then u can be shown to be d-periodic, and the diffraction problem for TE polarization can be transformed to a variational problem for u in the rectangle Ω. In fact, multiplying the differential equation (2.2) by some smooth function, applying Green's formula, and taking into account the transmission conditions at the material interfaces and the outgoing wave condition on $\Gamma^{\pm}$, it can be shown (cf. [23, 5, 12]) that the diffraction problem for TE polarization is equivalent to the variational equation

$$\begin{aligned} \mathcal{B}_{TE}(u, \varphi) &:= \int_\Omega \nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} - \int_\Omega k^2\, u\, \bar{\varphi} + \int_{\Gamma^+} (T_\alpha^+ u)\, \bar{\varphi} + \int_{\Gamma^-} (T_\alpha^- u)\, \bar{\varphi} \\ &= -\int_{\Gamma^+} 2\mathbf{i}\beta \exp(-\mathbf{i}\beta b^+)\, \bar{\varphi} \end{aligned} \tag{2.3}$$

for all φ, where $\nabla_\alpha = (\partial_{x_1,\alpha}, \partial_{x_2}) := \nabla + \mathbf{i}(\alpha, 0)$ and $T_\alpha^{\pm}$ are the usual hypersingular Dirichlet-to-Neumann maps for the solution in the outer domain. In particular, the

functions $T_\alpha^\pm u$ are defined on $\Gamma^\pm$ as

$$(T_\alpha^\pm u)(x_1, b^\pm) := -\sum_{n=-\infty}^{\infty} \mathbf{i}\beta_n^\pm \hat{u}_n^\pm \exp(\mathbf{i}nKx_1), \tag{2.4}$$

where $K := 2\pi/d$, and $\hat{u}_n^\pm$ denote the Fourier coefficients of $u(x_1, b^\pm)$:

$$\hat{u}_n^\pm := \frac{1}{d}\int_0^d u(x_1, b^\pm)\exp(-\mathbf{i}nKx_1)\,\mathrm{d}x_1.$$

The numbers $\beta_n^\pm$ are defined as

$$\beta_n^\pm = \beta_n^\pm(\alpha) := \sqrt{(k^\pm)^2 - \alpha_n^2}, \qquad 0 \le \arg\beta_n^\pm < \pi,$$

where as usual $\alpha_n := \alpha + nK$ and $k^\pm = k^\pm(x_1, b^\pm)$. Note that any solution of (2.3) satisfies on $\Gamma^\pm$ the nonlocal boundary conditions

$$\partial_n u|_{\Gamma^+} + T_\alpha^+ u|_{\Gamma^+} = -2\mathbf{i}\beta\exp(-\mathbf{i}\beta b^+), \quad \partial_n u|_{\Gamma^-} + T_\alpha^- u|_{\Gamma^-} = 0. \tag{2.5}$$

The variational equation (2.3) should be satisfied for all test functions $\varphi \in H^1_{per}(\Omega)$, that is, the function space of all complex-valued functions φ which are d-periodic in x_1 and which together with their first-order partial derivatives are square integrable in Ω (cf. [10] for the variational approach to classical elliptic boundary value problems).

The variational formulation (2.3) is very useful, because the transmission and outgoing wave conditions are enforced implicitly, and it allows us to seek the solution in the function space $H^1_{per}(\Omega)$, which is natural for second-order partial differential equations on nonsmooth domains. Here one can apply well-established methods for the analysis and numerical solution of the diffraction problems.

THEOREM 2.1 (cf., e.g., [12]). *Suppose that k satisfies condition* (2.1). *Then the sesquilinear form $\mathcal{B}_{TE}$ is strongly elliptic over $H^1_{per}(\Omega)$.*

We recall that a bounded sesquilinear form $\mathcal{B}_{TE}(\cdot,\cdot)$ given on the Hilbert space $H^1_{per}(\Omega)$ is called strongly elliptic if there exist a complex number ϕ, $|\phi| = 1$, a constant $c > 0$, and a compact form $\mathcal{Q}(\cdot,\cdot)$ such that

$$\Re\mathrm{e}\,\mathcal{B}_{TE}(\phi\,u, u) \ge c\|u\|_X^2 - \mathcal{Q}(u,u) \quad \forall\, u \in H^1_{per}(\Omega).$$

As usual, the sesquilinear form $\mathcal{B}_{TE}$ corresponds to a bounded linear operator B mapping $H^1_{per}(\Omega)$ into its dual $H^1_{per}(\Omega)'$ via $\mathcal{B}_{TE}(u,v) = \langle Bu, v\rangle$, $u, v \in H^1_{per}(\Omega)$. According to the proof of the last theorem (cf. [12]), the bilinear form $\mathcal{B}_{TE}$ splits into the compact form $\mathcal{C}_k(u,v) := -k^2\int_\Omega uv$, a strongly elliptic form $\mathcal{P}$ with $\mathcal{P}(u,u) \ge c\|u\|^2_{H^1_{per}(\Omega)}$ and constant $c > 0$, and a finite-dimensional form $\mathcal{T}$. Correspondingly, we get $B = P + T + C_k$ with $\langle Pu, u\rangle \ge c\|u\|^2_{H^1_{per}(\Omega)}$, with finite range operator T, and with compact C_k. From this splitting we infer that B is a Fredholm operator of index zero. Thus the strong ellipticity is the basis for proving the invertibility of operator B under additional conditions.

We write $B = B(k,\theta)$ to indicate the dependence of B on the wave number function k and on the incidence angle θ. The variational equation (2.3) is equivalent to the operator equation $B(k,\theta)u = w$ with $w \in H^1_{per}(\Omega)'$. Here $\langle w, \varphi\rangle$, i.e., the functional w applied to $\varphi \in H^1_{per}(\Omega)$ is defined by the right-hand side of (2.3). The operator $B(k,\theta)$ is a second-order differential operator. To get an equation with a

well-conditioned operator acting in the single space $H^1_{per}(\Omega)$, we multiply the equation by the inverse of $\widetilde{B}(\theta) := B(\tilde{k}, \theta)$ with a fixed simple wave number function $\tilde{k}$. Thus (2.3) is equivalent to $[\widetilde{B}(\theta)^{-1}B(k,\theta)]u = \widetilde{B}(\theta)^{-1}w$.

The invertibility of the operators $\widetilde{B}(\theta)$ and $B(k,\theta)$ will be supposed in the following. Partial results on this are reported, e.g., in [23, 5, 12]. Here we give only a stability result with respect to the wave number function.

THEOREM 2.2. *Suppose that the squared wave number functions k_n^2 form a weakly convergent sequence in the space $L^2(\Omega)$. (Note that the squared wave number function enters linearly into the scene.) If $B(k_0,\theta)$ is the operator defined with k_0 such that k_0^2 is the weak limit of the k_n^2 and if this $B(k_0,\theta)$ is invertible, then there exist an integer $n_0 > 0$ and a real $c > 0$ such that $\|B(k_n,\theta)u\|_{H^1_{per}(\Omega)'} \geq c\|u\|_{H^1_{per}(\Omega)}$ for any $u \in H^1_{per}(\Omega)$ and $n \geq n_0$. Since the $B(k_n,\theta)$ are Fredholm operators with index zero, the last estimate implies the invertibility of $B(k_n,\theta)$ if $n \geq n_0$.*

Proof. If the theorem were not true, then there is a sequence $\{u_n\} \subset H^1_{per}(\Omega)$ such that $\|u_n\|_{H^1_{per}(\Omega)} = 1$ and $\|B(k_n,\theta)u_n\|_{H^1_{per}(\Omega)'} \to 0$. Then, without loss of generality, we may suppose that u_n tends weakly to u_0 in $H^1_{per}(\Omega)$. Hence, $\|u_n - u_0\|_{L^p(\Omega)} \to 0$ for any p with $1 \leq p < \infty$. From the weak convergence of u_n we infer the weak convergence in $H^1_{per}(\Omega)'$ of $[P+T]u_n \rightharpoonup [P+T]u_0$. Indeed, $Tu_n \to Tu_0$ and $\langle Pu_n, \varphi\rangle = \langle u_n, P'\varphi\rangle \to \langle u_0, P'\varphi\rangle = \langle Pu_0, \varphi\rangle$ for all $\varphi \in H^1_{per}(\Omega)$.

Furthermore, since $\|u_n - u_0\|_{L^p(\Omega)} \to 0$ for any $p < \infty$, we obtain the relation $\|u_n\overline{\varphi} - u_0\overline{\varphi}\|_{L^2(\Omega)} \to 0$ for any $\varphi \in H^1_{per}(\Omega)$. Hence, $\mathcal{C}_k(u_n,\varphi) = -\int_\Omega k_n^2 u_n\overline{\varphi} \to -\int_\Omega k_0^2 u_0\overline{\varphi}$. Together with the weak convergence $[P+T]u_n \rightharpoonup [P+T]u_0$, we have $B(k_n,\theta)u_n \rightharpoonup B(k_0,\theta)u_0$. This implies $B(k_0,\theta)u_0 = 0$ and $u_0 = 0$. Consequently, $C_{k_n}u_n \to 0$ and $Tu_n \to 0$ together with $\|B(k_n,\theta)u_n\|_{H^1_{per}(\Omega)'} \to 0$ yield $Pu_n \to 0$, which contradicts $\langle Pu_n, u_n\rangle \geq c\|u_n\|^2_{H^1_{per}(\Omega)} = 1$. ∎

Note that any periodic solution of (2.3) can be represented as a Fourier series on $\Gamma^\pm$, i.e.,

$$
\begin{aligned}
u(x_1,b^+) &= \sum_{n=-\infty}^{\infty} A_n^+ \exp(\mathbf{i}\beta_n^+ b^+)\, \exp(\mathbf{i}nKx_1) + \exp(-\mathbf{i}\beta b^+),\\
u(x_1,b^-) &= \sum_{n=-\infty}^{\infty} A_n^- \exp(-\mathbf{i}\beta_n^- b^-)\, \exp(\mathbf{i}nKx_1)
\end{aligned}
\tag{2.6}
$$

for suitable coefficients $A_n^\pm$. It is not hard to see that the extensions of these series multiplied by the factor $\exp(\mathbf{i}\alpha x_1)$ (recall that $v(x_1,x_2) = u(x_1,x_2)\exp(\mathbf{i}\alpha x_1)$),

$$
\sum_{n=-\infty}^{\infty} A_n^\pm \exp(\pm\mathbf{i}\beta_n^\pm x_2)\, \exp(\mathbf{i}[\alpha+n]Kx_1), \quad x_2 \gtrless b^\pm,
\tag{2.7}
$$

define solutions of the Helmholtz equation satisfying the outgoing wave condition. The coefficients $A_n^\pm$ in the expansion (2.7) are called Rayleigh coefficients. The most interesting are those with $n \in \mathcal{U}^\pm$,

$$
\mathcal{U}^\pm := \begin{cases} \{n \in \mathbb{Z} : |n+\alpha| < k^\pm\} & \text{if } \Im\mathrm{m}\, k^\pm = 0,\\ \emptyset & \text{if } \Im\mathrm{m}\, k^\pm > 0. \end{cases}
$$

Indeed, these coefficients $A_n^\pm$ describe the magnitude and the phase shift of those

terms $A_n^{\pm}\exp(\mathbf{i}[\alpha+n]Kx_1)\exp(\pm\mathbf{i}\beta_n^{\pm}x_2)$ in the representation of $u(x_1,x_2)\exp(\mathbf{i}\alpha x_1)$ for $x_2 \gtrless b^{\pm}$, which correspond to propagating plane waves. The terms with $n \notin \mathcal{U}^{\pm}$ lead to evanescent waves only. Hence, the $A_n^{\pm}$ with $n \in \mathcal{U}^{\pm}$ can be considered to be the far field data of the diffraction problems at optical gratings. The optical efficiencies of the grating are defined by

$$e_n^{\pm} := (\beta_n^{\pm}/\beta)|A_n^{\pm}|^2,\ (n,\pm) \in \mathcal{U}^* := \left\{(n,+):\ n \in \mathcal{U}^+\right\} \cup \left\{(n,-):\ n \in \mathcal{U}^-\right\}, \tag{2.8}$$

which is the ratio of the energy of the nth propagating mode to the energy of the incident wave.

Restricting the solution $\Omega \ni (x_1,x_2) \mapsto u(k,\theta)(x_1,x_2)$ to $\Gamma^{\pm}$, we get the Rayleigh coefficients $A_n^{\pm}$ by computing the Fourier coefficients according to (2.6). The linear operator of restricting u to $\Gamma^{\pm}$ and of computing the Rayleigh coefficients $A_n^{\pm}$ will be denoted by $F(\theta)$, i.e.,

$$\begin{aligned} A &:= \left(A_n^{\pm}\right)_{(n,\pm)\in\mathcal{U}^*} \ = \ F(\theta)\ u + A^i, \\ A^i &:= \left(-\exp(-\mathbf{i}\beta b^+)\delta_{(n,\pm),(0,+)}\right)_{(n,\pm)\in\mathcal{U}^*}. \end{aligned}$$

If the refractive indices of the cover material above the grating and the substrate material beneath the grating are fixed, then the operator $F(\theta)$ is independent of the function k inside the grating.

3. The inverse problem. For the inverse problem, we suppose that the distribution of the material in the grating between the lines $\{(x_1,b_1^+):\ 0<x_1<d\}$ and $\{(x_1,b_1^-):\ 0<x_1<d\}$ is unknown. In other words, our task is to determine the unknown function $k(x_1,x_2)$ for $b_1^- < x_2 < b_1^+$. To get this, we illuminate the grating by plane waves $v^i(x_1,x_2) = \exp(\mathbf{i}k^+\sin\theta\, x_1 - ik^+\cos\theta\, x_2)$ under the incident angles $\theta = \theta_l,\ l = 1,\dots,L$, and measure the Rayleigh coefficients $A_{meas,n}^{\pm}(\theta_l)$ for $(n,\pm) \in \mathcal{U}^* = \mathcal{U}^*(\theta_l)$ and for each angle $\theta_l,\ l = 1,\dots,L$. We seek a material distribution and the corresponding wave number function $k(x_1,x_2)$ such that the Rayleigh coefficients $A_n^{\pm} = A_n^{\pm}(k,\theta_l)$, obtained by solving the variational equation (2.3) with respect to $u(x_1,x_2) = u(k,\theta_l)(x_1,x_2)$ and by computing the Fourier coefficients $A_n^{\pm}(k,\theta_l)$ of $u|_{\Gamma^{\pm}}$ according to (2.6), coincide with the measured data $A_{meas,n}^{\pm}(\theta_l)$ for $(n,\pm) \in \mathcal{U}^*(\theta_l)$ and $l = 1,\dots,L$. In other words, we seek an unknown squared wave number function k^2 and the corresponding solutions $u(k,\theta_l)$ from $[\widetilde{B}(\theta_1)^{-1}B(k,\theta_l)]u(k,\theta_l) = \widetilde{B}(\theta_1)^{-1}w(\theta_l)$ such that the computed Rayleigh coefficients $A(\theta_l) = F(\theta_l)u(k,\theta_l) + A^i(\theta_l)$ coincide with the measured $A_{meas}(\theta_l) := (A_{meas,n}^{\pm}(\theta_l))_{(n,\pm)\in\mathcal{U}^*(\theta_l)}$. Expressing our objective in formulae, we seek k^2 and $u_l = u(k,\theta_l)$ such that

$$\begin{aligned} &\sum_{l=1}^{L} \left\| \left[\widetilde{B}(\theta_1)^{-1}B(k,\theta_l)\right] u_l - \widetilde{B}(\theta_1)^{-1}w_l \right\|^2_{L^2(\Omega)} \ = 0, \\ &\sum_{l=1}^{L} \left\| \left[F(\theta_l)\ u_l + A^i(\theta_l)\right] - A_{meas}(\theta_l) \right\|^2_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))} = 0. \end{aligned}$$

Here $w_l = w(\theta_l)$ stands for the right-hand-side functional in (2.3) with θ replaced by θ_l. The symbol $\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))$ denotes the complex Euclidean space of vectors over the index set $\mathcal{U}^*(\theta_l)$.

The operator $F(\theta_l)$ is smoothing and the equation $F(\theta_l)u_l = A(\theta_l) - A^i(\theta_l)$ is severely ill-posed. To cope with measurement errors in the values of $A_{meas}(\theta_l)$ we need a regularization, i.e., we try to find solutions k^2 and u_l such that the left-hand sides of the last two equations are small and that, simultaneously, the solution is relatively smooth. Relatively smooth means that the H^1_{per} Sobolev norms of u_l and the $H^{1/2}_{per}$ Sobolev norm of k^2 do not blow up. This will be helpful also if the solution should not be unique. Finally, we define the nonlinear objective functional

$$\begin{aligned}\mathcal{F}\left(k^2, u_1, \dots, u_L; \gamma\right) := & \frac{\sum_{l=1}^{L} \left\| \widetilde{B}(\theta_1)^{-1} B(k,\theta_l) u_l - \widetilde{B}(\theta_1)^{-1} w_l \right\|^2_{L^2(\Omega)}}{\sum_{l=1}^{L} \left\| \widetilde{B}(\theta_1)^{-1} w_l \right\|^2_{L^2(\Omega)}} \\ & + c_d \frac{\sum_{l=1}^{L} \left\| \left[F(\theta_l)\, u_l + A^i(\theta_l)\right] - A_{meas}(\theta_l) \right\|^2_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))}}{\sum_{l=1}^{L} \|A_{meas}(\theta_l)\|^2_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))}} \\ & + c_v \gamma \left\|k^2\right\|^2_{H^{1/2}_{per}(\Omega)} + c_s \gamma \sum_{l=1}^{L} \|u_l\|^2_{H^1_{per}(\Omega)}. \end{aligned} \tag{3.1}$$

Here c_d, c_v, and c_s denote appropriate equilibration constants which are to be determined by numerical experiments. At first glance, one might think that the last regularization term in (3.1) is unnecessary. Indeed, the first term containing the constraint condition forces u_l to have a bounded L^2 norm. Unfortunately, our proof heavily relies on the boundedness of the stronger H^1 norm. In our numerical experiments (cf. section 6), we have tested various choices of c_s. We have found that a certain small $c_s\gamma$ leads to better results than $c_s = 0$. The number γ is a small positive regularization parameter which is to be chosen in dependence on the measurement error. Using the functional F, we finally arrive at the optimization problem

$$\begin{aligned} & \mathcal{F}\left(k^2, u_1, \dots, u_L; \gamma\right) \quad \longrightarrow \min, \\ & \text{with } k^2 \in H^{1/2}_{per}(\Omega), \\ & u_l \in H^1_{per}(\Omega),\ l = 1, \dots, L. \end{aligned} \tag{3.2}$$

This optimization problem will be discretized and solved numerically in the subsequent sections. For its solvability and its connection to the exact inverse problem, we get the following two theorems.

THEOREM 3.1. *For any fixed positive regularization parameter γ, there exists a minimizer $\{k_0^2,\ u_{l,0},\ l = 1, \dots, L\}$ of the optimization problem* (3.2).

Proof. Suppose $\{k_n^2,\ u_{l,n},\ l = 1, \dots, L\}_{n\in\mathbb{N}}$ is a minimizing sequence. Without loss of generality we may suppose $k_n^2 \rightharpoonup k_0^2$ in $H^{1/2}_{per}(\Omega)$, $k_n^2 \to k_0^2$ in $L^2(\Omega)$, and $u_{l,n} \rightharpoonup u_{l,0}$ weakly in $H^1_{per}(\Omega)$ since $\|k_n^2\|_{H^{1/2}_{per}(\Omega)}$ and $\|u_{l,n}\|_{H^1_{per}(\Omega)}$ are trivially bounded. Similarly to the proof of Theorem 2.2, we conclude $B(k_n,\theta_l)u_{l,n} \rightharpoonup B(k_0,\theta_l)u_{l,0}$ weakly in $H^1_{per}(\Omega)'$, and thus $\widetilde{B}(\theta_1)^{-1}B(k_n,\theta_l)u_{l,n} \rightharpoonup \widetilde{B}(\theta_1)^{-1}B(k_0,\theta_l)u_{l,0}$ weakly in $H^1_{per}(\Omega)$. Hence, $\widetilde{B}(\theta_1)^{-1}B(k_n,\theta_l)u_{l,n} \to \widetilde{B}(\theta_1)^{-1}B(k_0,\theta_l)u_{l,0}$ strongly in $L^2(\Omega)$. Moreover, $u_{l,n} \rightharpoonup u_{l,0}$ implies that $u_{l,n}|_{\Gamma^\pm} \to u_{l,0}|_{\Gamma^\pm}$ strongly in $L^2(\Gamma^\pm)$ and the

strong convergence $F(\theta_l)u_{l,n} \to F(\theta_l)u_{l,0}$. In other words, the first two terms in the objective functionals converge, and the limit relations for weakly convergent sequences $\|u_{l,n}\|_{H^1_{per}(\Omega)} \le \liminf \|u_{l,n}\|_{H^1_{per}(\Omega)}$ and $\|k_0^2\|_{H^{1/2}_{per}(\Omega)} \le \liminf \|k_n^2\|_{H^{1/2}_{per}(\Omega)}$ lead us to the upper estimate $\mathcal{F}(k_0^2, u_{1,0}, \dots, u_{L,0}; \gamma) \le \liminf \mathcal{F}(k_n^2, u_{1,n}, \dots, u_{L,n}; \gamma)$. Since $\{k_n^2,\ u_{l,n},\ l = 1, \dots, L\}_{n\in\mathbb{N}}$ is a minimizing sequence, we conclude that the value $\mathcal{F}(k_0^2, u_{1,0}, \dots, u_{L,0}; \gamma)$ is the attained minimum. ∎

THEOREM 3.2. *Suppose that, for the given data $A_{meas}(\theta_1), \dots, A_{meas}(\theta_L)$, there exists a wave number function $k_*^2 \in H^{1/2}_{per}(\Omega)$ such that the Rayleigh coefficients corresponding to k_* exactly match the values $A_{meas}(\theta_1), \dots, A_{meas}(\theta_L)$, i.e., $F(\theta_l)u(k_*, \theta_l) + A^i(\theta_l) = A_{meas}(\theta_l)$ for the solutions $u(k_*, \theta_l)$ of $B(k_*, \theta_l)u(k_*, \theta_l) = w_l$. Further suppose $0 < \gamma_m \to 0$ and that $\{k_m^2,\ u_{l,m},\ l = 1, \dots, L\}$ is a minimizer of the functional $\mathcal{F}(\dots; \gamma_m)$. Then there exists a $k_0^2 \in H^{1/2}_{per}(\Omega)$ and a subsequence of $\{k_m^2\}_{m\in\mathbb{N}}$ converging to k_0^2 weakly in $H^{1/2}_{per}(\Omega)$ and strongly in $L^2(\Omega)$. The corresponding solutions $u(k_0, \theta_l)$ of the variational equations (cf. (2.3)) or equivalently of $B(k_0, \theta_l)u(k_0, \theta_l) = w_l$ satisfy $F(\theta_l)u(k_0, \theta_l) + A^i(\theta_l) = A_{meas}(\theta_l)$; i.e., their Rayleigh coefficients coincide with the measured data $A_{meas}(\theta_l)$ for $l = 1, \dots, L$.*

Proof. From our assumption on the existence of k_* and from

$$
\begin{aligned}
c_v\gamma_m \left\|k_m^2\right\|^2_{H^{1/2}_{per}(\Omega)} + c_s\gamma_m \sum_{l=1}^{L} \|u_{l,m}\|^2_{H^1_{per}(\Omega)} &\le \mathcal{F}\left(k_m^2, u_{1,m}, \dots, u_{L,m}; \gamma_m\right) \\
&\le \mathcal{F}\left(k_*^2, u(k_*, \theta_1), \dots, u(k_*, \theta_L); \gamma_m\right) \\
&= c_v\gamma_m \left\|k_*^2\right\|^2_{H^{1/2}_{per}(\Omega)} \\
&\quad + c_s\gamma_m \sum_{l=1}^{L} \|u(k_*, \theta_1)\|^2_{H^1_{per}(\Omega)} \longrightarrow 0,
\end{aligned}
\tag{3.3}
$$

we obtain the uniform boundedness of $\|k_m^2\|_{H^{1/2}_{per}(\Omega)}$ and $\|u_{l,m}\|_{H^1_{per}(\Omega)}$. Therefore, we can switch to weakly convergent subsequences. Without loss of generality suppose that k_m^2 and $u_{l,m}$ converge weakly in the corresponding Sobolev spaces. Repeating the arguments of the proof to Theorem 3.1 and using (3.3) leads to

$$
\begin{aligned}
\mathcal{F}\left(k_m^2, u_{1,m}, \dots, u_{L,m}; \gamma_m\right) \longrightarrow\ & \frac{\displaystyle\sum_{l=1}^{L} \left\| \widetilde{B}(\theta_1)^{-1} B(k_0, \theta_l) u_{l,0} - \widetilde{B}(\theta_1)^{-1} w_l \right\|^2_{L^2(\Omega)}}{\displaystyle\sum_{l=1}^{L} \left\| \widetilde{B}(\theta_1)^{-1} w_l \right\|^2_{L^2(\Omega)}} \\
& + c_d\, \frac{\displaystyle\sum_{l=1}^{L} \left\| \left[F(\theta_l)\ u_{l,0} + A^i(\theta_l)\right] - A_{meas}(\theta_l) \right\|^2_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))}}{\displaystyle\sum_{l=1}^{L} \|A_{meas}(\theta_l)\|^2_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))}} \\
=\ & 0.
\end{aligned}
$$

The assertions of the theorem follow. ∎

COROLLARY 3.3. *Suppose the assumptions of the last theorem and, additionally, that the wave number function k_* is the unique solution of the inverse problem, i.e., that the relations $F(\theta_l)u(k, \theta_l) + A^i(\theta_l) = A_{meas}(\theta_l)$ and $B(k, \theta_l)u(k, \theta_l) = w_l$ for*

$k = 1, \dots, L$ imply $k_* = k$. Then the whole sequence $\{k_m^2\}_{m\in\mathbb{N}}$ converges to k_*^2 weakly in $H^{1/2}_{per}(\Omega)$ and strongly in $L^2(\Omega)$.

Proof. The proof is straightforward since a sequence is convergent if all subsequences have subsequences with a fixed limit. ∎

Remark 3.1. In general, the uniqueness assumption is hard to verify. For perfectly conducting gratings bounded by a curve of small oscillation represented as a finite Fourier series, uniqueness is proved in [16]. In [15] it has been shown that the knowledge of a finite number of Rayleigh coefficients even for all incident angles is not sufficient to determine the grating. The situation improves slightly if the measurement of Rayleigh coefficients is replaced by the measurement of the field u restricted to the lines $\{(x_1, b_2^+) : \ 0 < x_1 < d\}$ and $\{(x_1, b_2^-) : \ 0 < x_1 < d\}$ with $b_2^- < b^- < b^+ < b_2^+$. Note that the differences between the two data types are not so essential if the second data type is discretized. Moreover, the theoretical results of this section remain valid for the new kind of measurements. The case of smooth wave number functions depending only on the x_1 variable is treated in [15]. For grating structures corresponding to perfectly conducting gratings bounded by C^2 curves and for the reflected data measured in any direction of incidence, uniqueness is shown in [16]. A fixed incidence direction together with measured data corresponding to a finite number of wave lengths λ is treated in [19]. Gratings consisting of two materials (corresponding to the wave numbers $k^\pm$) separated by a Lipschitz curve and absorbing substrate materials are considered in the subsequent Theorem 3.5. If only local uniqueness in the inverse problem is known, then the optimization problem and the numerical methods in section 5 with suitable initial guess can be used to recover the grating. For local uniqueness, we refer to the local stability results and the papers quoted in [13].

COROLLARY 3.4. *Suppose the assumptions of Theorem* 3.2 *are satisfied. However, consider noisy data $A^{noisy}_{meas}(m, \theta_l) \in \ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))$ such that the error to the exactly measured data $A_{meas}(\theta_l)$ satisfies $\|A^{noisy}_{meas}(m, \theta_l) - A_{meas}(\theta_l)\|_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))} \le \gamma_m$. Suppose the minimizers are determined for the functional $\mathcal{F}(\dots; \gamma_m)$ with $A_{meas}(\theta_l)$ replaced by $A^{noisy}_{meas}(m, \theta_l)$. Then the assertions of Theorem* 3.2 *remains valid. If, additionally, k_* is the unique solution of the inverse problem, then the assertions of Corollary* 3.3 *stay in force.*

Proof. The proof is a straightforward modification of that to Theorem 3.2. ∎

THEOREM 3.5. *Assume that the graphs $\{(x_1, f_j(x_1)) : \ 0 < x_1 < d\}$ of two different Lipschitz continuous functions f_j $(j = 1, 2)$ cut Ω into an upper region $\{(x_1, x_2) : \ f_j(x_1) < x_2 < b^+\}$ with constant wave number $k^+ > 0$ and a lower region $\{(x_1, x_2) : \ b^- < x_2 < f_j(x_1)\}$ with constant wave number k^- such that $\Re\mathrm{e}\, k^- > 0$ and $\Im\mathrm{m}\, k^- > 0$. For these two gratings and for one planar incident wave $(L = 1)$, we assume that u_1 and u_2 are the solutions of the TE problem. (That is, $\triangle u_j - [k^+]^2 u_j = 0$ holds on $\{(x_1, x_2) : \ f_j(x_1) < x_2 < b^+\}$ and $\triangle u_j - [k^-]^2 u_j = 0$ holds on $\{(x_1, x_2) : \ b^- < x_2 < f_j(x_1)\}$, and the functions u_j and their normal derivatives are continuous across the surfaces $\{(x_1, f_j(x_1)) : \ 0 < x_1 < d\}$.) Then coincidence of the data $u_1|_{\Gamma^+} = u_2|_{\Gamma^+}$ and $u_1|_{\Gamma^-} = u_2|_{\Gamma^-}$ implies $f_1 = f_2$.*

Remark 3.2. This generalizes the uniqueness result by Bao [4] for a perfectly reflecting substrate material below the interface.

Proof. Setting $f(x_1) := \max\{f_1(x_1), f_2(x_1)\}$ and $g(x_1) := \min\{f_1(x_1), f_2(x_1)\}$, we consider the function $u := u_1 - u_2$. Then $u|_{\Gamma^+} = 0$, $u|_{\Gamma^-} = 0$, $\partial_\nu u|_{\Gamma^+} = 0$, and $\partial_\nu u|_{\Gamma^-} = 0$ (cf. (2.5)), which together with the unique continuation theorem implies $u = 0$ in the regions $\{(x_1, x_2) : \ f(x_1) < x_2 < b^+\}$ and $\{(x_1, x_2) : \ b^- < x_2 < g(x_1)\}$.

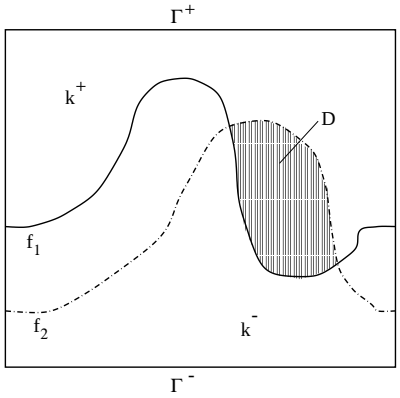


FIG. 3.1. *Gratings defined by two graphs.*

If D (cf. Figure 3.1) is a simply connected region bounded by the graphs of f and g (and possibly by the vertical lines $\{(x_1, x_2) : x_2 = 0\}$ and $\{(x_1, x_2) : x_2 = d\}$), then we have $\Delta u_1 + [k^+]^2 u_1 = 0$ and $\Delta u_2 + [k^-]^2 u_2 = 0$ in D, or vice versa. Additionally, we get $u_1 = u_2$ and $\partial_\nu u_1 = \partial_\nu u_2$ on the boundary ∂D of D, where ∂_ν stands for the normal derivative at the boundary points of ∂D. Applying Green's formula, which is justified for $u_j \in H^2(\Omega)$, $j = 1, 2$, we arrive at

$$0 = \int_D \{u_1 \Delta \overline{u}_1 - \overline{u}_1 \Delta u_1\} = \int_{\partial D} \{u_1 \partial_\nu \overline{u}_1 - \overline{u}_1 \partial_\nu u_1\} = \int_{\partial D} \{u_2 \partial_\nu \overline{u}_2 - \overline{u}_2 \partial_\nu u_2\}$$
$$= \int_D \{u_2 \Delta \overline{u}_2 - \overline{u}_2 \Delta u_2\} = 2\mathbf{i} \Im\mathrm{m}\, [k^-]^2 \int_D |u_2|^2 = 0.$$

Note that for the third equality we have used the quasi-periodicity of the solutions u_j leading to $\{u_j \partial_\nu \overline{u}_j - \overline{u}_j \partial_\nu u_j\} = 0$ over the vertical boundary parts of D. Therefore, $u_2 = 0$ in D. Consequently, $u_2 = 0$ in Ω, which is a contradiction to the fact that u_2 is the scattered wave component corresponding to a nonzero incident wave. ∎

4. The finite element solution. To define the finite element method, we split domain Ω into the union of triangles such that the diameter of each triangle is less than a prescribed mesh size h and that the triangles have no interior points in common. Moreover, we assume that any two triangles of the partition are either disjoint or their intersection is a common edge or a common corner point (no hanging nodes). By S_h^1 we denote the set of all piecewise linear functions subordinate to the partition. Then, the finite element solution u_h of the Helmholtz equation (2.2) in its variational form (2.3) is the unique solution $u_h \in S_h^1$ satisfying

$$\mathcal{B}_{TE}(u_h, \varphi_h) = \int_\Omega \nabla_\alpha u_h \cdot \overline{\nabla_\alpha \varphi_h} - \int_\Omega k^2\, u_h\, \overline{\varphi_h} + \int_{\Gamma^+} (T_\alpha^+ u_h)\, \overline{\varphi_h} + \int_{\Gamma^-} (T_\alpha^- u_h)\, \overline{\varphi_h}$$
$$= -\int_{\Gamma^+} 2\mathbf{i}\beta \exp(-\mathbf{i}\beta b^+)\, \overline{\varphi_h} \quad \forall \varphi_h \in S_h^1. \tag{4.1}$$

Clearly, choosing the usual hat function basis $\{\varphi_{h,j} : j = 1, \ldots, N\}$ of S_h^1, the last discrete variational equation is equivalent to an equation in the N-dimensional complex Euclidean space $\ell^2_{\mathbb{C}}(N)$, i.e., to the matrix equation $B_h \xi = \eta$ for the unknown

coefficients ξ_j of the function u_h, where

$$B_h := (\mathcal{B}_{TE}(\varphi_{h,j}, \varphi_{h,j'}))_{j',j=1,\dots,N},$$
$$\xi := (\xi_j)_{j=1,\dots,N}, \quad u_h(x_1,x_2) \;=\; \sum_{j=1}^{N} \xi_j\, \varphi_{h,j}(x_1,x_2),$$
$$\eta := (\eta_j)_{j=1,\dots,N}, \quad \eta_j := -\int_{\Gamma^+} 2\mathbf{i}\beta \exp(-\mathbf{i}\beta b^+)\, \overline{\varphi_{h,j}}\,.$$

In other words, if the function $k = k(x_1, x_2)$ is given, then we can determine an approximate solution u_h by solving $B_h\xi = \eta$. Clearly, the matrix, the right-hand side, and the solution depend on the angle of incidence θ and on the wave number function $k = k(x_1, x_2)$. To indicate this dependence, we write $u_h = u_h(k, \theta)$ and the matrix equation as $B_h(k, \theta)\xi(k, \theta) = \eta(\theta)$.

Restricting the solution $\Omega \ni (x_1, x_2) \mapsto u_h(k, \theta)(x_1, x_2)$ to $\Gamma^\pm$, we get the Rayleigh coefficients $A_n^\pm$ by computing the Fourier coefficients according to (2.6). We denote the so-obtained approximate values of $A_n^\pm$ by $A_{h,n}^\pm$ and get the approximate efficiencies by setting $e_{h,n}^\pm := (\beta_n^\pm/\beta)|A_{h,n}^\pm|^2$. The linear operator of restricting u_h to $\Gamma^\pm$ and of computing the Rayleigh coefficients $A_{h,n}^\pm$ will be denoted by $F_h(\theta)$, i.e.,

$$A_h := \left(A_{h,n}^\pm\right)_{(n,\pm)\in\mathcal{U}^*} \;=\; F_h(\theta)\,\xi + A^i(\theta).$$

If the refractive indices of the cover material above the grating and the substrate material beneath the grating are fixed, then the operator $F_h(\theta)$ is independent of the function k inside of the grating.

Finally, we remark that the linear system of equations $B_h(k, \theta)\xi(k, \theta) = \eta(\theta)$ is the discretization of a second-order differential equations. Consequently, the condition number of the finite element matrix $B_h(k, \theta)$ behaves like $\mathcal{O}(h^{-2})$ for h tending to zero. Therefore, a preconditioner is used for the iterative solution of $B_h(k, \theta)\xi(k, \theta) = \eta(\theta)$, i.e., we solve $[\widetilde{B}_h(\theta)^{-1}B_h(k, \theta)]\xi(k, \theta) = \widetilde{B}_h(\theta)^{-1}\eta(\theta)$ instead of $B_h(k, \theta)\xi(k, \theta) = \eta(\theta)$ with a matrix $\widetilde{B}_h(\theta)$ easy to invert and close to $B_h(k, \theta)$. Several preconditioning techniques are possible. In our special case, we can choose $\widetilde{B}_h(\theta)$, e.g., as the finite element matrix $\widetilde{B}_h(\theta) := B_h(\tilde{k}, \theta)$, where the wave number function $\tilde{k}(x_1, x_2)$ is equal to k^+ for $x_2 > (b^- + b^+)/2$ and equal to k^- for $x_2 < (b^- + b^+)/2$. If the partition of the finite element method is obtained by dividing the rectangles of a uniform rectangular partition of the rectangle Ω along the diagonals, then $B_h(\tilde{k}, \theta)$ is easy to invert. Indeed, if we group the degrees of freedom in clusters according to their x_2 coordinates, then $B_h(\tilde{k}, \theta)$ is a triangular block matrix with circular blocks.

5. The discretized inverse problem. For a numerical solution of the inverse problem of section 3, we switch to the discrete level, i.e., we seek the coefficient vectors $\xi(l)$ of the finite element solutions $u_h(k, \theta_l) = \sum \xi(l)_j \varphi_{h,j}$ from $\widetilde{B}_h(\theta_1)^{-1}B_h(k, \theta_l)\xi(l) \approx \widetilde{B}_h(\theta_1)^{-1}\eta(\theta_l)$ such that the computed Rayleigh coefficients $A_h(\theta_l) = F_h(\theta_l)\xi(l) + A^i(\theta_l)$ differ only slightly from the measured $A_{meas}(\theta_l) := (A_{meas,n}^\pm(\theta_l))_{(n,\pm)\in\mathcal{U}^*(\theta_l)}$. The unknown squared wave number function k^2 is to be approximated by a function from a discrete space. We fix a partition coarser than that of the finite element method and choose the space S_h^0 as the set of all functions which are piecewise constant subordinate to the fixed partition and which fulfil $k^2(x_1, x_2) = [k^+]^2$ for $0 < x_1 < d$ and $b_1^+ < x_2 < b^+$ as well as $k^2 = [k^-]^2$ for $0 < x_1 < d$ and $b^- < x_2 < b_1^-$. As usual, the corresponding basis of functions equal

to one over one triangle of the grid and to zero over the others will be denoted by $\{\chi_{h,j} : j = 1, \dots, M\}$. We can identify the functions $k^2 \in S_h^0$ with the vectors of coefficients $\kappa = (\kappa_j)_{j=1,\dots,M}$ satisfying $k(x_1, x_2)^2 = \sum_j \kappa_j \chi_{h,j}(x_1, x_2)$. In particular, we write $B_h(\kappa, \theta_l)$ for $B_h(\sum_j \sqrt{\kappa_j}\chi_{h,j}, \theta_l)$. Using the discretized and reduced $H^{1/2}_{per}(\Omega)$ norm

$$\|\kappa\|^2_{V(\Omega)} := \sum_j |\kappa_j|^2 w_j + \sum_{\substack{j,j': \\ \text{indices of} \\ \text{neighbors}}} |\kappa_j - \kappa_{j'}|^2 \sqrt{w_j}$$

with w_j the measure of the jth triangle, we define the discrete nonlinear objective functional by

$$\begin{aligned}\mathcal{F}_h(\kappa, \xi(1), \dots, \xi(L); \gamma) := {} & \frac{\sum_{l=1}^{L} \left\| \widetilde{B}_h(\theta_1)^{-1} B_h(\kappa, \theta_l)\xi(l) - \widetilde{B}_h(\theta_1)^{-1}\eta(\theta_l) \right\|^2_{\ell^2_{\mathbb{C}}(N)}}{\sum_{l=1}^{L} \left\| \widetilde{B}_h(\theta_1)^{-1}\eta(\theta_l) \right\|^2_{\ell^2_{\mathbb{C}}(N)}} \\ & + c_d \frac{\sum_{l=1}^{L} \left\| \left[F_h(\theta_l)\, \xi(l) + A^i(\theta_l)\right] - A_{meas}(\theta_l) \right\|^2_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))}}{\sum_{l=1}^{L} \|A_{meas}(\theta_l)\|^2_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))}} \\ & + c_v\gamma \|\kappa\|^2_{V(\Omega)} + c_s\gamma \sum_{l=1}^{L} \left\| \sum_{j=1}^{N} \xi(l)_j \varphi_{h,j} \right\|^2_{H^1_{per}(\Omega)}. \end{aligned} \tag{5.1}$$

Here c_d, c_v, and c_s denote appropriate equilibration constants which are to be determined by numerical experiments, and γ is the regularization parameter. Using the functional $\mathcal{F}_h$, we finally arrive at the discrete optimization problem

$$\begin{aligned} & \mathcal{F}_h(\kappa, \xi(1), \dots, \xi(L); \gamma) \quad \longrightarrow \min, \\ & \quad \text{with } \kappa \in \ell^2_{\mathbb{C}}(M), \\ & \xi(l) \in \ell^2_{\mathbb{C}}(N),\ l = 1, \dots, L. \end{aligned} \tag{5.2}$$

This will be solved using the following nonlinear conjugate gradient algorithm, which we prepare by giving formulae for the gradients. Note that the complex variables are treated as couples of real variables.

First we observe that the matrix-valued mapping $k^2 \mapsto [B_h(k, \theta_l) - B_h(0, \theta_l)] = (-\int_\Omega k^2 \varphi_{h,j}\overline{\varphi_{h,j'}})_{j',j}$ is linear and independent of θ_l. We easily get that

$$\nabla_\kappa B_h(\kappa, \theta_l) = \nabla_\kappa B_h = \left([\nabla_\kappa B_h]_j \right)_{j=1,\dots,M},$$

$$[\nabla_\kappa B_h]_j := \left(-\int_\Omega \varphi_{h,i}\overline{\varphi_{h,i'}}\chi_{h,j} \right)_{i',i=1,\dots,N},$$

$$\nabla_\kappa B_h\, \kappa' = \sum_{j=1}^{M} [\nabla_\kappa B_h]_j\, \kappa'_j.$$

If $\langle\cdot,\cdot\rangle_{\ell^2_{\mathbb{C}}(N)}$ stands for the scalar product in the N-dimensional Euclidean space, then the gradient of $\mathcal{F}_h$ is given by

$$\begin{aligned}
&\nabla\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)\left(\kappa',\xi(1)',\ldots,\xi(L)'\right)\\
&=\big(\nabla_\kappa\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)\,\kappa',\nabla_{\xi(1)}\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)\,\xi(1)',\\
&\quad\nabla_{\xi(2)}\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)\xi(2)',\ldots,\nabla_{\xi(L)}\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)\xi(L)'\big),
\end{aligned}$$

$$\begin{aligned}
&\nabla_\kappa\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)\kappa'\\
&=\Re\mathrm{e}\Big\langle 2\rho_1\sum_{l=1}^{L}\Big\langle\widetilde{B}_h(\theta_1)^{-1}\left[B_h(\kappa,\theta_l)\xi(l)-\eta(\theta_l)\right],\widetilde{B}_h(\theta_1)^{-1}\nabla_\kappa B_h\,\xi(l)\Big\rangle_{\ell^2_{\mathbb{C}}(N)},\kappa'\Big\rangle_{\ell^2_{\mathbb{C}}(M)}\\
&+2c_v\gamma\sum_j\Re\mathrm{e}\left[\kappa_j\overline{\kappa'_j}\right]\omega_j+2c_v\gamma\sum_{j,j^*}\Re\mathrm{e}\left[(\kappa_j-\kappa_{j^*})(\overline{\kappa'_j}-\overline{\kappa'_{j^*}})\right]\sqrt{\omega_j}\,,\\
&\quad\nabla_{\xi(l)}\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)\xi(l)'\\
&=\Re\mathrm{e}\big\langle 2\rho_1\Big[\widetilde{B}_h(\theta_1)^{-1}B_h(\kappa,\theta_l)\Big]^*\Big[\widetilde{B}_h(\theta_1)^{-1}B_h(\kappa,\theta_l)\xi(l)-\widetilde{B}_h(\theta_1)^{-1}\eta(\theta_l)\Big],\xi(l)'\big\rangle_{\ell^2_{\mathbb{C}}(N)}\\
&+\Re\mathrm{e}\big\langle 2\rho_2\,F_h(\theta_l)^*\left[\left[F_h(\theta_l)\,\xi(l)+A^i(\theta_l)\right]-A_{meas}(\theta_l)\right],\xi(l)'\big\rangle_{\ell^2_{\mathbb{C}}(N)}\\
&+\Re\mathrm{e}\left\langle 2\,c_s\gamma\left[\sum_{j=1}^{N}\xi(l)_j\varphi_{h,j}\right],\left[\sum_{j'=1}^{N}\xi(l)'_{j'}\varphi_{h,j'}\right]\right\rangle_{H^1_{per}(\Omega)},
\end{aligned}$$

$$\rho_1:=\frac{1}{\displaystyle\sum_{l=1}^{L}\left\|\widetilde{B}_h(\theta_1)^{-1}\eta(\theta_l)\right\|^2_{\ell^2_{\mathbb{C}}(N)}},\quad \rho_2:=\frac{c_d}{\displaystyle\sum_{l=1}^{L}\|A_{meas}(\theta_l)\|^2_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))}}\,.$$

Here $[\widetilde{B}_h(\theta_1)^{-1}B_h(\kappa,\theta_l)]^*$ is the adjoint (transposed and complex conjugate) of matrix $[\widetilde{B}_h(\theta_1)^{-1}B_h(\kappa,\theta_l)]$, and $F_h(\theta_l)^*$ that of $F_h(\theta_l)$. Treating the gradients as vectors, we arrive at

$$\begin{aligned}
&\nabla\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)\\
&\quad=\big(\nabla_\kappa\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right),\nabla_{\xi(1)}\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right),\\
&\qquad\nabla_{\xi(2)}\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right),\ldots,\nabla_{\xi(L)}\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)\big),\\
&\nabla_\kappa\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)=\Big(\nabla_\kappa\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)_j\Big)_{j=1,\ldots,M}\in\ell^2_{\mathbb{C}}(M),
\end{aligned}$$

$$\begin{aligned}
&\nabla_\kappa\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)_j:=2\rho_1\\
&\quad\times\sum_{l=1}^{L}\Big\langle\widetilde{B}_h(\theta_1)^{-1}B_h(\kappa,\theta_l)\xi(l)-\widetilde{B}_h(\theta_1)^{-1}\eta(\theta_l),\widetilde{B}_h(\theta_1)^{-1}\left[\nabla_\kappa B_h\right]_j\xi(l)\Big\rangle_{\ell^2_{\mathbb{C}}(N)}\\
&\qquad+2\,c_v\gamma\left\{w_j\kappa_j+\sqrt{w_j}\sum_{\substack{j':j,j'\text{ are}\\ \text{indices of}\\ \text{neighbors}}}\left[\kappa_j-\kappa_{j'}\right]-\sum_{\substack{j':j,j'\text{ are}\\ \text{indices of}\\ \text{neighbors}}}\sqrt{w_{j'}}\left[\kappa_j-\kappa_{j'}\right]\right\},
\end{aligned}$$

$$\nabla_{\xi(l)}\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)=\left(\nabla_{\xi(l)}\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)_j\right)_{j=1,\ldots,N}\in\ell^2_{\mathbb{C}}(N),$$

$$\begin{aligned}\nabla_{\xi(l)}\mathcal{F}_h\left(\kappa,\xi(1),\ldots,\xi(L);\gamma\right)_j:=\;&2\rho_1\\ &\times\left[\left[\widetilde{B}_h(\theta_1)^{-1}B_h(\kappa,\theta_l)\right]^*\left[\widetilde{B}_h(\theta_1)^{-1}B_h(\kappa,\theta_l)\xi(l)-\widetilde{B}_h(\theta_1)^{-1}\eta(\theta_l)\right]\right]_j\\ &+2\rho_2\left[F_h(\theta_l)^*\left[\left[F_h(\theta_l)\,\xi(l)+A^i(\theta_l)\right]-A_{meas}(\theta_l)\right]\right]_j\\ &+2\,c_s\gamma\left\langle\left[\sum_{j'=1}^{N}\xi(l)_{j'}\varphi_{h,j'}\right],\varphi_{h,j}\right\rangle_{H^1_{per}(\Omega)}.\end{aligned}$$

Now the nonlinear conjugate gradient algorithm of Fletcher–Reeves modified by Polak–Ribière (cf., e.g., [20]) takes the following form.

Conjugate Gradient Algorithm.

Given the constant $0<c_1=10^{-3}$*;*

Given the initial guess $(\kappa_0,\xi_0(1),\ldots,\xi_0(L))$*;*

Evaluate $\mathcal{F}_{h,0}:=\mathcal{F}_h\left(\kappa_0,\xi_0(1),\ldots,\xi_0(L);\gamma\right)$ *and the gradient* $\nabla\mathcal{F}_{h,0}:=\nabla\mathcal{F}_h\left(\kappa_0,\xi_0(1),\ldots,\xi_0(L);\gamma\right)$*;*

Set the first search direction $p_0:=\left(\kappa_0^d,\xi_0^d(1),\ldots,\xi_0^d(L)\right)=-\nabla\mathcal{F}_{h,0}$ *and set* $j=0$*;*

while $\nabla\mathcal{F}_{h,j}=\nabla\mathcal{F}_h\left(\kappa_j,\xi_j(1),\ldots,\xi_j(L);\gamma\right)\neq 0$

Compute step size α_j *of the correction* $\left(\alpha_j\kappa_j^d,\alpha_j\xi_j^d(1),\ldots,\alpha_j\xi_j^d(L)\right)$ *such that* α_j *is the largest number in* $\{256,128,64,32,16,\ldots\}$ *with*

$$\begin{aligned}&\mathcal{F}_h\left(\kappa_j+\alpha_j\kappa_j^d,\xi_j(1)+\alpha_j\xi_j^d(1),\ldots,\xi_j(L)+\alpha_j\xi_j^d(L);\gamma\right)\\ &\leq\mathcal{F}_h\left(\kappa_j,\xi_j(1),\ldots,\xi_j(L);\gamma\right)+c_1\alpha_j\nabla\mathcal{F}_{h,j}^T\left(\kappa_j^d,\xi_j^d(1),\ldots,\xi_j^d(L);\gamma\right);\end{aligned}$$

Set the new iterate solution $(\kappa_{j+1},\xi_{j+1}(1),\ldots,\xi_{j+1}(L))$ *to*

$$\left(\kappa_j+\alpha_j\kappa_j^d,\xi_j(1)+\alpha_j\xi_j^d(1),\ldots,\xi_j(L)+\alpha_j\xi_j^d(L)\right);$$

Evaluate gradient $\nabla\mathcal{F}_{h,j+1}:=\nabla\mathcal{F}_h\left(\kappa_{j+1},\xi_{j+1}(1),\ldots,\xi_{j+1}(L);\gamma\right)$*;*

Set $\beta_{j+1}=\max\left\{\dfrac{\nabla\mathcal{F}_{h,j+1}^T(\nabla\mathcal{F}_{h,j+1}-\nabla\mathcal{F}_{h,j})}{\|\nabla\mathcal{F}_{h,j}\|^2},0\right\}$*;*

Set new search direction $p_{j+1}=\left(\kappa_{j+1}^d,\xi_{j+1}^d(1),\ldots,\xi_{j+1}^d(L)\right)$ *to*

$$p_{j+1}:=-\nabla\mathcal{F}_{h,j+1}+\beta_{j+1}\left(\kappa_j^d,\xi_j^d(1),\ldots,\xi_j^d(L)\right);$$

Set $j=j+1$

end(while)

The line search part, i.e., the determination of α_j, can be improved. In fact, instead of changing α_j to half its value, we can take the argument of the minimum of a quadratic interpolation to $\alpha_j\mapsto\mathcal{F}_h(\kappa_j+\alpha_j\kappa_j^d,\xi_j(1)+\alpha_j\xi_j^d(1),\ldots,\xi_j(L)+\alpha_j\xi_j^d(L);\gamma)$ as the next value for α_j.

Usually, this conjugate gradient method converges to a local minimum of the objective function $\mathcal{F}_h$. The determination of the global minimum for high-dimensional optimization is a difficult and expensive problem. Note that a high number of degrees of freedom is required for the finite element method in order to resolve the oscillations of the Helmholtz equation. Even if a fast method for the computation of the global minimum were available, we would have to be careful. Indeed, the global solution of the optimization problem with regularization parameter γ set to zero might be

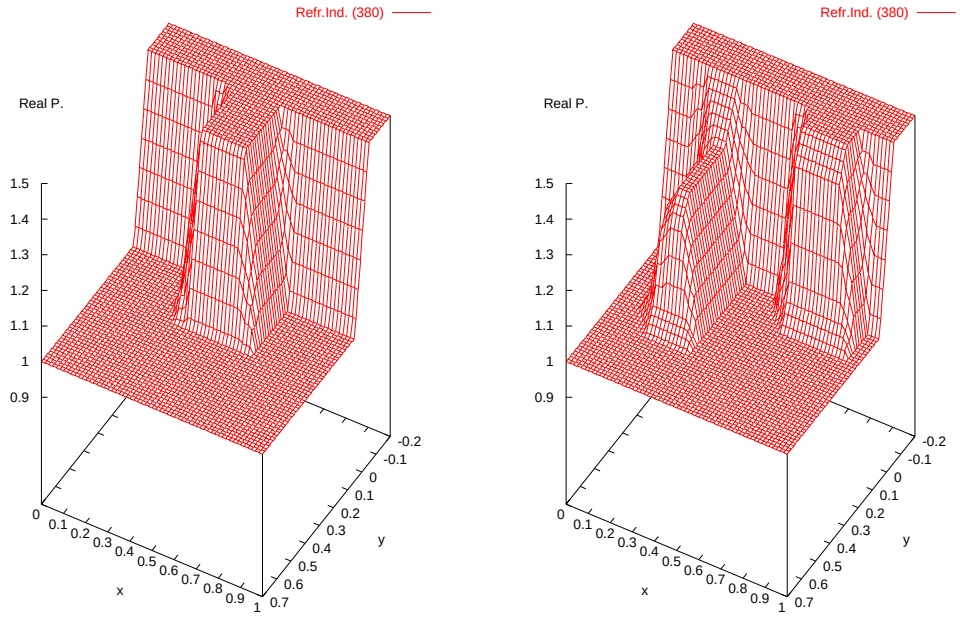


FIG. 6.1. *The two gratings: Rectangular and two towers.*

close to a local minimum of the regularized problem ($\gamma > 0$) different from the global minimum. In any case, due to the locality of the conjugate gradient solution, the choice of the initial solution is very important. Fortunately, for our numerical experiments, the choice of the initial guess as the mean value wave number function and the corresponding solutions of the Helmholtz solutions, i.e.,

$$\kappa_{0,j} := k_0^2,\ j = 1, \dots, M, \quad k_0 := \frac{k^- + k^+}{2},$$
$$\xi_0(l) := [B_h(k_0, \theta_l)]^{-1} \eta(\theta_l),\ l = 1, \dots, L,$$

was satisfactory.

6. The numerical experiment. The conjugate gradient approach. For our numerical tests we consider two gratings. Both are chosen with $b^- = -0.2\,\mu$m, $b_1^- = 0\,\mu$m, $b_1^+ = 0.5\,\mu$m, and $b^+ = 0.7\,\mu$m. The period is $d = 1\,\mu$m. The grating materials are characterized by the refractive index ν, which determines the value of the wave number function by the formula $k = \nu d/\lambda$. The wave length of light is $\lambda = 635\,$nm. The cover material over the grating (for $x_2 > b_1^+$) is air with $\nu = 1$. The index of the substrate material (for $x_2 < b_1^-$) is $\nu = 1.5$. The first grating is rectangular (cf. Figure 6.1, where a continuous linear interpolation of the piecewise constant function is plotted), i.e., the refractive index is

$$\nu = \nu(x_1, x_2) := \begin{cases} 1.5 & \text{for } \left|x_1 - \frac{1}{2}\right| \le \frac{1}{6} \text{ and } x_2 \le \frac{1}{4}, \\ 1.0 & \text{for } \left|x_1 - \frac{1}{2}\right| > \frac{1}{6} \text{ or } x_2 > \frac{1}{4}\,. \end{cases}$$

The second (cf. Figure 6.1) is a two tower grating with

$$\nu = \nu(x_1, x_2) := \begin{cases} 1.5 & \text{for } \left|x_1 - \frac{3}{4}\right| \le \frac{1}{6} \text{ and } x_2 \le \frac{1}{8}, \\ 1.35 & \text{for } \left|x_1 - \frac{1}{4}\right| \le \frac{1}{6} \text{ and } x_2 \le \frac{3}{8}, \\ 1.5 & \text{for } x_2 < 0, \\ 1.0 & \text{else.} \end{cases}$$

For the two gratings, we have computed the Rayleigh coefficients corresponding to nonevanescent modes under the angles of incidence θ_l, $l = 1, \dots L$.

$$\{\theta_l : \ l = 1, \dots L\} := \begin{cases} \{0\} & \text{if } L = 1, \\ \{-60, 0, 60\} & \text{if } L = 3, \\ \{-60, -40, -20, 0, 20, 40, 60\} & \text{if } L = 7, \\ \{-60, -55, -50, -45, \dots, 45, 50, 55, 60\} & \text{if } L = 25. \end{cases}$$

Depending on the angle of incidence, these Rayleigh coefficients are three numbers of the A_n^+, $n = -2, -1, 0, 1, 2$, and five numbers of the A_n^-, $n = -3, -2, -1, 0, 1, 2, 3$. From these numbers we have to recover the grating by the inverse algorithm described in section 5.

Since our simulated measurement data should be obtained by a method different from that involved in the inverse algorithm, we have computed the $A_n^\pm$ by a finite element method over a high level nonuniform triangulation. Actually we have employed a standard grid generator and more than 200 000 unknown finite elements. The finite element operator $B_h(k, \theta)$ used for the algorithm of section 5 is based on a coarse uniform triangulation. More precisely, we split the domain $\Omega = (0, 1) \times (-0.2, 0.7)$ into 40×36 equal squares and divide each square into two triangles by a cut along the diagonal. Taking into account the periodicity, the resulting number of finite elements is 1600. The unknown wave number function is sought as a function piecewise continuous over the triangulation resulting from halving the squares of a 20×18 uniform rectangular partition. This means there are 720 triangles in Ω and exactly 400 unknowns for the wave number function corresponding to the triangles falling into the strip $(0, d) \times (b_1^-, b_1^+) = (0, 1) \times (0, 0.5)$.

The constants c_s, c_d, c_v, and γ have to be adapted to the special case at hand. So one should take a typical example with known wave number solution and determine the constants such that the resulting approximation of the wave number function is the closest to the known exact solution. Then the unknown gratings should be recovered using the just obtained constants. Note that, in general, the determination of the regularization parameter is a hard problem. It is probably easiest to adjust it for a simple case with known solution and to reuse it for the general case.

Following this philosophy, we have determined the "optimal" constants for the first rectangular grating. We have set $c_d = 0.005$, and the other numbers, including the number of necessary conjugate gradient iterations, are given in Table 6.1. More precisely, we have chosen the equilibration parameter c_d such that the first term in the functional (5.1) is about the same size as the second (cf. the choice of the analogous parameter γ following (5.56) in [11]). The equilibrium parameters c_v and c_d have been optimized together with the regularization parameter γ. In a series of calculations we have doubled and halved, independently, the values $c_v\gamma$ and $c_s\gamma$ and found the "optimal" values presented in Table 6.1. Here "optimality" means that the approximate values of the refractive index at the points (0.0083,0.1417), (0.5167,0.1583), and (0.8167,0.2583) are closest to the exact values 1., 1.5, and 1.,

TABLE 6.1
Constants for the objective functional.

L	$c_v\gamma$	$c_s\gamma$	Iterations
1	0.000 000 9	0.000 001 5	2 000
3	0.000 000 5	0.000 000 25	8 000
7	0.000 000 03	0.000 000 005	25 000
25	0.000 000 2	0.000 000 000 05	50 000

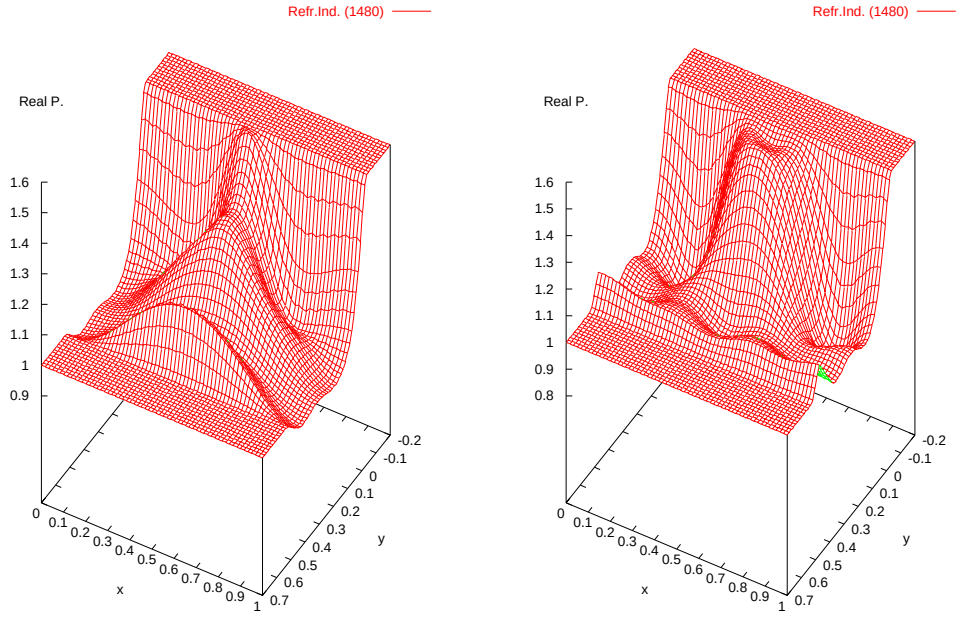


FIG. 6.2. *Reconstructed rectangular grating.* $L = 1$ *and* $L = 3$. *Conjugate gradient method.*

respectively. The plots of the resulting reconstructed wave number functions are shown in Figures 6.2 and 6.3. With larger L, i.e., with more measurement data, the recovered wave number improves slightly.

Next we have taken the optimal parameters of the rectangular grating and employed them in the algorithm for the two towers grating. Figures 6.4 and 6.5 show the results of the reconstruction which are close to the exact function (cf. Figure 6.1).

The SQP approach. Clearly, the conjugate gradient algorithm for the optimization problem (5.2) can be replaced by different optimization methods. For example, we consider the implementation SNOPT 5.3-4 of the SQP method [17]. Since this method is capable of dealing with constraints, we define

$$\mathcal{F}_h^{sqp}(\kappa, \xi(1), \dots, \xi(L); \gamma) := c_d \frac{\displaystyle\sum_{l=1}^{L} \left\| \left[F_h(\theta_l)\, \xi(l) + A^i(\theta_l) \right] - A_{meas}(\theta_l) \right\|^2_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))}}{\displaystyle\sum_{l=1}^{L} \|A_{meas}(\theta_l)\|^2_{\ell^2_{\mathbb{C}}(\mathcal{U}^*(\theta_l))}} + c_v\gamma \|\kappa\|^2_{V(\Omega)} + c_s\gamma \sum_{l=1}^{L} \left\| \sum_{j=1}^{N} \xi(l)_j \varphi_{h,j} \right\|^2_{H^1_{per}(\Omega)} \tag{6.1}$$

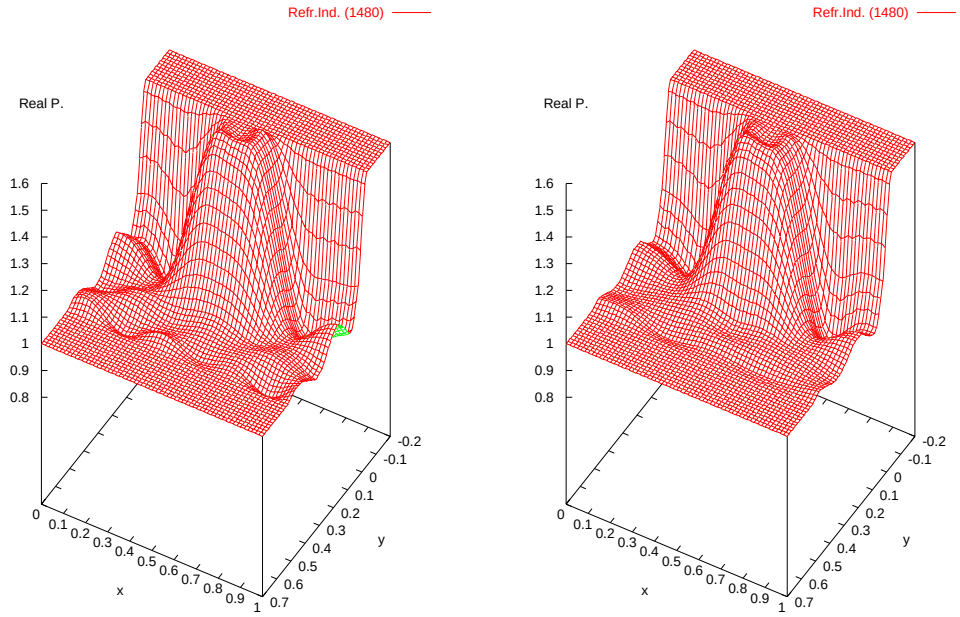


FIG. 6.3. *Reconstructed rectangular grating.* $L = 7$ *and* $L = 25$. *Conjugate gradient method.*

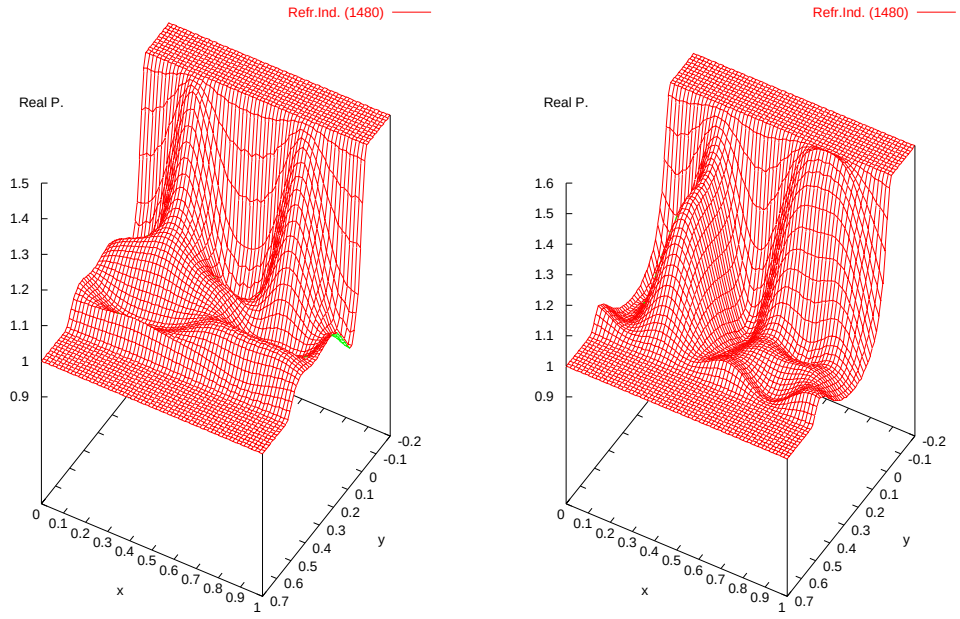


FIG. 6.4. *Reconstructed two towers.* $L = 1$ *and* $L = 3$. *Conjugate gradient method.*

and solve the optimization problem with constraints

$$\begin{aligned} \mathcal{F}_h^{sqp}(\kappa, \xi(1), \dots, \xi(L); \gamma) \quad &\longrightarrow \min, \\ \text{with } \kappa \in \ell^2_{\mathbb{C}}(M), & \\ \xi(l) \in \ell^2_{\mathbb{C}}(N),\ l = 1, \dots, L, & \\ B_h(\kappa, \theta_l)\xi(l) = \eta(\theta_l). & \end{aligned} \tag{6.2}$$

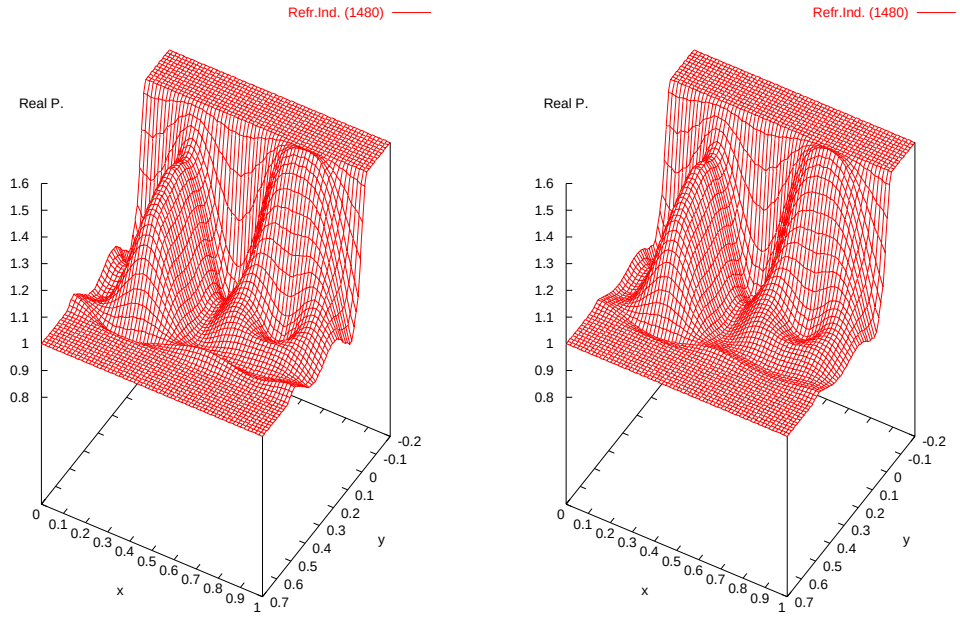


FIG. 6.5. *Reconstructed two towers.* $L = 7$ *and* $L = 25$. *Conjugate gradient method.*

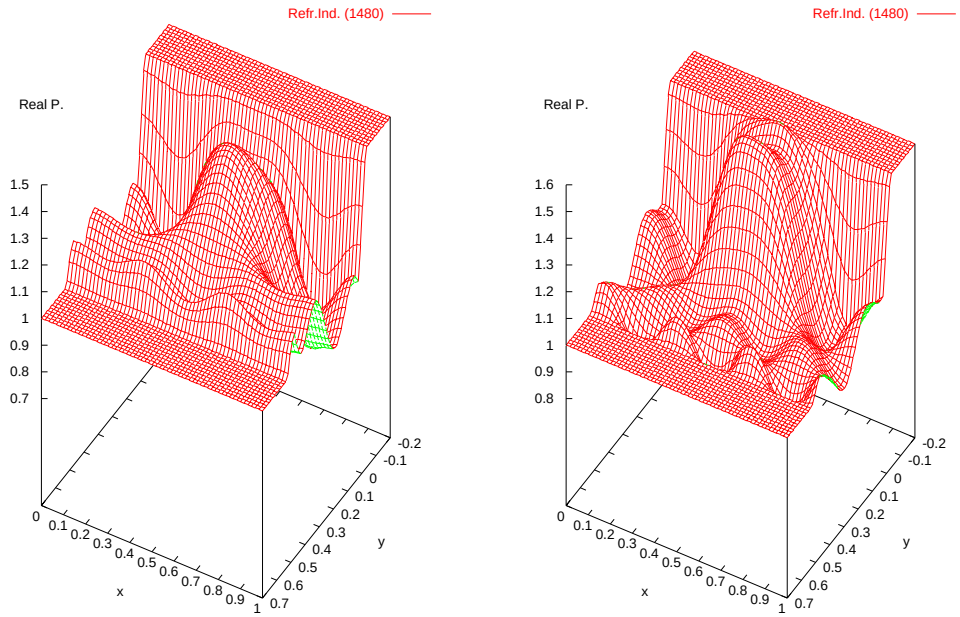


FIG. 6.6. *Reconstructed rectangular grating.* $L = 3$ *and* $L = 7$. *SQP method.*

Taking $c_d = 0.005$ and the parameters from Table 6.1, we arrive at visually the same pictures of reconstructed gratings (cf. Figures 6.6 and 6.7 and compare with Figures 6.1–6.5). Due to the extra effort for solving the constraint equations $B_h(\kappa, \theta_l)\xi(l) = \eta(\theta_l)$ exactly, the SQP method is much slower. Moreover, the SQP algorithm requires more storage capacity. However, to be fair, we have to admit that our simple test

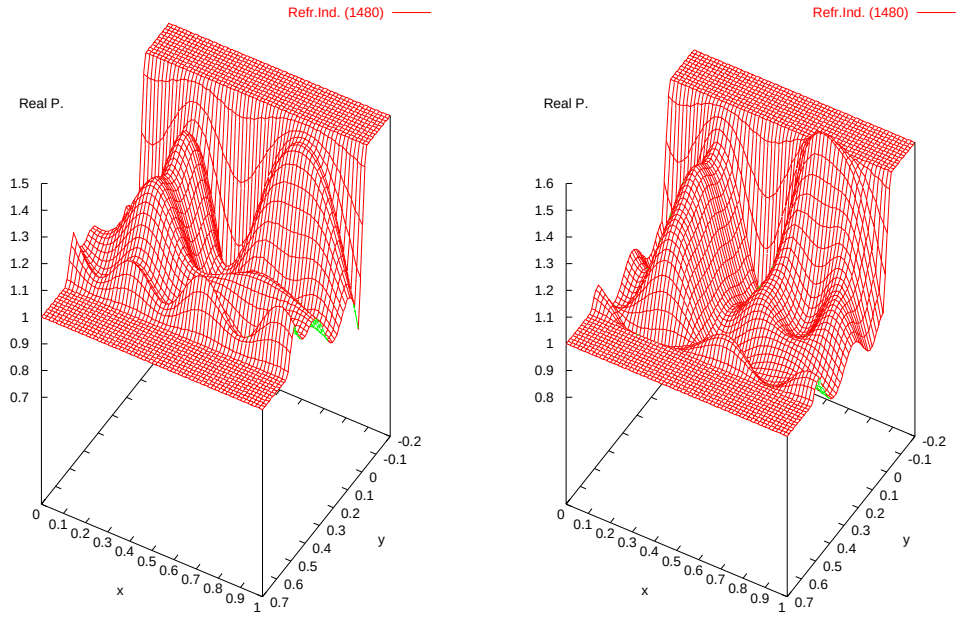


FIG. 6.7. *Reconstructed two towers. $L = 3$ and $L = 7$. SQP method.*

is performed employing a general code and default parameters. Changing the SQP parameters and including preconditioners might improve the performance.

Acknowledgments. The authors thank A. Möller and R. Henrion for many helpful consultations concerning the implementation of the optimization algorithms.

REFERENCES

[1] Y. ACHDOU AND O. PIRONNEAU, *Optimization of a photocell*, Optimal Control Appl. Methods, 12 (1991), pp. 221–246.

[2] T.S. ANGELL, G.C. HSIAO, AND L. WEN, *On the two-dimensional inverse scattering problems in electromagnetics*, Appl. Anal., 82 (2003), pp. 483–497 .

[3] T. ARENS AND A. KIRSCH, *The factorization method in inverse scattering from periodic structures*, Inverse Problems, 19 (2003), pp. 1195–1211.

[4] G. BAO, *A uniqueness theorem for an inverse problem in periodic diffractive optics*, Inverse Problems, 10 (1994), pp. 335–340.

[5] G. BAO, D.C. DOBSON, AND J.A. COX, *Mathematical studies in rigorous grating theory*, J. Opt. Soc. Amer. A, 12 (1995), pp. 1029–1042.

[6] A.-S. BONNET-BENDHIA AND F. STARLING, *Guided waves by electromagnetic gratings and non–uniqueness examples for the diffraction problem*, Math. Methods Appl. Sci., 17 (1994), pp. 305–338.

[7] G. BRUCKNER AND J. ELSCHNER, *A two-step algorithm for the reconstruction of perfectly reflecting periodic profiles*, Inverse Problems, 19 (2003), pp. 315–329.

[8] G. BRUCKNER, J. ELSCHNER, AND M. YAMAMOTO, *An optimization method for grating profile reconstruction*, WIAS preprint 682, Weierstrass Institute for Applied Analysis and Stochastics, Berlin, 2002; Proceedings of the 3rd Congress of the International Society for Analysis, Applications and Computation (ISAAC 2001), to appear.

[9] J. CHANDEZON, A.YE. POYEDINCHUK, AND N.P. YASHINA, *Reconstruction of periodic boundary between dielectric media*, in Proceedings of the 9th International Conference on Mathematical Methods in Electromagnetic Theory, Kiev, Ukraine, Vol. 2, Kontrast Publishing, Kharkov, Ukraine, 2002, pp. 416–419.

[10] P. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, New York, 1978.

[11] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Applied Math. Sci. 93, Springer-Verlag, Berlin, Heidelberg, 1998.

[12] J. Elschner and G. Schmidt, *Diffraction in periodic structures and optimal design of binary gratings.* I. *Direct problems and gradient formulas*, Math. Methods Appl. Sci., 21 (1998), pp. 1297–1342.

[13] J. Elschner and G. Schmidt, *Inverse scattering for periodic structures: Stability of polygonal interfaces*, Inverse Problems, 17 (2001), pp. 1817–1829.

[14] K. Ito and F. Reitich, *A high-order perturbation approach to profile reconstruction:* I. *Perfectly conducting gratings*, Inverse Problems, 15 (1999), pp. 1067–1085.

[15] A. Kirsch, *An inverse scattering problem for periodic structures*, in Inverse Scattering and Potential Problems in Mathematical Physics, Methoden Verfahren Math. Phys. 40, R.E. Kleinman, R. Kress, and E. Martensen, eds., Lang, Frankfurt am Main, 1995, pp. 75–93.

[16] A. Kirsch, *Uniqueness theorems in inverse scattering theory for periodic structures*, Inverse Problems, 10 (1994), pp. 145–152.

[17] P.E. Gill, W. Murray, and M.A. Saunders, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM J. Optim., 12 (2002), pp. 979–1006.

[18] F. Hettlich, *Iterative regularization schemes in inverse scattering by periodic structures*, Inverse Problems, 18 (2002), pp. 701–718.

[19] F. Hettlich and A. Kirsch, *Schiffer's theorem in inverse scattering theory for periodic structures*, Inverse Problems, 13 (1997), pp. 351–361.

[20] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.

[21] R. Petit, ed., *Electromagnetic Theory of Gratings*, Springer-Verlag, Berlin, Heidelberg, New York, 1980.

[22] J. Turunen and F. Wyrowski, eds., *Diffractive Optics for Industrial and Commercial Applications*, Akademie-Verlag, Berlin, 1997.

[23] H.P. Urbach, *Convergence of the Galerkin method for two-dimensional electromagnetic problems*, SIAM J. Numer. Anal., 28 (1991), pp. 697–710.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 546–564

ON THE NATURE OF INITIAL-BOUNDARY VALUE SOLUTIONS FOR DISPERSIVE EQUATIONS*

NATASHA FLYER† AND BENGT FORNBERG‡

Abstract. If the initial and boundary data for a partial differential equation (PDE) do not obey an infinite set of compatibility conditions, singularities will arise in its solutions. For dissipative equations, these singularities are well localized in both time and space, and an effective numerical remedy is available for accurate computation of initial transients. This study analyzes the nature of similar corner discrepancies for dispersive equations, such as $u_t - u_{xxx} = 0$ and $iu_t - u_{xx} = 0$.

Key words. time-space corner singularities, dispersive equations, initial-boundary value problems

AMS subject classifications. 35B05, 35B30, 35B65, 35A20

DOI. 10.1137/S0036139902415853

1. Introduction.

Solutions to initial-boundary value problems (IBVPs) will feature "corner singularities" in the space-time domain where initial and boundary data meet, unless these two data sets are connected by an infinite number of compatibility conditions [2]. Since the two data sets usually arise from different considerations, these singularities are almost always present. Although the issue has been analyzed at least since the 1950s (as surveyed in [2] and [3]), the focus has mostly been theoretical rather than numerical. For dissipative equations, the irregularities that are caused by these corner singularities are short-lived in time and remain local in space. These features allowed for the development of a highly effective strategy for restoring full numerical accuracy with little extra computational cost, as described in [4]. For dispersive PDEs, the irregularities do not stay local in space, and it depends on the equation whether or not they will be short-lived in time. Methods for effective numerical treatment will likely vary from equation to equation. The goal of the present paper is to give illustrating examples of dispersive corner singularities, largely by means of finding corner basis functions, which illuminate the mixing of temporal and/or spatial scales that occurs initially. If the boundaries are introduced to the problem only for the purpose of truncating what otherwise would have been an infinite domain, the preferred strategy would quite certainly be to create artificial boundary conditions in such a way that these space-time domain corner singularities do not arise.

Section 2 introduces the concept of corner basis functions, first for the well-known model equations $u_t + u_x = 0$ and $u_t - u_{xx} = 0$. It is shown how corner basis functions describe the nature of the corner singularities for these equations. Since the general character of solutions to dispersive equations may be less familiar, section 3 starts with some illustrative solutions for the linearized KdV equation, $u_t - u_{xxx} = 0$, and then proceeds with establishing its corner basis functions. Section 4 contains a similar discussion for the linear Schrödinger equation, $iu_t - u_{xx} = 0$. Based on these corner basis functions, we discuss in section 5 the character of IBV solutions for the two

*Received by the editors October 4, 2002; accepted for publication (in revised form) June 3, 2003; published electronically December 31, 2003. This work was supported by NSF grants DMS-9810751 (VIGRE), DMS-0073048, and DMS-0309803.
http://www.siam.org/journals/siap/64-2/41585.html
†National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80305 (flyer@ucar.edu).
‡University of Colorado, Department of Applied Mathematics, 526 UCB, Boulder, CO 80309 (fornberg@colorado.edu).

dispersive equations just mentioned. The final section offers some concluding remarks, summarizing the nature of corner singularities for PDEs of the type $u_t \pm u_{nx} = 0$, $n = 1, 2, 3, \ldots$, in terms of corner basis functions.

2. Corner basis functions for $u_t + u_x = 0$ and $u_t - u_{xx} = 0$. The quarter plane problem ($x > 0,\ t > 0$)

$$\begin{aligned} &\text{PDE:} && u_t + u_x = 0, \\ &\text{IC:} && u(x,0) = f(x), \\ &\text{BC:} && u(0,t) = g(t) \end{aligned} \tag{2.1}$$

has the analytic solution

$$u(x,t) = \begin{cases} f(x-t), & x > t, \\ g(t-x), & x < t. \end{cases} \tag{2.2}$$

Assuming that $f(x)$ and $g(t)$ are smooth functions, the solution (2.2) is smooth for all times if and only if an infinite sequence of compatibility conditions holds in the corner at $x = 0$, $t = 0$ [2], [3]:

$$\begin{aligned} &g(0) - f(0) && = 0, && \text{continuity} \\ &g_t(0) + f_x(0) && = 0, && \text{PDE} \\ &g_{tt}(0) - f_{xx}(0) && = 0, && \text{differentiated versions of the PDE} \\ &g_{ttt}(0) + f_{xxx}(0) && = 0, && \quad\downarrow \\ &\vdots \\ &g_{nt}(0) + (-1)^{n+1} f_{nt}(0) && = 0. \end{aligned} \tag{2.3}$$

If we are given a problem (2.1) for which any of the equalities in (2.3) fail to hold, one strategy for transforming the problem to one with a smooth solution (better suited for most numerical methods) is to create an explicitly known function $s(x,t)$ which also satisfies the PDE and which possesses an identical corner singularity. To construct $s(x,t)$, we introduce the concept of corner basis functions $u_n(x,t)$ with the properties

(2.4) (i) $u_n(x,t)$ satisfies the PDE away from the corner,

(ii) $u_n(x,0) \equiv 0$,

(iii) $$\frac{1}{n!}\frac{\partial^j u_n(0,t)}{\partial t^j} = \begin{cases} 1 & \text{for } j = k, \\ 0 & \text{for } j \neq k. \end{cases}$$

The corner basis functions are derived by Taylor expanding the boundary condition (BC) in time and then for each term in the expansion solving the PDE with zero initial condition (IC), as shown below in the case of (2.1).

$$\begin{aligned} u_0(x,t) &= \begin{cases} 0, & x > t, \\ 1, & x < t, \end{cases} \\ u_1(x,t) &= \begin{cases} 0, & x > t, \\ t - x, & x < t, \end{cases} \\ u_2(x,t) &= \begin{cases} 0, & x > t, \\ (t-x)^2, & x < t, \end{cases} \\ &\cdots, \end{aligned} \tag{2.5}$$

i.e.,

$$u_n(x,t) = \begin{cases} 0, & x > t, \\ (t-x)^n, & x < t, \end{cases} \quad n = 0, 1, 2, \dots .$$

If the right-hand sides of (2.3) were not equal to 0, 0, 0, 0, ... but instead equal to $r_0,\ r_1,\ r_2,\ r_3,\ \dots$, the function

$$s(x,t) = r_0\, u_0(x,t) + \frac{r_1}{1!}\, u_1(x,t) + \frac{r_2}{2!}\, u_2(x,t) + \frac{r_3}{3!}\, u_3(x,t) + \cdots \tag{2.6}$$

would have exactly the same corner singularity as the solution $u(x,t)$. Standard numerical methods can then be applied to the difference function

$$v(x,t) = u(x,t) - s(x,t), \tag{2.7}$$

which is infinitely smooth and well suited for numerics (satisfying the same PDE and IC, and having known BCs). However, in practice, we are limited to machine precision and thus need to use only a finite number p of compatibility conditions, corresponding to a truncated version of (2.6). The difference in (2.7) will be of size $O(t^p)$, the first neglected term in the expansion $s(x,t)$. For small t, we can make this difference arbitrarily small by choosing p sufficiently large.

This idea of creating corner singularity functions $u_n(x,t),\ n = 0, 1, 2, \dots$, and then subtracting a combination of them is of no particular utility for (2.1) since the analytic solution (2.2) is almost as simple algebraically as are the corner functions (2.5). Furthermore, the corner irregularity will persist for all times. If (2.1) is generalized to variable coefficients, the singularity will travel along a curved characteristic path, and cancellation based solely on corner information is not feasible.

Turning to the heat equation, it may at first appear that corner corrections are not needed. Figures 2.1(a), (b) show the analytic solution to the IBVP

$$\begin{aligned} &\text{PDE:} \quad u_t - u_{xx} = 0, \\ &\text{IC:} \quad\ u(x,0) \quad = 0, && 0 \le x \le 1, \\ &\text{BCs:} \begin{cases} u(0,t) & = \sin 2\pi t, \\ u_x(1,t) & = 0, \end{cases} && t > 0, \end{aligned} \tag{2.8}$$

over $0 \le x \le 1,\ 0 \le t \le 1$ and $0 \le x \le 10^{-3},\ 0 \le t \le 10^{-3}$, respectively. No matter how much we zoom in on the area near the origin, the solution surface will graphically look indistinguishable from the one shown to the right (Figure 2.1(b)).

However, this apparent regularity of the solution near the origin is severely misleading. The seemingly smooth solution in fact features a sharp irregularity, as the plot over $0 \le x \le 1,\ 0 \le t \le 10^{-3}$ in Figure 2.2(a) reveals.

A 21-node numerical Chebyshev solution (implemented without grid clustering at the right boundary, cf. [6, section 5.1, Example 3], and using a fourth-order Runge–Kutta method in time) will feature errors of the order 10^{-4} near the origin during the first moments, due to the fact that the PDE is not satisfied in the corner $(0,0)$ by the solution. This numerical observation is theoretically proven in [3]. At later times, the error decreases to around 10^{-12}. In Figure 2.2(b), the numerical corner error has already decayed to around 10^{-6} by the first displayed time level.

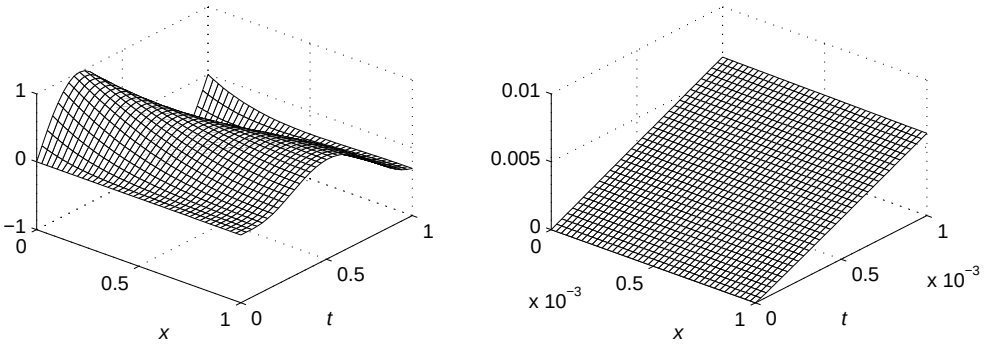


FIG. 2.1. *Analytic solution to the IBV problem* (2.8) *shown over* (a) $0 \le x \le 1,\ 0 \le t \le 1$ *and* (b) $0 \le x \le 10^{-3},\ 0 \le t \le 10^{-3}$.

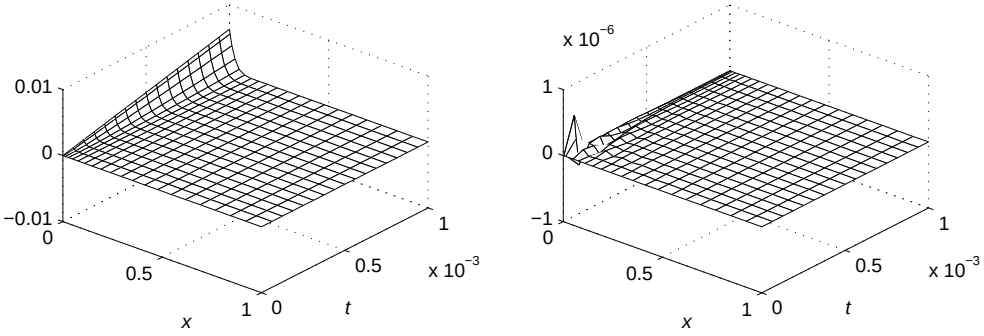


FIG. 2.2. (a) *Analytic solution to IBV problem* (2.8). (b) *Error in Chebyshev numerical solution. Both are shown over* $0 \le x \le 1,\ 0 \le t \le 10^{-3}$ *and displayed on a grid that is quadratically refined at the left edge.*

Although corner irregularities for the heat equation persist only a very short time, corrections for them are needed in order to obtain an accurate solution of initial transients. For the constant coefficient case

$$\begin{aligned} &\text{PDE:} && u_t - u_{xx} = 0, && \\ &\text{IC:} && u(x,0) = f(x), && x > 0, \\ &\text{BC:} && u(0,t) = g(t), && t > 0, \end{aligned} \tag{2.9}$$

we need to replace the compatibility conditions (2.3) by

$$\begin{aligned} g(0) - f(0) &= 0, && \text{continuity} \\ g_t(0) - f_{xx}(0) &= 0, && \text{PDE} \\ g_{tt}(0) - f_{xxxx}(0) &= 0, && \text{differentiated versions of the PDE} \\ \dots, & && \downarrow \\ g_{nt}(0) - f_{(2n)x}(0) &= 0, && \end{aligned}$$

and the corner functions (2.5) by

$$
\begin{aligned}
u_0(x,t) &= \operatorname{Erfc}\left(\tfrac{x}{2\sqrt{t}}\right),\\
u_1(x,t) &= -\sqrt{\tfrac{t}{\pi}}\; x\; e^{-x^2/(4t)} + \left(t+\tfrac{x^2}{2}\right)\operatorname{Erfc}\left(\tfrac{x}{2\sqrt{t}}\right),\\
u_2(x,t) &= -\tfrac{1}{6}\sqrt{\tfrac{t}{\pi}}\; x\; (10\,t+x^2)\; e^{-x^2/(4t)} + \left(t^2+t\,x^2+\tfrac{x^4}{12}\right)\operatorname{Erfc}\left(\tfrac{x}{2\sqrt{t}}\right),\\
u_3(x,t) &= -\tfrac{1}{60}\sqrt{\tfrac{t}{\pi}}\; x\; (132\,t^2+28\,t\,x^2+x^4)\; e^{-x^2/(4t)}\\
&\quad + \left(t^3+\tfrac{3}{2}t^2x^2+\tfrac{1}{4}t\,x^4+\tfrac{x^6}{120}\right)\operatorname{Erfc}\left(\tfrac{x}{2\sqrt{t}}\right),\\
&\ldots .
\end{aligned}
\tag{2.10}
$$

These functions all satisfy the PDE with the IC and BC $u_n(x,0)=0,\ u_n(0,t)=t^n,\ n=0,1,2,\ldots$.

One way to derive (2.10) is to note that the change of variables $\xi=\frac{x}{\sqrt{t}},\ \tau=\log t$ transforms $u_t-u_{xx}=0$ into $u_\tau=u_{\xi\xi}+\frac{\xi}{2}u_\xi$. The $u_0(x,t)$ solution corresponds to an equilibrium solution of the transformed PDE. With the BCs $u(0)=1,\ u(\infty)=0$ we find $u(\xi)=\operatorname{Erfc}(\frac{\xi}{2})$ $(=1-\frac{2}{\sqrt{\pi}}\int_0^{\xi/2}e^{-\varsigma^2}d\varsigma)$, and consequently $u_0(x,t)=\operatorname{Erfc}(\frac{x}{2\sqrt{t}})$. The subsequent corner functions can then (like for all other PDEs) be generated recursively:

$$
u_n(x,t) = n\int_0^t u_{n-1}(x,t)\,dt, \quad n=1,2,\ldots . \tag{2.11}
$$

Alternatively, we can obtain a general expression for all the $u_n(x,t)$ functions in terms of Kummer's confluent ${}_1F_1$ hypergeometric functions:

$$
u_n(x,t) = t^n\left\{{}_1F_1(-n,\tfrac{1}{2},-\tfrac{x^2}{4t}) - \tfrac{x\,\Gamma(n+1)}{\sqrt{t}\,\Gamma(n+\frac{1}{2})}\;{}_1F_1(\tfrac{1}{2}-n,\tfrac{3}{2},-\tfrac{x^2}{4t})\right\},\ n=0,1,2,\ldots . \tag{2.12}
$$

To arrive at (2.12), we generalize the observation above regarding the $u_0(x,t)$ corner function by noting that $\frac{u_n(x,t)}{t^n}$ becomes a function of one variable $\xi=\frac{x}{\sqrt{t}}$ only, which we write as $u_n(\xi)$. From its definition and the governing PDE, this function will need to satisfy

$$
(u_n)_{\xi\xi}+\frac{\xi}{2}(u_n)_\xi - n\,u_n = 0 \quad \text{with} \quad \begin{cases} u_n(0) &= 1,\\ u_n(\infty) &= 0. \end{cases}
$$

The general solution to the ODE can be written

$$
u_n(\xi) = c_1\;\; {}_1F_1(-n,\tfrac{1}{2},-\tfrac{\xi^2}{4}) + c_2\;\xi\;\; {}_1F_1(\tfrac{1}{2}-n,\tfrac{3}{2},-\tfrac{\xi^2}{4}).
$$

The condition $u_n(0)=1$ says that $c_1=1$, and leading order asymptotics of the ${}_1F_1$ functions (see [1]) demonstrate that cancellation of growths as $\xi\to\infty$ requires $\frac{c_2}{c_1}=-\frac{\Gamma(n+1)}{\Gamma(n+\frac{1}{2})}$.

Figure 2.3 shows $u_0(x,t),\ u_1(x,t)$, and $u_2(x,t)$ displayed over two different time intervals. The irregularity remains local in both time and space. For dissipative equations like (2.9), corner functions form a very effective means of improving the accuracy of numerical calculations since, as is shown in [4],

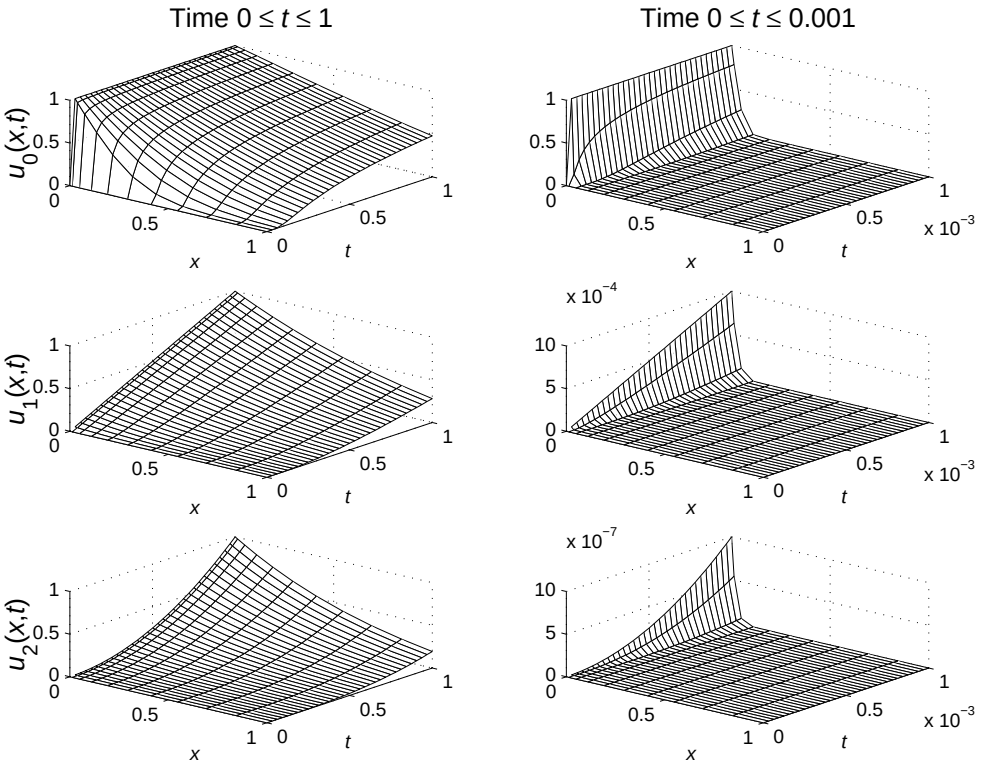


FIG. 2.3. *The corner functions* $u_0(x,t)$, $u_1(x,t)$, *and* $u_2(x,t)$ *for the heat equation* $u_t - u_{xx} = 0$, *shown over* $0 \le x \le 1$ *and (left column)* $0 \le t \le 1$, *(right column)* $0 \le t \le 0.001$. *The grid is again quadratically refined towards the left edge.*

1. only 3–4 correction functions typically suffice for correction to machine precision, and
2. generalizations to variable coefficients are straightforward.

3. Illustrative solutions and corner functions for $u_t - u_{xxx} = 0$. Similar to the heat equation, IBV solutions to the linearized KdV equation

$$u_t - u_{xxx} = 0 \tag{3.1}$$

will typically feature two scales: (1) a slow, long-term part and (2) a high-frequency part emanating from the corners and described by corner basis functions. Initially, the high-frequency part of the solution is of infinitesimal size but then expands to cover the whole domain. To illustrate the first part and to provide a background for discussing the latter part, we first consider different half-plane problems containing only slow long-term scales.

3.1. Traveling wave solutions in different half-planes.

3.1.1. The upper half-plane ($t > 0$). With the IC

$$\text{IC:} \quad u(x,0) = \cos(kx), \tag{3.2}$$

the solution of (3.1) becomes

$$\text{solution:} \quad u(x,t) = \cos(kx - k^3 t). \tag{3.3}$$

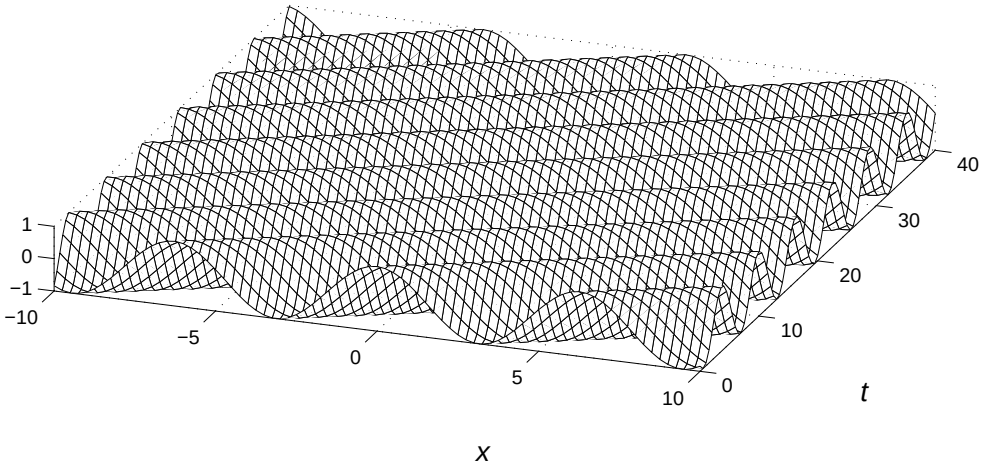


FIG. 3.1. *Upper half-plane solution to* $u_t - u_{xxx} = 0$ *with IC* $u(x,0) = \cos(x)$.

This is a single Fourier mode whose phase speed increases with the wave number k as $c = k^2$. We can note that *all* waves travel to the right, as shown in Figure 3.1 for $k = 1$.

3.1.2. The right half-plane ($x > 0$). For the equation (3.1) we need to impose two BCs on the left side (taken to be $x = 0$). Two cases can be noted. The first case has a sinusoidal forcing on the boundary, with the first derivative $u_x(0,t) = 0$.

Case 1.

$$\text{BCs:} \quad \begin{cases} u(0,t) = \sin(k^3 t), \\ u_x(0,t) = 0, \end{cases}$$

$$\text{solution:} \quad u(x,t) = \tfrac{1}{\sqrt{3}} \left[\cos(kx - k^3 t + \tfrac{\pi}{3}) - e^{-\frac{\sqrt{3}}{2}kx} \cos(\tfrac{1}{2}kx + k^3 t + \tfrac{\pi}{3}) \right].$$

In the second case, the solution is zero and it is the first derivative that has a sinusoidal forcing:

Case 2.

$$\text{BCs:} \quad \begin{cases} u(0,t) = 0, \\ u_x(0,t) = \sin(k^3 t), \end{cases}$$

$$\text{solution:} \quad u(x,t) = \tfrac{1}{\sqrt{3}\,k} \left[\cos(kx - k^3 t + \tfrac{\pi}{6}) - e^{-\frac{\sqrt{3}}{2}kx} \cos(\tfrac{1}{2}kx + k^3 t - \tfrac{\pi}{6}) \right].$$

Figures 3.2 and 3.3 show Cases 1 and 2, respectively. We notice in Figure 3.2 how the crests of the waves emerge perpendicularly to the left boundary in order to accommodate the zero first derivative BC. Similarly, Figure 3.3 shows how the waves again are deformed near the boundary, this time to accommodate the condition $u(0,t) = 0$. As these two cases demonstrate, forcing the left boundary with a Fourier mode will produce outgoing waves with the same wave number, differing between the two cases only in amplitude and phase shift. It is therefore possible to create a

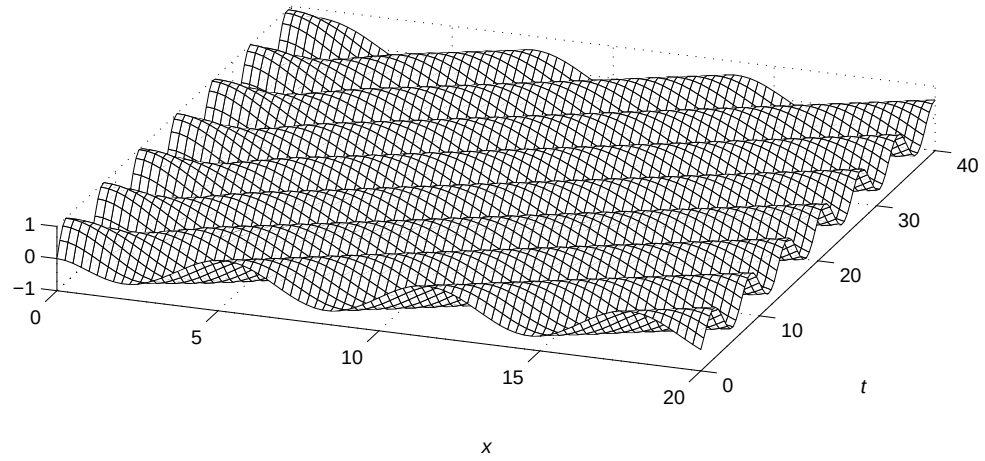


FIG. 3.2. *Solution to the right half-plane problem for* $u_t - u_{xxx} = 0$ *with the left BCs* $u(0,t) = \sin t,\ u_x(0,t) = 0$.

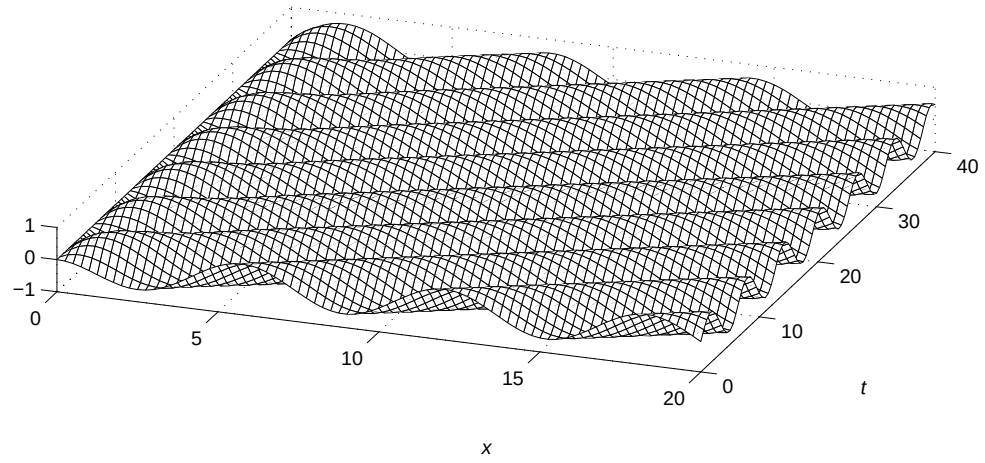


FIG. 3.3. *Solution to the right half-plane problem for* $u_t - u_{xxx} = 0$ *with left BCs* $u(0,t) = 0,\ u_x(0,t) = \sin t$.

BC so that the outgoing waves cancel, and only the exponential decay to the right remains. However, this is a very special case; sinusoidal forcing will in general produce sinusoidal waves traveling to the right.

3.1.3. The left half-plane ($x < 0$). For the right half-plane problems, we considered forcing on the left side. We now consider forcing on the right side, requiring only one BC for the PDE:

$$\begin{aligned} &\text{BC:} && u(0,t) = \sin(k^3 t), \\ &\text{solution:} && u(x,t) = e^{\frac{\sqrt{3}}{2}kx} \sin(\tfrac{1}{2}kx + k^3 t). \end{aligned}$$

There can be no waves traveling to the left for (3.1), and the solution therefore decays exponentially away from the boundary, as is seen in Figure 3.4. This also implies that

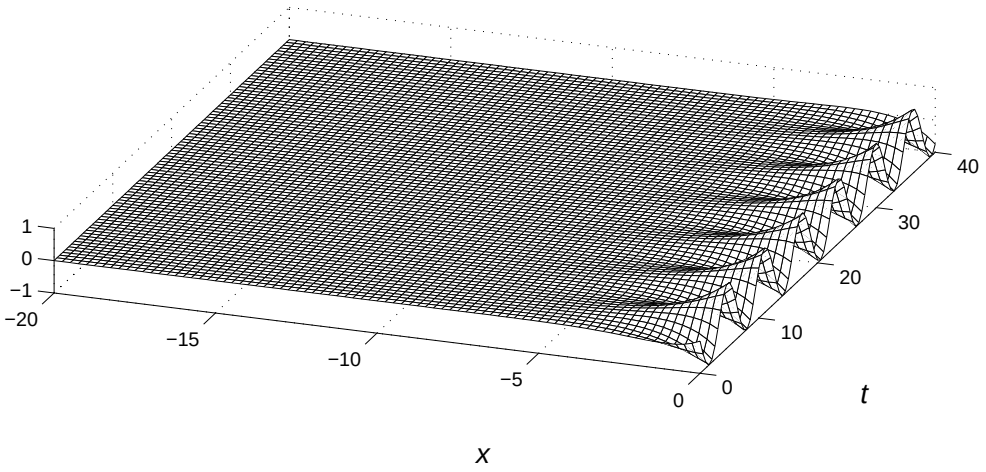


FIG. 3.4. *Solution to the left half-plane problem for* $u_t - u_{xxx} = 0$ *with forcing* $u(0,t) = \sin t$ *on the right boundary.*

when waves arrive from the left to a right boundary, there can be no reflection, but only decay. With incoming waves of the form $u(x,t) = \cos(kx - k^3 t)$, closed form solutions for the cases with Neumann and Dirichlet right BCs become as follows.

Case 1.

$$\begin{aligned} &\text{BC:} && u_x(0,t) = 0, \\ &\text{solution:} && u(x,t) = \cos(kx - k^3 t) + e^{\frac{\sqrt{3}}{2}kx} \cos(\tfrac{1}{2}kx + k^3 t + \tfrac{\pi}{3}). \end{aligned}$$

Case 2.

$$\begin{aligned} &\text{BC:} && u(0,t) = 0, \\ &\text{solution:} && u(x,t) = \cos(kx - k^3 t) - e^{\frac{\sqrt{3}}{2}kx} \cos(\tfrac{1}{2}kx + k^3 t). \end{aligned}$$

These solutions are shown in Figures 3.5 and 3.6.

In both cases, the incoming wave from the left undergoes a transition near the right boundary in order to accommodate the BCs at the right edge. The half-plane solutions for (3.1) can be summed up as waves traveling *solely* to the right, with at most a thin transition region at a right-hand boundary. The character of these half-plane solutions set the stage for solving the quarter-plane problem, leading us to sets of left and right corner basis functions.

3.2. Left corner functions. Since (3.1) needs two BCs to the left, we need to obtain two independent sets of corner functions $u_n(x,t)$ and $v_n(x,t)$. These functions should all obey the PDE, the IC $u_n(x,0) = v_n(x,0) = 0$, and the BCs

$$\begin{cases} u_n(0,t) = t^n, & \frac{\partial}{\partial x} u_n(0,t) = 0, \\ v_n(0,t) = 0, & \frac{\partial}{\partial x} v_n(0,t) = t^n, \end{cases} \qquad n = 0,\ 1,\ 2,\ \ldots .$$

Following the approach which led us to the corner functions (2.12) for the heat

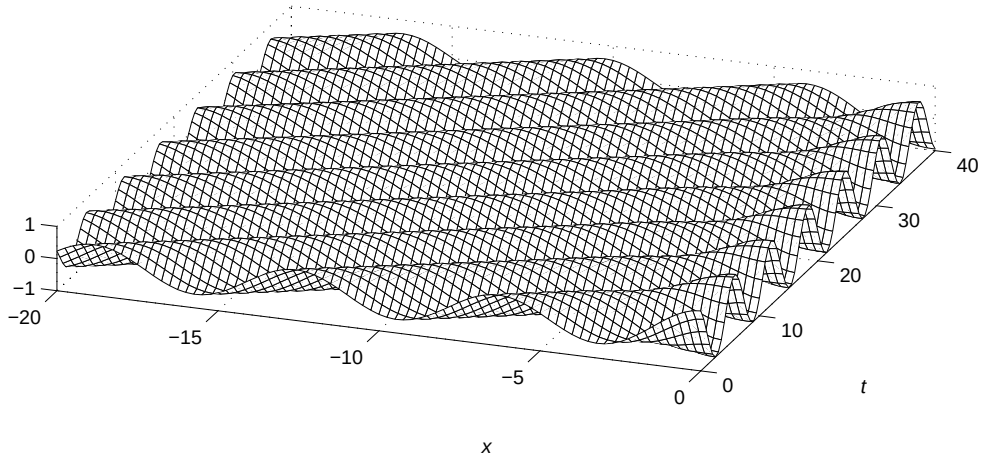


FIG. 3.5. *Solution to the left half-plane problem for* $u_t - u_{xxx} = 0$ *with incoming sinusoidal wave and right BC* $u_x(0,t) = 0$.

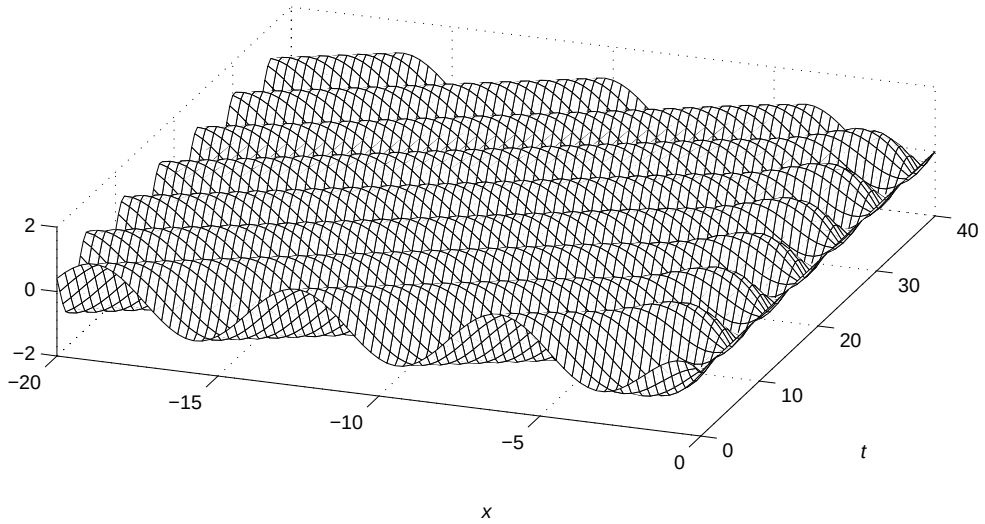


FIG. 3.6. *Solution to the left half-plane problem for* $u_t - u_{xxx} = 0$ *with incoming sinusoidal wave and right BC* $u(0,t) = 0$.

equation, we note that $\frac{u_n(x,t)}{t^n} = u_n(\xi)$ and $\frac{v_n(x,t)}{t^{n+1/3}} = v_n(\xi)$ both are functions of $\xi = \frac{x}{\sqrt[3]{t}}$ only, satisfying

$$(u_n)_{\xi\xi\xi} + \frac{\xi}{3}(u_n)_\xi - n\, u_n = 0, \;\; \{u_n(0) = 1,\ (u_n)_\xi(0) = 0,\ u_n(\infty) = 0\}$$

and

$$(v_n)_{\xi\xi\xi} + \frac{\xi}{3}(v_n)_\xi - (n + \tfrac{1}{3})\, v_n = 0, \;\; \{v_n(0) = 0,\ (v_n)_\xi(0) = 1,\ v_n(\infty) = 0\},$$

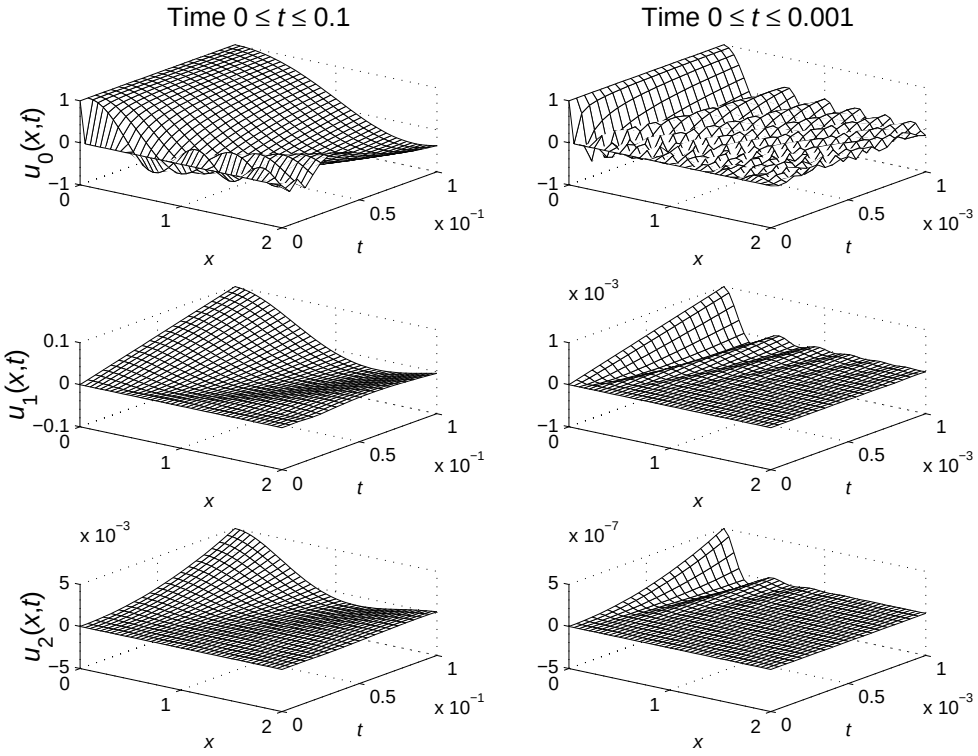


FIG. 3.7. *First three corner functions* $u_n(x,t)$ *for* $u_t - u_{xxx} = 0$, *displayed over* $0 \le t \le 0.1$ *and* $0 \le t \le 0.001$ *(left and right column, respectively).*

respectively, leading to the general expressions for the corner functions:

$$u_n(x,t) = t^n \left\{ {}_1F_2(-n, \{\tfrac{1}{3}, \tfrac{2}{3}\}, -\tfrac{x^3}{27t}) - \tfrac{x^2}{2t^{2/3}} \tfrac{\Gamma(n+1)}{\Gamma(n+\frac{1}{3})}\ {}_1F_2(\tfrac{2}{3} - n, \{\tfrac{4}{3}, \tfrac{5}{3}\}, -\tfrac{x^3}{27t}) \right\}, \quad n = 0, 1, 2, \dots,$$

and

$$v_n(x,t) = t^{n+\frac{1}{3}} \left\{ {}_1F_2(-n, \{\tfrac{2}{3}, \tfrac{4}{3}\}, -\tfrac{x^3}{27t}) - \tfrac{x^2}{2t^{2/3}} \tfrac{\Gamma(n+1)}{\Gamma(n+\frac{2}{3})}\ {}_1F_2(\tfrac{1}{3} - n, \{\tfrac{4}{3}, \tfrac{5}{3}\}, -\tfrac{x^3}{27t}) \right\}, \quad n = 0, 1, 2, \dots.$$

Figures 3.7 and 3.8 display the first three corner functions of each of the two types.

3.3. Right corner functions. Since the PDE is incapable of transporting any waves to the left, waves reaching a right side boundary will get absorbed no matter what BC is used there. As a consequence, the right corner functions on the domain $x < 0$, $t > 0$ will be nonoscillatory and reminiscent of the ones for the heat equation. Denoting these by $w_n(x,t)$, $n = 0, 1, 2, \dots$, we find by the same means as in the previous section

$$w_n(x,t) = t^n \left\{ {}_1F_2(-n, \{\tfrac{1}{3}, \tfrac{2}{3}\}, -\tfrac{x^3}{27t}) + \tfrac{x}{t^{1/3}} \tfrac{\Gamma(n+1)}{\Gamma(n+\frac{2}{3})}\ {}_1F_2(\tfrac{1}{3} - n, \{\tfrac{2}{3}, \tfrac{4}{3}\}, -\tfrac{x^3}{27t}) \right.$$
$$\left. + \tfrac{(-1)^n\sqrt{3}}{4\pi} \tfrac{x^2}{t^{2/3}} \Gamma(n+1)\ \Gamma(\tfrac{2}{3} - n)\ {}_1F_2(\tfrac{2}{3} - n, \{\tfrac{4}{3}, \tfrac{5}{3}\}, -\tfrac{x^3}{27t}) \right\},\ n = 0, 1, 2, \dots.$$

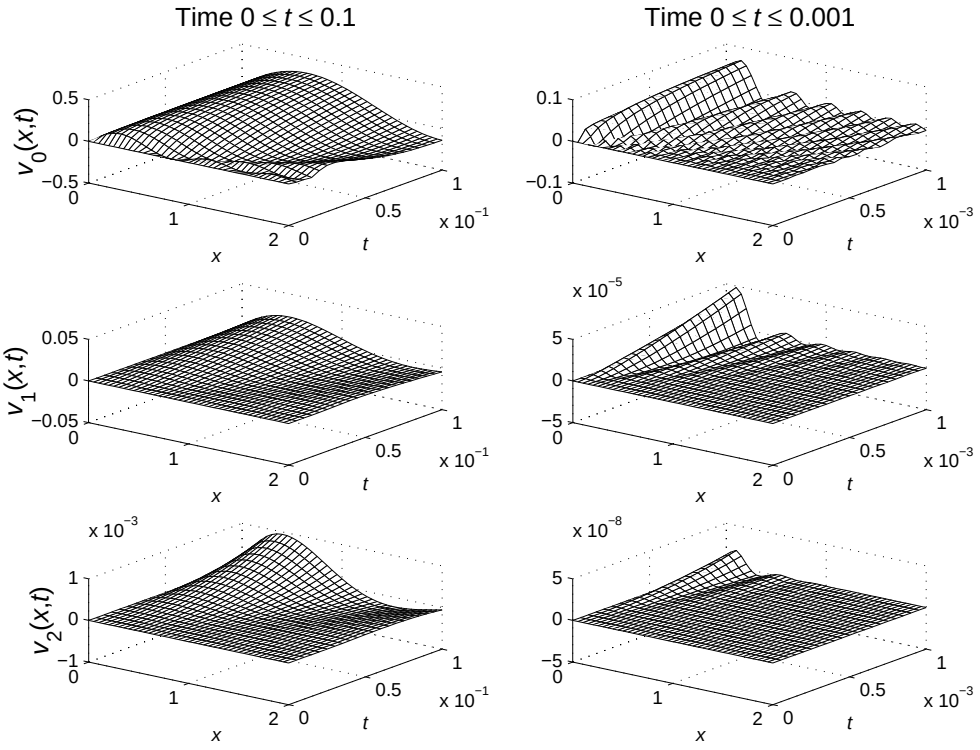


FIG. 3.8. *First three corner functions* $v_n(x,t)$ *for* $u_t - u_{xxx} = 0$, *displayed over* $0 \le t \le 0.1$ *and* $0 \le t \le 0.001$ *(left and right column, respectively).*

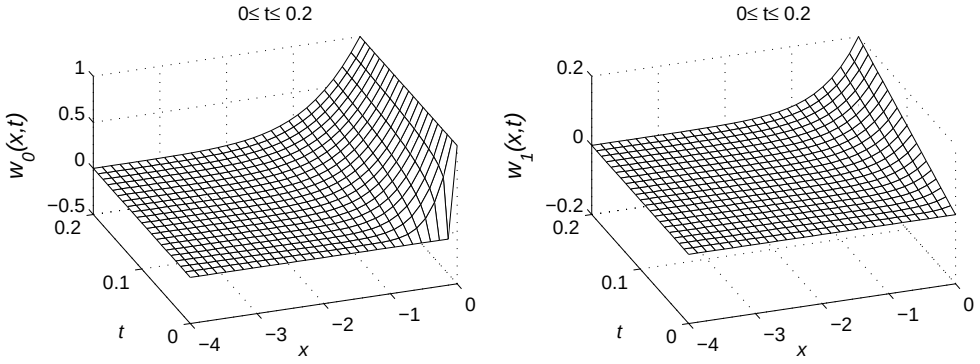


FIG. 3.9. *The right corner functions* $w_0(x,t)$ *and* $w_1(x,t)$ *to* $u_t - u_{xxx} = 0$.

Figure 3.9 displays the first two $w_n(x,t)$-functions.

For all the corner functions we have derived ($u_n(x,t)$ for the heat equation and $u_n(x,t)$, $v_n(x,t)$, $w_n(x,t)$ for the linearized KdV equation), the first hypergeometric function has $-n$ as its first parameter. This implies that, for $n = 0, 1, 2, \dots$, its usually infinite Taylor series truncates to become a finite degree polynomial.

To conclude this discussion of right corner functions, we note that a general

solution to (3.1) with

$$\begin{aligned} &\text{IC:} \quad u(x,0) = f(x), \quad x < 0, \\ &\text{BC:} \quad u(0,t) = g(t), \quad t > 0, \end{aligned}$$

can be expressed in terms of coupled contour principal value integrals [5].

4. Corner analysis for $iu_t - u_{xx} = 0$. The next example we consider is the linear Schrödinger equation

$$iu_t - u_{xx} = 0. \tag{4.1}$$

Although this also is a dispersive PDE, it will be shown that the character of IBV solutions for this equation is fundamentally different than for the linear KdV equation (3.1). Like the diffusion equation $u_t - u_{xx} = 0$, equation (4.1) requires only one BC on each side. Also, since the analysis is similar to that for the diffusion equation, we will here not consider any introductory half-plane problems.

4.1. Corner functions. Since $iu_t - u_{xx} = 0$ differs from the heat equation only by a factor of i, the same similarity transformation $\xi = \frac{x}{\sqrt{t}}$ and $\tau = \log t$ will again lead us to the corner functions. Substituting these transformations into $iu_t - u_{xx} = 0$ yields

$$iu_\tau - \frac{i\xi}{2} u_\xi - u_{\xi\xi} = 0. \tag{4.2}$$

The equilibrium solution satisfies

$$u_{\xi\xi} + \frac{i\xi}{2} u_\xi = 0,$$

leading to

$$u_0(x,t) = \operatorname{Erfc}\left(\sqrt{i}\frac{x}{2\sqrt{t}}\right) = \frac{1+i}{\sqrt{2\pi}} \int_{\frac{x}{\sqrt{t}}}^{\infty} e^{-i\frac{\eta^2}{4}} d\eta. \tag{4.3}$$

Separating (4.3) into real and imaginary parts results in

$$u_0(x,t) = 1 - S\left(\frac{x}{\sqrt{2\pi t}}\right) - C\left(\frac{x}{\sqrt{2\pi t}}\right) + i\left[S\left(\frac{x}{\sqrt{2\pi t}}\right) - C\left(\frac{x}{\sqrt{2\pi t}}\right)\right],$$

where S and C are the Fresnel sine, $\int_0^z \sin(\pi t^2/2)dt$, and cosine, $\int_0^z \cos(\pi t^2/2)dt$, functions. Higher-order corner functions are again most easily expressed in terms of hypergeometric functions. In analogy to (2.12), we obtain

$$u_n(x,t) = t^n \left\{ {}_1F_1(-n, \tfrac{1}{2}, \tfrac{-ix^2}{4t}) - \tfrac{x\sqrt{i}}{\sqrt{t}} \tfrac{\Gamma(n+1)}{\Gamma(n+\frac{1}{2})} {}_1F_1(\tfrac{1}{2} - n, \tfrac{3}{2}, \tfrac{-ix^2}{4t}) \right\}, \; n = 0, 1, \ldots,$$

which satisfies (4.1) with IC $u(x,0) = 0$ and the BCs $u(0,t) = t^n$, $u(\infty,t) = 0$.

The real and imaginary parts of the corner functions $u_0(x,t)$ and $u_1(x,t)$ are plotted in Figure 4.1.

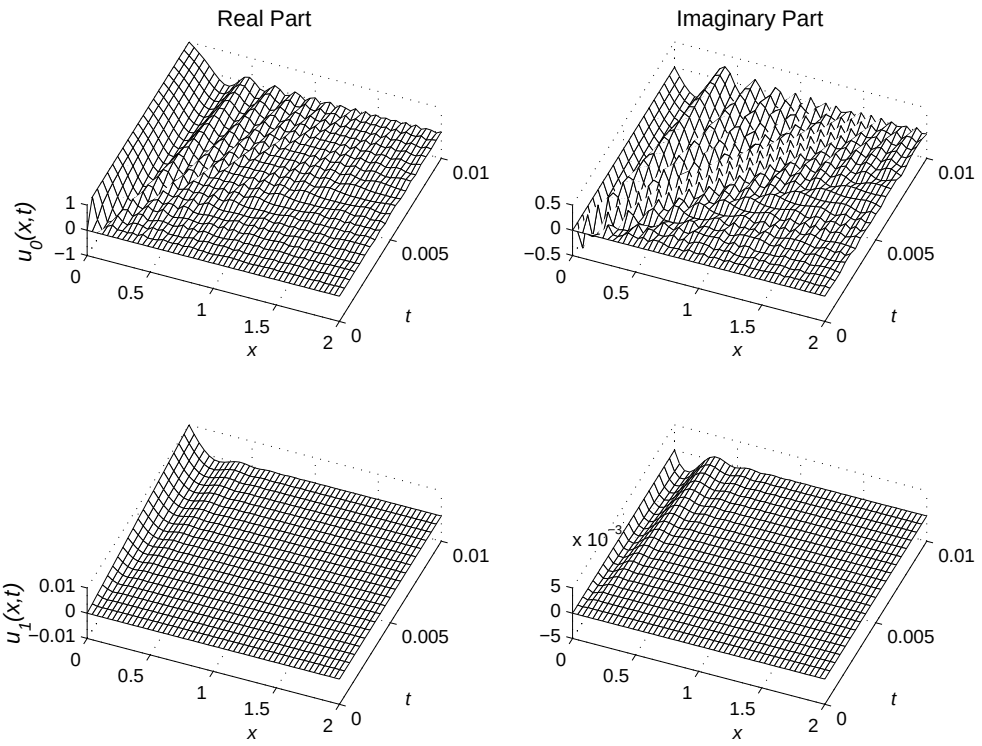


FIG. 4.1. *Real and imaginary part of the first two corner functions u_0 and u_1 to $iu_t - u_{xx} = 0$.*

5. Qualitative solution features in the case of two boundaries. In this section, the properties of $u_t - u_{xxx} = 0$ are contrasted with those of $iu_t - u_{xx} = 0$. In the former case ($u_t - u_{xxx} = 0$), high-frequency waves race across the interval and become absorbed at the opposite boundary. Like for the heat equation, the solutions are infinitely differentiable for all $t > 0$. In the latter case ($iu_t - u_{xx} = 0$), the waves are reflected off the boundaries for all times, resulting in a solution that is several times differentiable only for rare values of $t > 0$ (when recurrences to the IC happen to occur). This lack of smoothness has severe impact on the accuracy of straightforward numerical calculations.

5.1. Features of the solution to $u_t - u_{xxx} = 0$ in the case of two boundaries. Although the IBV problem

$$\begin{aligned} &\text{PDE:} && u_t - u_{xxx} = 0, && \\ &\text{IC:} && u(x,0) = 0, && 0 < x \le 1, \\ &\text{BCs:} && u(0,t) = f(t),\ u_x(0,t) = g(t),\ u(1,t) = 0, && t > 0, \end{aligned} \tag{5.1}$$

does not appear to admit a simple closed form solution for general functions $f(t)$ and $g(t)$, it can be verified that the function

$$u(x,t) = \frac{3}{2\pi} \int_0^\infty \frac{e^{\frac{1}{2}rx - r^3 t}}{r} \sin\left(\frac{\sqrt{3}}{2} r(x-1)\right) \left(e^{-3r/2} + 2\cos\left(\frac{\sqrt{3}}{2} r\right)\right) dr \tag{5.2}$$

satisfies it for some particular choice of $f(t)$ and $g(t)$. We note the following:

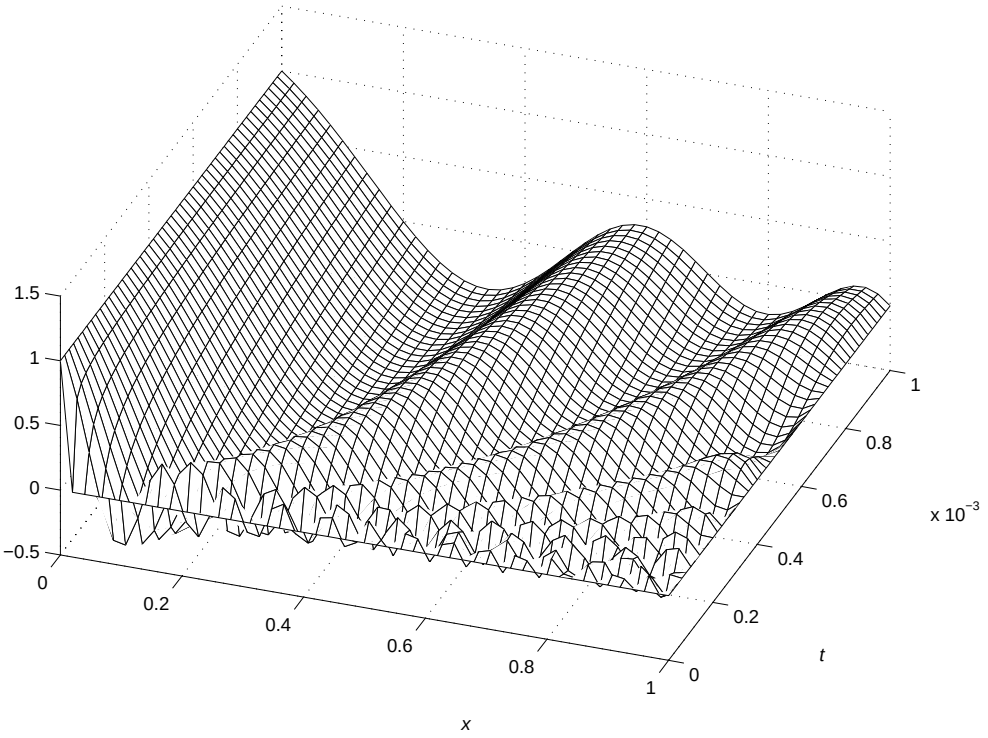


FIG. 5.1. *Solution* $u(x,t)$ (5.2) *to the IBV problem for* $u_t - u_{xxx} = 0$, *displayed for time* $0 \le t \le 10^{-3}$.

- $\lim_{t\to 0+} u(x,t) = 0$ for $0 < x \le 1$ (although the integral for $u(x,t)$ diverges if $t = 0$ is substituted directly into it).
- The function $f(t)$ (as obtained from (5.2)) is not identically equal to one although it satisfies $f(0) = 1$ and $f^k(0) = 0,\ k = 1, 2, \ldots$.

Figure 5.1 shows $u(x,t)$ for $0 \le t \le 10^{-3}$, illustrating how high-frequency waves emerge out of the singular corner and then get absorbed (with no reflections) at the right edge.

5.2. Features of the solution to $iu_t - u_{xx} = 0$ in the case of two boundaries. Consider the IBV problem

$$\begin{aligned} &\text{PDE:} && iu_t - u_{xx} = 0, && \\ &\text{IC:} && u(x,0) = 0, && 0 \le x \le 1, \\ &\text{BCs:} && u(0,t) = \sin t,\ u(1,t) = 0, && t > 0. \end{aligned} \tag{5.3}$$

The long-term solution

$$u_L(x,t) = \frac{i}{2}\left(e^{it}\frac{\sin(1-x)}{\sin 1} - e^{-it}\frac{\sinh(1-x)}{\sinh 1}\right)$$

satisfies the PDE and the BCs. The fast scale solution, emanating from the corner, is

$$u_T(x,t) = u(x,t) - u_L\,(x,t)$$

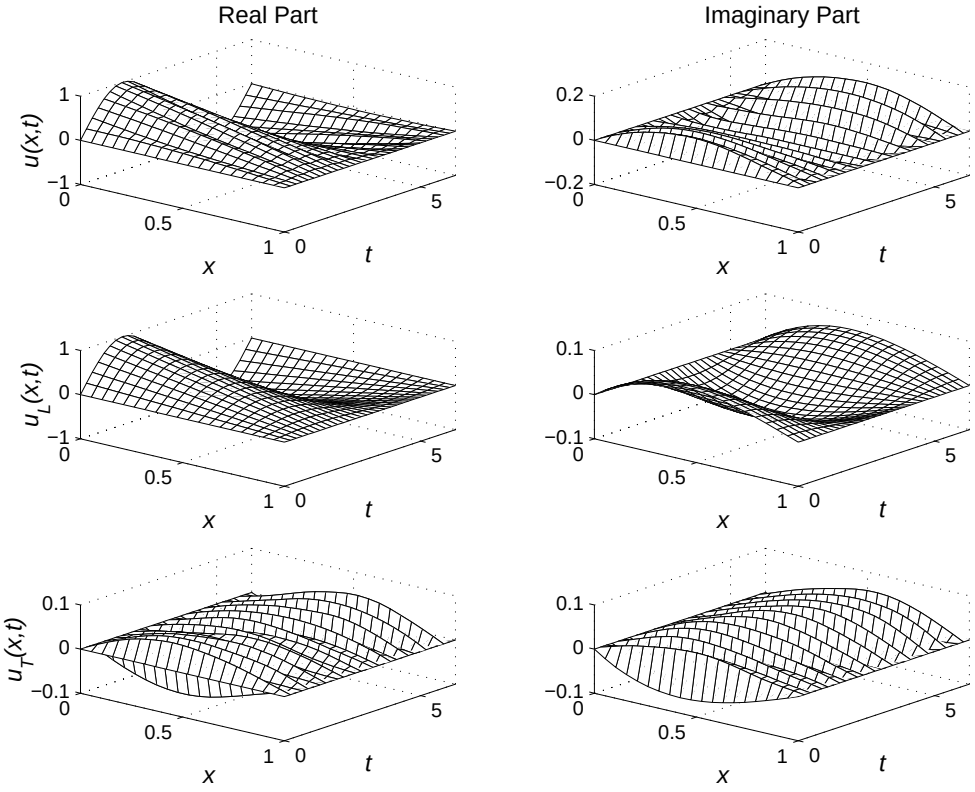


FIG. 5.2. *Full solution $u(x,t)$, long-term solution $u_L(x,t)$, and transient solution $u_T(x,t)$ for* (5.3).

and will again need to satisfy the PDE but with different IC and BC:

$$\begin{aligned} &\text{IC:} \quad u_T(x,0) = -\tfrac{i}{2}\left(\tfrac{\sin(1-x)}{\sin 1} - \tfrac{\sinh(1-x)}{\sinh 1}\right), && 0 \le x \le 1, \\ &\text{BC:} \ u_T(0,t) = u_T(1,t) = 0, && t > 0. \end{aligned}$$

It can be written as a simple sine series expansion:

$$u_T(x,t) = 2\pi i \sum_{k=1}^{\infty} \frac{k}{1-(k\pi)^4} e^{i(k\pi)^2 t} \sin k\pi x. \tag{5.4}$$

Figure 5.2 shows the real and imaginary parts of the $u(x,t)$, $u_L(x,t)$, and $u_T(x,t)$. Figure 5.3 shows the full solution over a short time interval, revealing

1. emanating waves from the corner, as described by the $u_n(x,t)$ corner functions (cf. Figure 4.1) and
2. the reflection of all waves at the boundaries.

The latter fact means that, in contrast to the heat equation, $u_t - u_{xx} = 0$, or the linear KdV equation, $u_t - u_{xxx} = 0$, the solution will not become smoother with time. Indeed, (5.4) shows that $\frac{\partial^4 u}{\partial x^4}$ and $\frac{\partial^2 u}{\partial t^2}$ will fail to exist at almost all x and t. Unless (5.3) is modified to contain some form of dissipation (interior or at the boundaries), accurate numerical solutions would appear to be quite difficult to obtain.

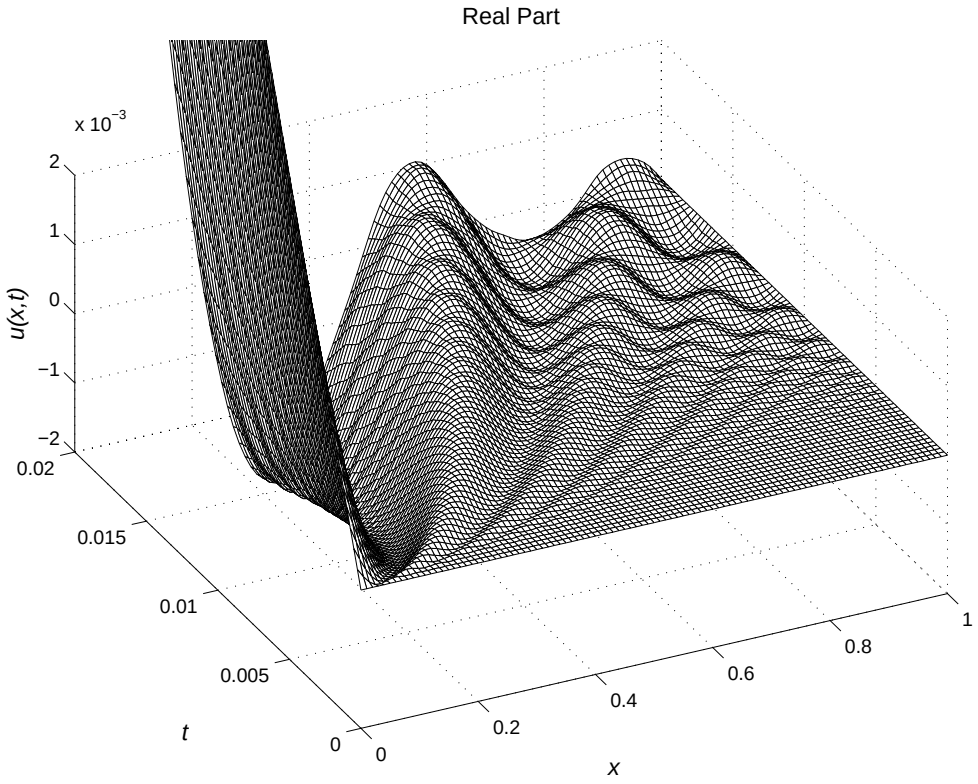


FIG. 5.3. *Real part of the solution $u(x,t)$ to the IBVP test problem for $iu_t - u_{xx} = 0$, shown for $0 \le t \le 0.02$.*

6. Concluding remarks. Since the initial data and the boundary data for PDEs typically arise from different considerations, discrepancies will almost always occur in the corners of the time-space domain. Unless infinitely many compatibility conditions hold in the corners, the solutions will feature singularities which may or may not remain local in time and space. With modern high-order or spectral methods, these discrepancies are often the dominant source of numerical error. It is thus essential to

1. identify and understand the nature of the corner singularities for the IBV problem being solved, and
2. devise numerical remedies to restore expected levels of accuracy.

Both issues have been addressed for second-order convective-diffusive equations in [4]. This study has focused on the first point above for dispersive equations, showing that the concept of corner basis functions, introduced in [4], is critical in understanding the nature of the singularities.

To summarize the different characters of the corner singularities for the PDEs considered, analytic expressions for the first corner basis function, $u_0(x,t)$, are given in Table 6.1 and are graphically contrasted over two different time intervals in Figure 6.1. The analytical form of the corner basis functions for the general PDE $u_t \pm u_{nx} = 0$, $n = 1, 2, 3, \ldots$, has also been included in the table. The constants c_k are determined by the BC $u_0(0,t) = 1$ and all higher derivative BCs equal to zero. The dissipative case $u_t - u_{4x} = 0$ is also illustrated.

TABLE 6.1
Analytic expressions for the $u_0(x,t)$ corner functions for some PDEs of the form $u_t \pm u_{nx} = 0$.

Equation	Elementary form	Hypergeometric form
$u_t - u_x = 0$	0 if $x > t$ 1 if $x < t$	—
$u_t - u_{xx} = 0$	$\operatorname{Erfc}\left(\frac{x}{2\sqrt{t}}\right)$	$1 - \frac{x}{\sqrt{\pi t}}\ {}_1F_1\left(\frac{1}{2}, \frac{3}{2}, -\frac{x^2}{4t}\right)$
$u_t - u_{xxx} = 0$	—	$1 - \frac{\sqrt{3}\,\Gamma(\frac{2}{3})\,x^2}{4\pi\, t^{2/3}}\ {}_1F_2\left(\frac{2}{3}, \left\{\frac{4}{3}, \frac{5}{3}\right\}, -\frac{x^3}{27t}\right)$
$u_t + u_{4x} = 0$	—	$1 - \frac{1}{2\sqrt{\pi}}\frac{x^2}{t^{2/4}}\ {}_1F_3\left(\frac{2}{4}, \left\{\frac{3}{4}, \frac{5}{4}, \frac{6}{4}\right\}, \frac{x^4}{256t}\right)$ $+\frac{\Gamma(\frac{3}{4})}{6\pi}\frac{x^3}{t^{3/4}}\ {}_1F_3\left(\frac{3}{4}, \left\{\frac{5}{4}, \frac{6}{4}, \frac{7}{4}\right\}, \frac{x^4}{256t}\right)$
$\vdots$	$\vdots$	$\vdots$
$u_t \pm u_{nx} = 0$	—	$\sum_{k=0}^{n-1}\left[c_k \frac{x^k}{t^{k/n}}\right.$ $\left.\times\ {}_1F_{n-1}\left(\frac{k}{n}, \left\{\frac{k+1}{n}, \frac{k+2}{n}, \dots, \frac{k+n}{n}\right\}, \pm\frac{x^n/t}{n^n}\right)\right]$, where the c_k are constants; the entry $= 1$ is omitted in the sequence $\left\{\frac{k+1}{n}, \frac{k+2}{n}, \dots, \frac{k+n}{n}\right\}$

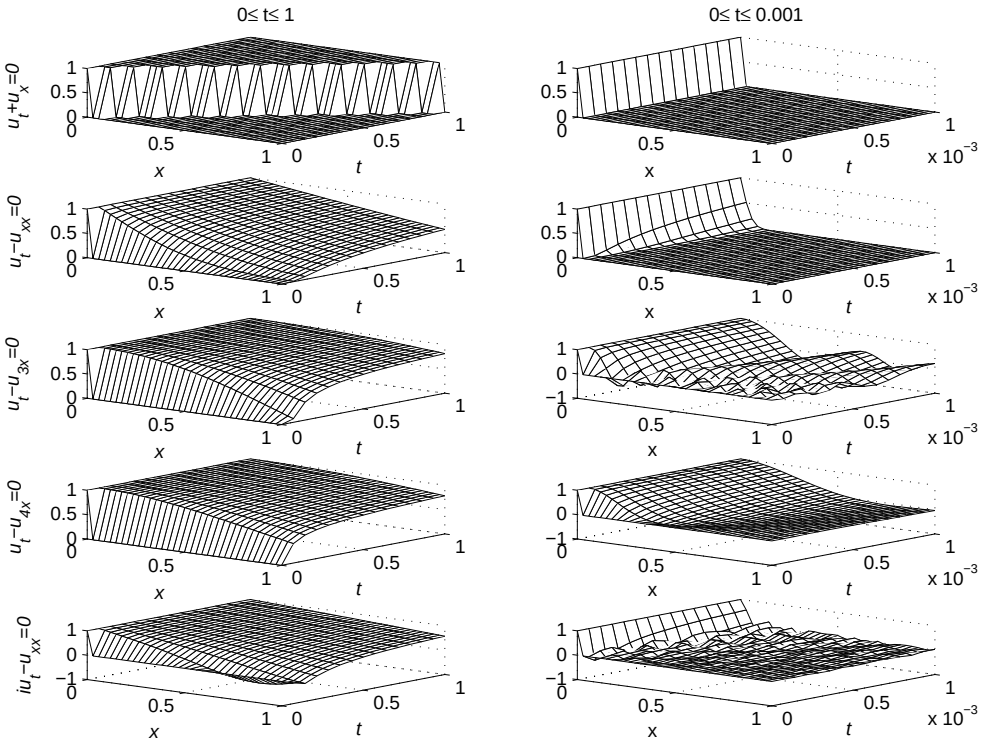


FIG. 6.1. *The $u_0(x,t)$ corner functions for the equations, $u_t \pm u_{nx} = 0$, $n = 1, \dots, 4$, and $iu_t - u_{xx} = 0$ (real part).*

The nature of the corner singularities for each PDE is different, including propagation of the discontinuity throughout the domain, dissipation of it locally, and high-frequency oscillations which either get absorbed or reflected at boundaries. In subsequent studies, numerical techniques for restoring accuracy of high-order and spectral methods for dispersive IBV problems will be explored.

REFERENCES

[1] E.W. Barnes, *The asymptotic expansion of integral functions defined by generalized hypergeometric series*, Proc. London Math. Soc. Ser. 2, 5 (1906), pp. 59–116.

[2] J.P. Boyd and N. Flyer, *Compatibility conditions for time-dependent partial differential equations and the rate of convergence of Chebyshev and Fourier spectral methods*, Comput. Methods Appl. Mech. Engrg., 175 (1999) pp. 281–309.

[3] N. Flyer and P.N. Swarztrauber, *The convergence of spectral and finite difference methods for initial-boundary value problems*, SIAM J. Sci. Comput., 23 (2002), pp. 1731–1751.

[4] N. Flyer and B. Fornberg, *Accurate numerical resolution of transients in initial-boundary value problems for the heat equation*, J. Comput. Phys, 184 (2003), pp. 526–539.

[5] A. Fokas and B. Pelloni, *The solution of certain initial boundary-value problems for the linearized Korteweg-deVries equation*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 454 (1998), pp. 645–657.

[6] B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge Monogr. Appl. Comput. Math. 1, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 565–582

NEW FREQUENCY-AVERAGED APPROXIMATIONS TO THE EQUATIONS OF RADIATIVE HEAT TRANSFER*

EDWARD W. LARSEN†, GUIDO THÖMMES‡, AND AXEL KLAR‡

Abstract. We develop new direction- and frequency-averaged approximations to the equations of radiative heat transfer in glass for optically thick, diffusive regimes. These approximations, which are based on the SP_N approach given in [E. Larsen, G. Thömmes, A. Klar, M. Seaïd, and T. Götz, *J. Comput. Phys.*, 183 (2002), pp. 652–675], represent asymptotic corrections to the familiar Rosseland, or equilibrium diffusion, approximation. Numerical results for realistic problems in the simulation of radiative heat transfer in glass cooling confirm the accuracy and efficiency of the new approximations.

Key words. radiative transfer equations, simplified P_N (SP_N) approximations, frequency averages, asymptotic analysis

AMS subject classifications. 85A25, 35K20, 80M35

DOI. 10.1137/S0036139902413842

1. Introduction.

We consider industrial radiative heat transfer (RHT) processes, such as those that occur in gas turbines, or in the cooling of hot glass molds. These phenomena are governed by the radiative transfer equations [7], which in scaled form contain the dimensionless parameter

$$\varepsilon = \frac{1}{\kappa_{ref} x_{ref}}, \tag{1.1}$$

where κ_{ref} and x_{ref} are reference absorption and length scales. Physically, ε is the ratio of a typical photon mean free path to a typical length scale of the problem. For a spatial point x in a domain $V \subset \mathbb{R}^3$ consisting of glass, the following radiative transfer equations hold for all $\varepsilon > 0$:

$$\varepsilon^2 \frac{\partial T}{\partial t} = \varepsilon^2 \nabla \cdot k \nabla T - \int_{\nu_1}^{\infty} \int_{S^2} \kappa (B - I)\, d\Omega d\nu, \tag{1.2a}$$

$$\Omega \in S^2 : \quad \varepsilon \Omega \cdot \nabla I = \kappa (B - I). \tag{1.2b}$$

Here, B denotes Planck's function for black body radiation in glass:

$$B(\nu, T) = n_1^2 \frac{2 h_P \nu^3}{c^2} \left(e^{h_P \nu / k_B T} - 1 \right)^{-1},$$

which contains Planck's constant h_P, Boltzmann's constant k_B, the index of refraction in glass n_1, the speed of light c, and the material temperature T. On the system boundary ($x \in \partial V$), the incident radiation intensity I is prescribed by semitransparent, i.e., partly reflective, boundary conditions

$$\forall n \cdot \Omega < 0 : \quad I(\Omega) = \rho(n \cdot \Omega) I(\Omega') + \left[1 - \rho(n \cdot \Omega) \right] B^{(a)}(\nu, T_b), \tag{1.2c}$$

*Received by the editors September 2, 2002; accepted for publication (in revised form) June 26, 2003; published electronically December 31, 2003. This work was supported by DFG grant KL 1105/7 and SFB 568.

http://www.siam.org/journals/siap/64-2/41384.html

†Department of Nuclear Engineering and Radiological Sciences, University of Michigan, Ann Arbor, Michigan 48109–2104 (edlarsen@eugin.umich.edu).

‡Department of Mathematics, Technical University of Darmstadt, 64289 Darmstadt, Germany (theommes@mathematik.tu-darmstadt.de, klar@mathematik.tu-darmstadt.de).

where the direction Ω' of the exiting ray in (1.2c), which specularly reflects into the incident ray Ω, is given by

$$\Omega' = \Omega - 2(n \cdot \Omega)n.$$

Note that we have to use Planck's function $B^{(a)}$ for air,

$$B^{(a)}(\nu, T) = n_2^2 \frac{2h_P \nu^3}{c^2} \Big(e^{h_P \nu / k_B T} - 1 \Big)^{-1},$$

in the boundary conditions. The temperature at the boundary obeys

$$\varepsilon k \, n \cdot \nabla T = h(T_b - T) + \alpha\pi \int_0^{\nu_1} \Big[B^{(a)}(\nu, T_b) - B^{(a)}(\nu, T) \Big] \, d\nu. \tag{1.2d}$$

At the initial time $t = 0$, T is prescribed for all $x \in V$ by

$$T(x, 0) = T_0(x). \tag{1.2e}$$

In (1.2), $I(x, t, \Omega, \nu)$ denotes the *specific radiation* intensity at point $x \in V$ traveling in direction $\Omega \in S^2$ with frequency $\nu > 0$ at time $t \geq 0$. $T(x, t)$ denotes the *material temperature* at point $x \in V$ at time $t \geq 0$.

The outside radiation is assumed to be a Planckian for air, $B^{(a)}(\nu, T_b)$, at specified temperature T_b for incident directions on the boundary (satisfying $n \cdot \Omega < 0$, where n denotes the outward normal on ∂V). Equations (1.2) contain the known opacity $\kappa(\nu)$, heat conductivity k, and convective heat transfer coefficient h. The integration in the second term of the temperature boundary condition (1.2d) is performed on the opaque interval of the spectrum $[0, \nu_1]$, where radiation is completely absorbed. The refractive indices at the boundary are n_1 (glass) $> n_2$ (exterior air).

The reflectivity $\rho \in [0, 1]$ is, from (1.2c), the fraction of exiting radiation that is reflected back into V. This is defined by

$$\rho(\mu) = \begin{cases} 1, & \sin\theta_1 \geq n_2/n_1, \\ \frac{1}{2}\Big(\frac{\tan^2(\theta_1 - \theta_2)}{\tan^2(\theta_1 + \theta_2)} + \frac{\sin^2(\theta_1 - \theta_2)}{\sin^2(\theta_1 + \theta_2)} \Big), & \sin\theta_1 < n_2/n_1, \end{cases} \tag{1.3}$$

where the refraction angles θ_1 and θ_2 are given by $\cos\theta_1 = |n \cdot \Omega| = \mu$ and

$$n_1 \sin\theta_1 = n_2 \sin\theta_2.$$

Finally, the hemispheric emissivity α is defined in terms of the reflectivity ρ by

$$\alpha = 2n_1 \int_0^1 [1 - \rho(\mu)] \, d\mu.$$

In astrophysical problems, an extra term $\partial I / c\partial t$ is often included in the left-hand side of (1.2b). However, for the problems of glass annealing considered in this paper, this term is negligibly small, so we ignore it. For a full discussion of the above equations and their applications we refer, for example, to [6], [10], [5] and the monographs [3] and [8].

In the optically thick regime, where the opacity κ is large and the radiation is conveyed in a diffusion-like manner, it is appropriate to consider the above equations for $\varepsilon \ll 1$. In this "diffusive regime, the solution of the transport equation asymptotically limits to the solution of the *equilibrium diffusion equation* (see [6])

$$\frac{\partial T}{\partial t} = \nabla \cdot k \nabla T + \nabla \cdot \Big(\frac{4\pi}{3} \int_{\nu_1}^{\infty} \frac{1}{\kappa} \frac{\partial B}{\partial T} \, d\nu \Big) \nabla T, \tag{1.4}$$

which is independent of the direction and frequency variables.

Considering numerical solutions of the above equations, the solution of the radiative transfer equations is much more complex and costly than the solution of the equilibrium diffusion equation. However, in many situations, the equilibrium diffusion solution is not sufficiently accurate. For example, in the annealing (cooling) of glass, the radiative transfer process produces boundary layers in T at the outer glass surface. These give rise to mechanical stresses—due to the thermal contraction of the glass upon cooling—which can cause defects or breaking. The equilibrium diffusion approximation does not accurately describe these important boundary layer regions.

Therefore, approximations to the full radiative transfer model that are computationally less time consuming, yet sufficiently more accurate than equilibrium diffusion theory, have been sought. In [7], *simplified P_N* ($\mathrm{SP_N}$) approximations are developed as alternatives to the full radiative transfer equations and the usual $\mathrm{P_N}$ (spherical harmonic) approximations. The $\mathrm{SP_N}$ approximations are derived by asymptotic and variational analyses and employ frequency-dependent diffusion equations, or coupled systems of frequency-dependent diffusion equations. Approximate models of this type contain boundary layer effects and can be remarkably accurate—much more accurate than equilibrium diffusion theory; see [7].

A drawback of the $\mathrm{SP_N}$ approximations, from the point of view of efficient simulations, is the large number of frequency bands often encountered in applications. For the RHT equations and their $\mathrm{P_N}$ and $\mathrm{SP_N}$ approximations, this leads to a coupled system of equations, since each frequency band is treated in a separate equation. In the present paper, we develop a new *grey simplified P_N* ($\mathrm{GSP_N}$) approximation that eliminates the frequency variable. (More precisely, the new grey approximation compresses all the frequency bands into one. The term "grey," borrowed from the astrophysical literature, means "frequency-independent.") We also develop "partially averaged" approximations, in which the number of frequency bands is reduced, but not to one. (A "fully averaged" approximation would have only one band, and would be "grey.") The elimination or reduction of frequency significantly reduces computation time. Yet, our numerical results show that this process yields solutions that are much more accurate than the equilibrium diffusion solution and are often nearly as accurate as the $\mathrm{SP_N}$ solution, even in the important boundary layers. We believe that the resulting approximations will provide practical industrial tools for simulating the annealing of glass.

The remainder of this paper is organized as follows: In section 2, the *outer* asymptotic expansion to derive the frequency-dependent $\mathrm{SP_N}$ approximations to (1.2a) and (1.2b) are reviewed. In sections 3 and 4, the GSPN procedure—in which the frequency is eliminated as an independent variable—is presented. This procedure has the practical advantage that it leads to equations that are much less expensive to compute. However, it represents a significant step beyond standard $\mathrm{SP_N}$ theory. In section 5, numerical comparisons of the $\mathrm{SP_N}$, $\mathrm{GSP_N}$, and equilibrium diffusion approximations are presented. We conclude with a discussion in section 6.

2. The SP_1 and SP_2 approximations. To solve (1.2a) in the domain V, we write the transport equation as

$$\Big(1+\frac{\varepsilon}{\kappa}\Omega\cdot\nabla\Big)I(x,t,\Omega,\nu)=B(\nu,T)$$

and use a Neumann series to formally solve for I. Assuming a homogeneous medium, one obtains

$$\begin{aligned} I &= \Big(1+\frac{\varepsilon}{\kappa}\Omega\cdot\nabla\Big)^{-1}B \\ &\widetilde{=} \Big[1-\frac{\varepsilon}{\kappa}\Omega\cdot\nabla+\frac{\varepsilon^2}{\kappa^2}(\Omega\cdot\nabla)^2-\frac{\varepsilon^3}{\kappa^3}(\Omega\cdot\nabla)^3+\frac{\varepsilon^4}{\kappa^4}(\Omega\cdot\nabla)^4\cdots\Big]B. \end{aligned} \tag{2.1}$$

Integrating with respect to Ω and using

$$\int_{S^2}(\Omega\cdot\nabla)^n f\ d\Omega=[1+(-1)^n]\,\frac{2\pi}{n+1}\nabla^n f,$$

where $\nabla^2=\nabla\cdot\nabla$ is the spatial Laplacian and f is sufficiently smooth, we obtain

$$\begin{aligned} \phi(x,t,\nu) &= \int_{S^2} I(x,t,\Omega,\nu)\ d\Omega \\ &= 4\pi\Big[1+\frac{\varepsilon^2}{3\kappa^2}\nabla^2+\frac{\varepsilon^4}{5\kappa^4}\nabla^4+\frac{\varepsilon^6}{7\kappa^6}\nabla^6\cdots\Big]B+\mathcal{O}(\varepsilon^8). \end{aligned}$$

Hence

$$\begin{aligned} 4\pi B &= \Big[1+\frac{\varepsilon^2}{3\kappa^2}\nabla^2+\frac{\varepsilon^4}{5\kappa^4}\nabla^4+\frac{\varepsilon^6}{7\kappa^6}\nabla^6\Big]^{-1}\phi+\mathcal{O}(\varepsilon^8) \\ &= \Bigg\{1-\Big[\frac{\varepsilon^2}{3\kappa^2}\nabla^2+\frac{\varepsilon^4}{5\kappa^4}\nabla^4+\frac{\varepsilon^6}{7\kappa^6}\nabla^6\Big] \\ &\quad+\Big[\frac{\varepsilon^2}{3\kappa^2}\nabla^2+\frac{\varepsilon^4}{5\kappa^4}\nabla^4+\frac{\varepsilon^6}{7\kappa^6}\nabla^6\Big]^2 \\ &\quad-\Big[\frac{\varepsilon^2}{3\kappa^2}\nabla^2+\frac{\varepsilon^4}{5\kappa^4}\nabla^4+\frac{\varepsilon^6}{7\kappa^6}\nabla^6\Big]^3\cdots\Bigg\}\phi+\mathcal{O}(\varepsilon^8), \\ &= \Big[1-\frac{\varepsilon^2}{3\kappa^2}\nabla^2-\frac{4\varepsilon^4}{45\kappa^4}\nabla^4-\frac{44\varepsilon^6}{945\kappa^6}\nabla^6\Big]\phi+\mathcal{O}(\varepsilon^8). \end{aligned} \tag{2.2}$$

If we discard terms of $\mathcal{O}(\varepsilon^4)$ or $\mathcal{O}(\varepsilon^6)$—and rearrange asymptotically, if necessary, to obtain robust second-order partial differential equations—we obtain the SP_1 and SP_2 approximations, respectively, for $\phi(x,t,\nu)$. We now summarize these different approximate equations and boundary conditions, derived in [7] (see also [9], [2]).

The SP_1 approximation, which is equivalent to the P_1 spherical harmonic approximation, is

$$-\varepsilon^2\nabla\cdot\Big(\frac{1}{3\kappa}\nabla\phi\Big)+\kappa\phi=\kappa(4\pi B)+\mathcal{O}(\varepsilon^4),\quad x\in V, \tag{2.3}$$

$$\frac{\partial T}{\partial t}=\nabla\cdot(k\nabla T)+\int_{\nu_1}^{\infty}\nabla\cdot\Big(\frac{1}{3\kappa}\nabla\phi\Big)d\nu+\mathcal{O}(\varepsilon^2), \tag{2.4}$$

$$\phi(x,\nu,t)+\Big(\frac{1+3r_2}{1-2r_1}\Big)\Big(\frac{2\varepsilon}{3\kappa}\Big)n\cdot\nabla\phi(x,\nu,t)=4\pi B^{(a)}(\nu,T_b),\quad x\in\partial V. \tag{2.5}$$

The SP_2 approximation is

$$-\varepsilon^2\nabla\cdot\left(\frac{3}{5\kappa}\nabla\phi\right)+\kappa\phi=\kappa(4\pi B)+\mathcal{O}(\varepsilon^6),\quad x\in V, \tag{2.6}$$

$$\frac{\partial T}{\partial t}=\nabla\cdot(k\nabla T)+\int_{\nu_1}^{\infty}\nabla\cdot\left(\frac{1}{3\kappa}\ \nabla\phi\right)d\nu+\mathcal{O}(\varepsilon^4), \tag{2.7}$$

$$\begin{aligned}\phi(x,\nu,t)+\left(\frac{1+3r_2}{1-4r_3}\right)\left(\frac{4\varepsilon}{5\kappa}\right)n\cdot\nabla\phi(x,\nu,t)\\ =4\pi B^{(a)}(\nu,T)+\left(\frac{1-2r_1}{1-4r_3}\right)\left(\frac{24\pi}{5}\right)[B^{(a)}(\nu,T_b)-B^{(a)}(\nu,T)],\quad x\in\partial V.\end{aligned} \tag{2.8}$$

The constants r_1, r_2, and r_3 in (2.5) and (2.8) are moments of $\rho(\mu)$:

$$r_n=\int_0^1\ \mu^n\rho(-\mu)\ d\mu,\quad n=1,2,3.$$

The boundary and initial conditions for T in both the SP_1 and SP_2 approximations are given by (1.2d) and (1.2e).

The SP_1 approximation (2.3)–(2.5) and the SP_2 approximation (2.6)–(2.8) do not contain the direction variable Ω, but they do contain the frequency variable ν. Our goal in this paper is to develop accurate frequency-independent (grey) and partially frequency-averaged approximations to these frequency-dependent SP_N approximations.

In [7], the differential equations (2.3), (2.4), (2.6), and (2.7) are derived by an asymptotic analysis based on (2.2). The formal asymptotic order of error is indicated in each of these above equations. However, the boundary conditions (2.5) and (2.8) were derived by an approximate variational analysis, not by an asymptotic analysis, and it is not known if a rigorous asymptotic order of error can be assigned to them. In spite of this lack of strict asymptotic "pedigree," the SP_1 and SP_2 approximations have been shown to be accurate approximations to the full radiative transfer equations for small and even intermediate values of ε (see [7]).

In this paper, we treat the SP_1 and SP_2 approximations as fundamental equations and derive formal asymptotic approximations to them. Specifically, we derive (i) the GSP_1 *approximation*, a grey (frequency-independent) approximation to the SP_1 equations with $\mathcal{O}(\varepsilon^2)$ error, and (ii) the GSP_2 *approximation*, a grey approximation to the SP_2 equations with $\mathcal{O}(\varepsilon^3)$ error. The GSP_1 approximation is equivalent to the well-known *Rosseland*, or *equilibrium diffusion* approximation, while the GSP_2 approximation is new and represents an asymptotic correction to the Rosseland approximation. We also derive "partially averaged" approximations, in which frequency is not eliminated from the problem, but the number of frequency bands is significantly reduced.

If the SP_1 equations are an $\mathcal{O}(\varepsilon^2)$ approximation to the full radiative transfer model (this assumption amounts to asserting that (2.5) has $\mathcal{O}(\varepsilon^2)$ error), then the GSP_1 equations approximate the radiative transfer model with $\mathcal{O}(\varepsilon^2)$ error. Likewise, if the SP_2 equations are an $\mathcal{O}(\varepsilon^3)$ approximation to the full radiative transfer model (this assumption amounts to asserting that (2.8) has $\mathcal{O}(\varepsilon^3)$ error), then the GSP_2 equations approximate the radiative transfer model with $\mathcal{O}(\varepsilon^3)$ error.

We do not attempt to calculate a grey approximation to the SP_1 equations with error $\mathcal{O}(\varepsilon^3)$ or smaller, because the resulting grey approximation can have no better

than $\mathcal{O}(\varepsilon^2)$ error relative to the full radiative transfer equations. It does make sense to attempt to calculate a grey approximation to the SP_2 equations with error $\mathcal{O}(\varepsilon^4)$, because the resulting grey approximation *could* then have $\mathcal{O}(\varepsilon^4)$ error relative to the full radiative transfer equations. However, the best that we could manage is the $\mathcal{O}(\varepsilon^3)$ approximation presented in this paper.

Regardless of these asymptotic subtleties, numerical simulations presented in section 5 demonstrate that the GSP_2 approximation is significantly more accurate than the GSP_1 (Rosseland) approximation, yet is only slightly more expensive to compute. For very difficult problems, the partially averaged equations are shown to be even more accurate than the grey equations, typically almost as accurate as the original SP_N equations on which the approximations developed in this paper are directly based. As might be expected, the cost of the partially averaged simulations is greater than the cost of the grey calculations but less than the cost of the full SP_N calculations.

3. The GSP_1 approximation. The starting point for deriving the GSP_1 approximation is the system of SP_1 equations (2.3)–(2.5) stated above.

In (2.3) and (2.4), we have indicated that the formal errors are $\mathcal{O}(\varepsilon^4)$ and $\mathcal{O}(\varepsilon^2)$, respectively (see [7] for details). Overall, the error in the SP_1 equations is expected to be $\mathcal{O}(\varepsilon^2)$. If we delete the $\mathcal{O}(\varepsilon^2)$ diffusion term in (2.3) and also the boundary condition (2.5)—since these are not strictly needed—(2.3) yields, with $\mathcal{O}(\varepsilon^2)$ error,

$$\phi = 4\pi B. \tag{3.1}$$

Then (2.4) becomes, also with $\mathcal{O}(\varepsilon^2)$ error,

$$\begin{aligned}\frac{\partial T}{\partial t} - \nabla\cdot(k\nabla T) &= \int_{\nu_1}^{\infty} \nabla\cdot\frac{4\pi}{3\kappa}\,\nabla\, B(\nu,T)d\nu \\ &= \nabla\cdot\int_{\nu_1}^{\infty}\frac{4\pi}{3\kappa}\,\frac{\partial B}{\partial T}(\nu,T)\,\nabla T d\nu \\ &= \nabla\cdot\Big(\frac{4\pi}{3}\int_{\nu_1}^{\infty}\frac{1}{\kappa}\frac{\partial B}{\partial T}(\nu,T)d\nu\Big)\nabla T.\end{aligned}$$

Equivalently,

$$\frac{\partial T}{\partial t} = \nabla\cdot[k + k_R(T)]\,\nabla T, \tag{3.2}$$

with

$$k_R(T) = \frac{4\pi}{3}\int_{\nu_1}^{\infty}\frac{1}{\kappa(\nu)}\frac{\partial B}{\partial T}(\nu,T)d\nu. \tag{3.3}$$

Equations (3.2), (3.3), (1.2d), and (1.2e) constitute the familiar *Rosseland*, or *equilibrium diffusion*, approximation. In our lexicon, we also call this the *grey simplified* P_1 (GSP_1) approximation, because it is a frequency-independent (grey) approximation to the frequency-dependent SP_1 equations. The formal error in this approximation is $\mathcal{O}(\varepsilon^2)$, which is the same as the formal error in the SP_1 approximation.

However, in discarding the $\varepsilon^2\nabla\cdot\frac{1}{3\kappa}\nabla\phi$ term in (2.3) and the boundary condition (2.5) on ϕ, we thereby discarded the boundary layers in the SP_1 solution. The frequency-dependent SP_1 solution contains boundary layers, but the frequency-independent GSP_1 (Rosseland) solution does not.

4. The $\mathbf{GSP_2}$ approximation.

4.1. $\mathbf{GSP_2}$ equations. The starting point now is the system of SP_2 (2.6)–(2.8). Equation (2.6) for a homogeneous medium gives

$$\left(1 - \frac{3\varepsilon^2}{5\kappa^2}\nabla^2\right)\phi = 4\pi B + \mathcal{O}(\varepsilon^6),$$

so

$$\phi = \left(1 + \frac{3\varepsilon^2}{5\kappa^2}\nabla^2\right)4\pi B + \mathcal{O}(\varepsilon^4). \tag{4.1}$$

Then,

$$\frac{\phi}{3\kappa} = \frac{4\pi}{3}\frac{B}{\kappa} + \varepsilon^2\nabla^2\frac{4\pi}{5}\frac{B}{\kappa^3} + \mathcal{O}(\varepsilon^4) \tag{4.2}$$

and

$$\frac{1}{3}\int_{\nu_1}^{\infty}\frac{\phi}{\kappa}\,d\nu = \frac{4\pi}{3}\int_{\nu_1}^{\infty}\frac{B}{\kappa}\,d\nu + \varepsilon^2\nabla^2\frac{4\pi}{5}\int_{\nu_1}^{\infty}\frac{B}{\kappa^3}\,d\nu + \mathcal{O}(\varepsilon^4). \tag{4.3}$$

Thus, if we define

$$W(x,t) = \frac{1}{3}\int_{\nu_1}^{\infty}\frac{\phi(x,\nu,t)}{\kappa(\nu)}\,d\nu \tag{4.4}$$

and

$$f_n(T) = \frac{4\pi}{n+2}\int_{\nu_1}^{\infty}\frac{B(\nu,T)}{\kappa^n(\nu)}\,d\nu \qquad (n = 1, 3), \tag{4.5}$$

which satisfy $f_n'(T) > 0$, then (2.7) and (4.3) become, with $\mathcal{O}(\varepsilon^4)$ error,

$$\frac{\partial T}{\partial t} = \nabla\cdot(k\nabla T) + \nabla^2 W, \tag{4.6}$$

$$W = f_1(T) + \varepsilon^2\nabla^2 f_3(T). \tag{4.7}$$

It is now trivial to use (4.7) to eliminate W from (4.6), producing a nonlinear fourth-order (in space) differential equation for T:

$$\frac{\partial T}{\partial t} = \nabla\cdot(k\nabla T) + \nabla^2 f_1(T) + \varepsilon^2(\nabla^2)^2 f_3(T), \tag{4.8}$$

which has formal error $\mathcal{O}(\varepsilon^4)$. However, this equation is ill-conditioned, so we do not propose to attempt to solve it for T. Instead we convert (4.7) into a well-conditioned second-order differential equation for W, with formal $\mathcal{O}(\varepsilon^4)$ asymptotic error.

To do this, we first write (4.7) as

$$W = f_1(T) + \mathcal{O}(\varepsilon^2),$$

so

$$\nabla W = f_1'(T)\nabla T + \mathcal{O}(\varepsilon^2),$$

and hence

$$\nabla T = \frac{1}{f_1'(T)}\nabla W + \mathcal{O}(\varepsilon^2). \tag{4.9}$$

Equations (4.7) and (4.9) now give, with $\mathcal{O}(\varepsilon^4)$ error,

$$\begin{aligned} W &= f_1(T) + \varepsilon^2 \nabla^2 f_3(T) \\ &= f_1(T) + \varepsilon^2 \nabla \cdot f_3'(T) \nabla T \\ &= f_1(T) + \varepsilon^2 \nabla \cdot \frac{f_3'(T)}{f_1'(T)} \nabla W, \end{aligned}$$

or

$$-\varepsilon^2 \nabla \cdot \left(\frac{f_3'(T)}{f_1'(T)} \nabla W \right) + W = f_1(T). \tag{4.10}$$

In (4.10), the positive function

$$\frac{f_3'(T)}{f_1'(T)} \equiv D(T)$$

has the role of a space-dependent diffusion coefficient. This is physically appropriate, as W is an intensity-like quantity, and the diffusion coefficient for W should intuitively depend on T.

The GSP_2 equations, which are defined throughout the physical system V, consist of the two coupled (4.6) and (4.10) for T and W. The formal error in these equations is $\mathcal{O}(\varepsilon^4)$. Next, we derive initial and boundary conditions for these equations. Unfortunately, the error in the boundary condition for W is $\mathcal{O}(\varepsilon^3)$, not $\mathcal{O}(\varepsilon^4)$. Still, this is more accurate than the $\mathcal{O}(\varepsilon^2)$ error in the Rosseland approximation.

4.2. $\mathbf{GSP_2}$ initial and boundary conditions. As before, the initial and boundary conditions for (4.6) are (1.2e) and (1.2d), respectively. Using the boundary condition (2.8), we shall derive a boundary condition for

$$W(x,t) = \frac{1}{3} \int_{\nu_1}^{\infty} \frac{\phi(x,\nu,t)}{\kappa(\nu)} d\nu,$$

i.e., for (4.10), with $\mathcal{O}(\varepsilon^3)$ error.

When we introduce the constants

$$\begin{aligned} \alpha_1 &\equiv \frac{4}{5} \left(\frac{1+3r_2}{1-4r_3} \right), \\ \alpha_2 &\equiv \frac{6}{5} \left(\frac{1-2r_1}{1-4r_3} \right), \end{aligned}$$

then (2.8) may be written

$$\phi + \frac{\varepsilon \alpha_1}{\kappa} n \cdot \nabla \phi = 4\pi \left[B^{(a)} + \alpha_2 (B_b^{(a)} - B^{(a)}) \right], \quad x \in \partial V. \tag{4.11}$$

Dividing by 3κ and integrating over $\nu_1 \le \nu < \infty$, we obtain, using (4.4) and (4.5),

$$W + \frac{\varepsilon \alpha_1}{3} n \cdot \nabla \int_{\nu_1}^{\infty} \frac{\phi}{\kappa^2} \, d\nu = f_1^{(a)}(T) + \alpha_2 [f_1^{(a)}(T_b) - f_1^{(a)}(T)]. \tag{4.12}$$

However, (4.1) gives

$$\phi = 4\pi B + \mathcal{O}(\varepsilon^2).$$

Dividing by κ^2 and integrating over $\nu_1 \le \nu < \infty$, we obtain

$$\int_{\nu_1}^{\infty} \frac{\phi}{\kappa^2}\, d\nu = 4\pi \int_{\nu_1}^{\infty} \frac{B}{\kappa^2}\, d\nu + \mathcal{O}(\varepsilon^2) = 4f_2(T) + \mathcal{O}(\varepsilon^2).$$

Hence (4.12) becomes, with $\mathcal{O}(\varepsilon^3)$ error,

$$W + \frac{4\alpha_1\varepsilon}{3} n \cdot \nabla f_2(T) = f_1^{(a)}(T) + \alpha_2[f_1^{(a)}(T_b) - f_1^{(a)}(T)], \quad x \in \partial V. \tag{4.13}$$

We also have from (4.10)

$$W = f_1(T) + \mathcal{O}(\varepsilon^2), \quad x \in V. \tag{4.14}$$

The previous two equations imply

$$W + \frac{4\alpha_1\varepsilon}{3}(f_2)'(T)\; n \cdot \nabla T = f_1^{(a)}(T) + \alpha_2[f_1^{(a)}(T_b) - f_1^{(a)}(T)],$$
$$n \cdot \nabla W = (f_1)'(T)\; n \cdot \nabla T + \mathcal{O}(\varepsilon^2).$$

Eliminating $n \cdot \nabla T$, we obtain, with $\mathcal{O}(\varepsilon^3)$ error,

$$W + \left(\frac{4\alpha_1\varepsilon}{3}\, \frac{(f_2)'(T)}{(f_1)'(T)}\right) n \cdot \nabla W = f_1^{(a)}(T) + \alpha_2[f_1^{(a)}(T_b) - f_1^{(a)}(T)], \quad x \in \partial V. \tag{4.15}$$

This boundary condition for W can be used with (4.10). The positive quantity

$$\frac{4\alpha_1\varepsilon}{3}\, \frac{(f_2)'(T)}{(f_1)'(T)}$$

in (4.15) plays the role of an *extrapolation distance*, which is familiar from neutron transport.

4.3. GSP$_2$ approximation: Summary. To summarize, we propose the following GSP$_2$ equations for $T(x,t)$ and $W(x,t)$, which hold for all $x \in V$ and $t > 0$:

$$\frac{\partial T}{\partial t} = \nabla \cdot (k \nabla T) + \nabla^2 W, \tag{4.16}$$

$$-\varepsilon^2 \nabla \cdot \left(\frac{f_3'(T)}{f_1'(T)} \nabla W\right) + W = f_1(T). \tag{4.17}$$

The following boundary conditions for T and W hold for $x \in \partial V$ and $t > 0$:

$$\varepsilon k\, n \cdot \nabla T = h(T_b - T) + \alpha\pi \int_0^{\nu_1} \left[B^{(a)}(\nu, T_b) - B^{(a)}(\nu, T)\right]\, d\nu, \tag{4.18}$$

$$W + \left(\frac{4\alpha_1\varepsilon}{3}\, \frac{(f_2)'(T)}{(f_1)'(T)}\right) n \cdot \nabla W = f_1^{(a)}(T) + \alpha_2[f_1^{(a)}(T_b) - f_1^{(a)}(T)]. \tag{4.19}$$

The following initial condition for T holds for $x \in V$:

$$T(x,0) = T_0(x). \tag{4.20}$$

The GSP_2 equations formally approximate the SP_2 equations with $\mathcal{O}(\varepsilon^3)$ error. If the SP_2 equations themselves approximate the radiative transfer equations with $\mathcal{O}(\varepsilon^3)$ error—this is tantamount to an assumption on the accuracy of the SP_2 boundary condition—then the GSP_2 equations also approximate the radiative transfer equations with $\mathcal{O}(\varepsilon^3)$ error.

If one chooses to approximate the GSP_2 equations by allowing $\mathcal{O}(\varepsilon^2)$ errors, then the derivative term on the left side of (4.17) can be dropped, and the boundary condition (4.19) for W can then be dropped because it is not needed. The resulting equations reduce to the GSP_1, or Rosseland equations. Hence the above GSP_2 equations are an $\mathcal{O}(\varepsilon^3)$ correction to the $\mathcal{O}(\varepsilon^2)$ Rosseland equation.

In our derivation of the GSP_2 equations and boundary conditions, we assumed that all quantities involved in the manipulations (i.e., the solution and its derivatives) are $\mathcal{O}(1)$. The goal of our asymptotic manipulations has been to obtain a set of equations and boundary conditions that

(a) is well-conditioned,
(b) is asymptotically accurate if the solution and all its derivations are $\mathcal{O}(1)$, and
(c) is robust if derivatives of the solution are large (possibly, in boundary layer regions).

Thus there is no guarantee that the above GSP_2 equations are accurate for problems with boundary layers. Indeed, no boundary layer scaling or analysis has been used in the derivation of these equations! Nevertheless, we show next that numerical solutions of the GSP_2 equations are (i) much more accurate than the Rosseland solution, even in boundary layer regions, and (ii) often of comparable accuracy to the SP_2 equations, even in boundary layer regions. Thus, even though our derivation of the GSP_2 equations does not explicitly account for boundary layers, it seems to accomplish this in some implicit manner.

5. Partial averaging. To obtain more accurate results the averaging procedure may be slightly modified. We proceed in the same way as described above; however, we integrate only over certain intervals in frequency space instead of integrating the respective terms over the whole frequency interval $[\nu_1, \infty)$. This procedure does not lead to a single-band or grey approximation but to a multiband approximation.

Starting with a multiband model (i.e., a piecewise constant absorption coefficient) with N bands, the number of bands can be strongly reduced with such a procedure. We denote the number of frequency bands after the averaging procedure with M.

The procedure leads to the following partially averaged SP_2 equations for $T(x,t)$ and $W_i(x,t)$, $i = 1, \ldots, M$, which hold for all $x \in V$ and $t > 0$:

$$\frac{\partial T}{\partial t} = \nabla \cdot (k \nabla T) + \nabla^2 \sum_{i=1}^{M} W_i, \tag{5.1}$$

$$-\varepsilon^2 \nabla \cdot \left(\frac{(f_3^{(i)})'(T)}{(f_1^{(i)})'(T)} \nabla W_i \right) + W_i = f_1^{(i)}(T), \tag{5.2}$$

with

$$f_n^{(i)}(T) = \frac{4\pi}{n+2} \int_{\nu_i}^{\nu_{i+1}} \frac{B(\nu, T)}{\kappa^n(\nu)} \, d\nu \qquad (n = 1, 3, \;\; i = 1, \ldots, M), \tag{5.3}$$

where $\nu_{M+1} = \infty$.

The following boundary conditions for T and W_i hold for $x \in \partial V$ and $t > 0$:

$$\varepsilon k\, n \cdot \nabla T = h(T_b - T) + \alpha\pi \int_0^{\nu_1} \left[B^{(a)}(\nu, T_b) - B^{(a)}(\nu, T) \right] \, d\nu, \tag{5.4}$$

$$W_i + \left(\frac{4\alpha_1 \varepsilon}{3} \, \frac{(f_2^{(i)})'(T)}{(f_1^{(i)})'(T)} \right) n \cdot \nabla W_i = f_1^{(a,i)}(T) + \alpha_2 [f_1^{(a,i)}(T_b) - f_1^{(a,i)}(T)]. \tag{5.5}$$

The following initial condition for T holds for $x \in V$:

$$T(x, 0) = T_0(x). \tag{5.6}$$

As can be seen in section 6, choosing suitable intervals in frequency space and averaging in the way described above leads to strongly improved results.

6. Numerical comparisons. We now compare solutions of the full radiative transfer equations with solutions of the approximate SP_2 equations, GSP_2 equations, and Rosseland (GSP_1) equation. We investigate problems in one-dimensional slab geometry, in which the temperature and radiation depend spatially on the x-coordinate but not on the y- or z-coordinates. The direction-dependent radiation is assumed to be rotationally symmetric around the x-axis.

The problems considered correspond to the cooling of a uniform slab of glass surrounded by air at room temperature $T_b = 300K$. Initially, the temperature inside the glass is taken to be $T_0(x) = 1000K$. The radiation outside the glass is assumed to be a Planckian, i.e., isotropic with $I_b(\Omega, \nu) = B^{(a)}(\nu, T_b,)$. For $t > 0$, the glass cools down by the processes considered in this paper: regular heat conduction (linear diffusion) and thermal RHT.

We used standard finite difference techniques to discretize the diffusion equations, with uniform space and time grids. We chose $\Delta x = 0.02$ in the scaled interval $[0, 1]$ and 100 equal time steps $\Delta t = 10^{-5}$ to reach the final time $t = 10^{-3}$.

We assumed the scaled physical parameters k and h to have the values $k = 1$ and $h = 1$. The refractive coefficients were chosen to be $n_1 = 1.46$ (glass) and $n_2 = 1$ (air). The corresponding hemispheric emissivity is $\alpha = 0.92$.

We approximated the opacity $\kappa(\nu)$ by a piecewise constant step function defined on the frequency bands $[\nu_i, \nu_{i+1}]$, $i = 1, \ldots, N$ ($\nu_{N+1} = \infty$), and with associated absorption rates $\kappa(\nu) = \kappa_i$ for $\nu \in [\nu_i, \nu_{i+1}]$.

We first considered several problems with only two frequency bands. For $\nu_1 = 4.28 \cdot 10^{14} Hz$ and $\nu_2 = 9.99 \cdot 10^{14} Hz$ (corresponding to $\lambda_1 = 3\mu m$ and $\lambda_2 = 7\mu m$, respectively), we considered the following cases: $\kappa_1 = 1$ and $\kappa_2 = 2$, $\kappa_1 = 10$ and $\kappa_2 = 20$, $\kappa_1 = 50$ and $\kappa_2 = 100$, $\kappa_1 = 1$ and $\kappa_2 = 10$, $\kappa_1 = 1$ and $\kappa_2 = 100$ (see Figures 6.1–6.5.) In all these problems, the SP_2 solution agrees very well with the full radiative transfer solution, and the GSP_2 solution is significantly closer to these than is the Rosseland solution.

We also studied a more realistic problem with eight frequency bands. The absorption coefficients were chosen according to Table 6.1; these data have been used for practical simulations of annealing in glass. The edge of the opaque part of the spectrum is located at the wavelength $\lambda = 7\mu m$, thus giving $\nu_1 = c/\lambda_1 = 4.28 \cdot 10^{13} s^{-1}$.

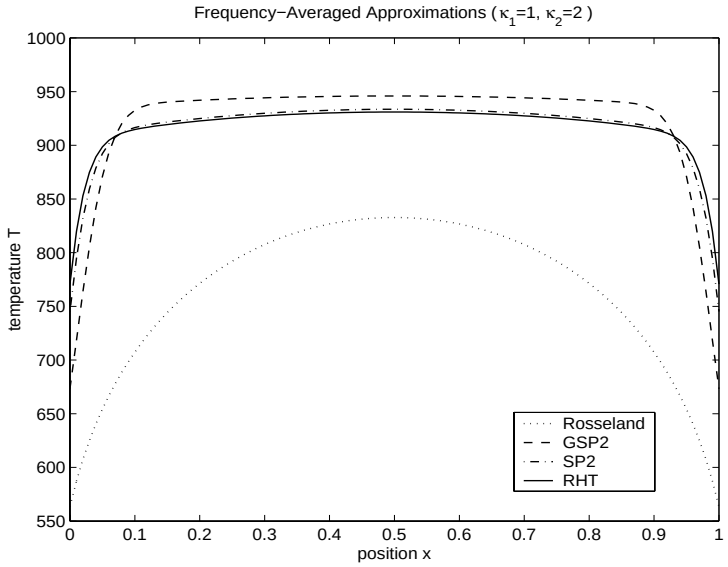


FIG. 6.1. *Comparison of the different approximations for* $\kappa_1 = 1$ *and* $\kappa_2 = 2$.

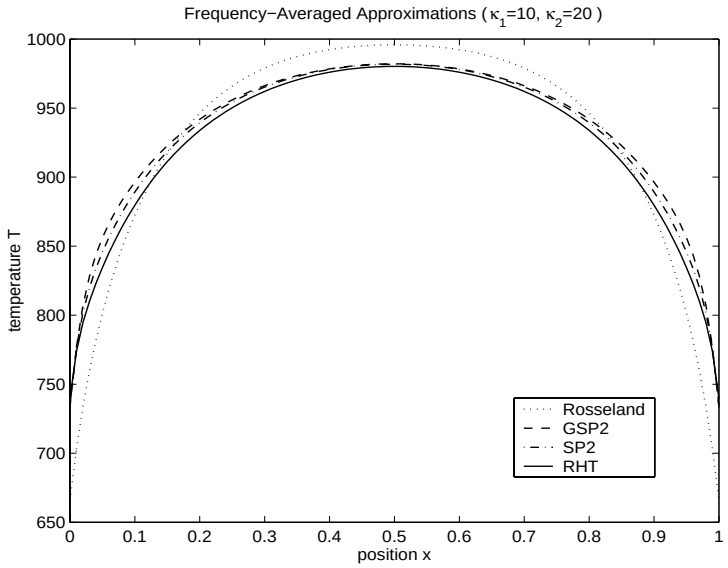


FIG. 6.2. *Comparison of the different approximations for* $\kappa_1 = 10$ *and* $\kappa_2 = 20$.

Figure 6.6 shows the frequency-averaged GSP_2 solution in comparison with the full RHT solution, the (grey) Rosseland solution, and the (frequency-dependent) SP_2 solution. Again, the GSP_2 solutions are in much closer agreement with the full radiative transfer solution and the SP_2 solution than is the Rosseland solution.

In additional to the GSP_2 solution in Figure 6.6, averaging was also done partially with respect to certain frequency bands as explained above. In the following two figures the results of partial averaging are plotted. Figure 6.7 shows results where averaging is done over the first two and the last six of the frequency bands, thus

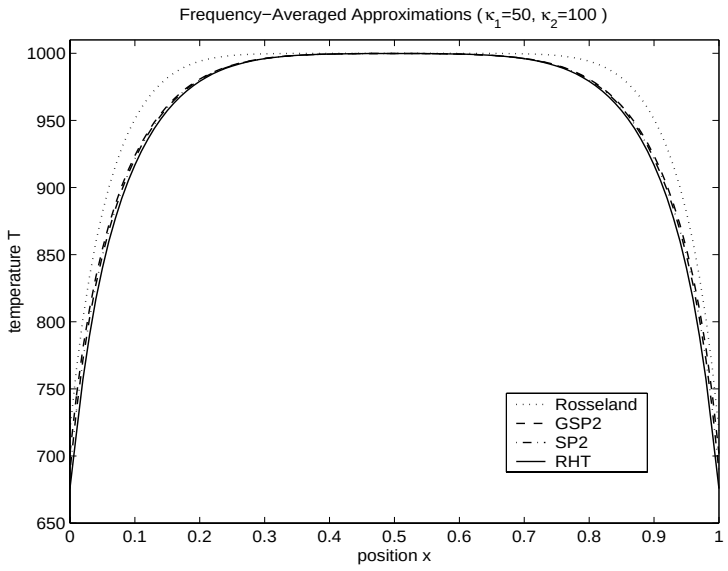


FIG. 6.3. *Comparison of the different approximations for* $\kappa_1 = 50$ *and* $\kappa_2 = 100$.

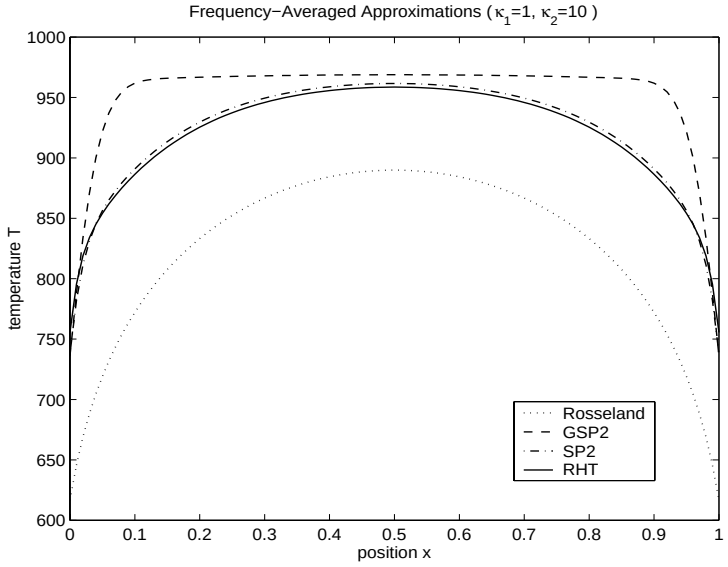


FIG. 6.4. *Comparison of the different approximations for* $\kappa_1 = 1$ *and* $\kappa_2 = 10$.

resulting in a 2-band model. Figure 6.8 shows results with averaging over bands 1 and 2, bands 3 and 4, and bands 5 to 8, resulting in a 3-band model. In both cases very good agreement of the averaged solutions with the radiative heat transfer solution can be observed. Research on how to choose the frequency bands for partial averaging in an optimal way is referred to future work.

However, we do observe, for each of the above problems, that when the GSP_2 solution disagrees with the full radiative transfer solution, the main region of disagreement is in the outer boundary layers; in these cases, the GSP_2 boundary layers

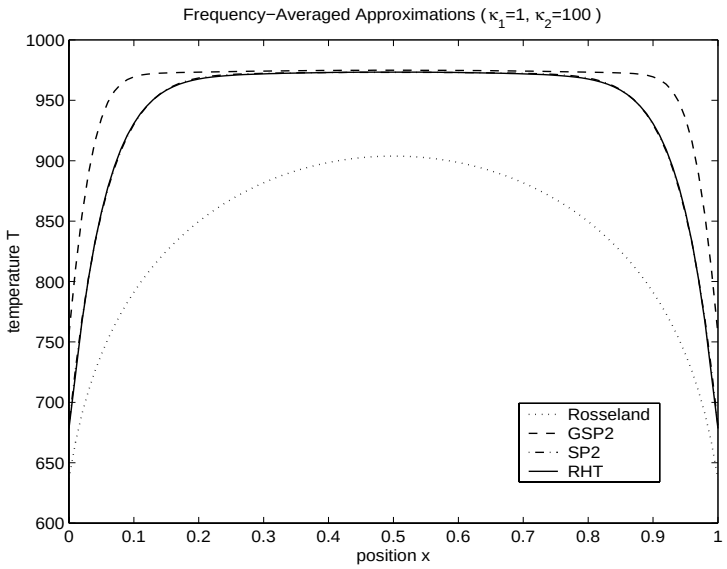


FIG. 6.5. *Comparison of the different approximations for* $\kappa_1 = 1$ *and* $\kappa_2 = 100$.

TABLE 6.1

Opacities of the eight-band model for glass. The bands are defined by wavelength intervals (data kindly provided by Fraunhofer ITWM Kaiserslautern).

Band i	λ_i $[\mu m]$	λ_{i+1} $[\mu m]$	κ_i $[1/m]$
1	0	0.20	0.40
2	0.20	3.00	0.50
3	3.00	3.50	7.70
4	3.50	4.00	15.45
5	4.00	4.50	27.98
6	4.50	5.50	267.98
7	5.50	6.00	567.32
8	6.00	7.00	7136.06
–	7.00	∞	opaque

are steeper than the those in the exact radiative transfer solution. The Rosseland solution, on the other hand, tends to underpredict the steepness of the boundary layers. The more expensive partially averaged solutions generally provide much more accurate estimates of the temperature in the outer boundary layers. At this time we are not certain whether these trends apply consistently in all problems.

Finally, Table 6.2 below displays a run-time comparison of the different models for the 8-band problem, with data measured on a PC with the AMD-K6 200 processor, running MATLAB 5 under Linux 2.2. The Rosseland solution is (as expected) the least expensive, the GSP_2 solution is only slightly more expensive, the frequency-dependent SP_2 calculation is significantly more expensive, and the frequency-and-angle-dependent radiative transfer calculation is by far the most expensive of these simulations. The computation time for the partially averaged models is proportional to the number of equations (number of frequency bands plus temperature equation) used in the models. It ranges from the computation time needed for GSP_2 up to the time needed for SP_2.

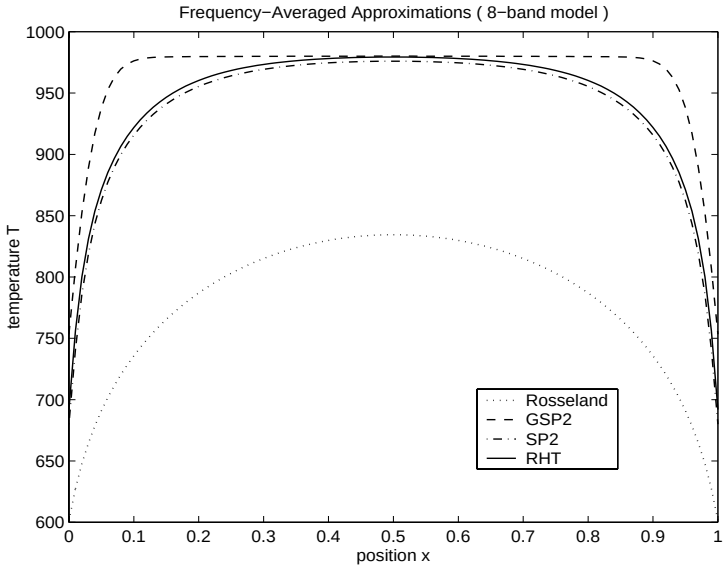


FIG. 6.6. *Numerical results for the temperature of the averaged models and the RHT model in the case of eight frequency bands.*

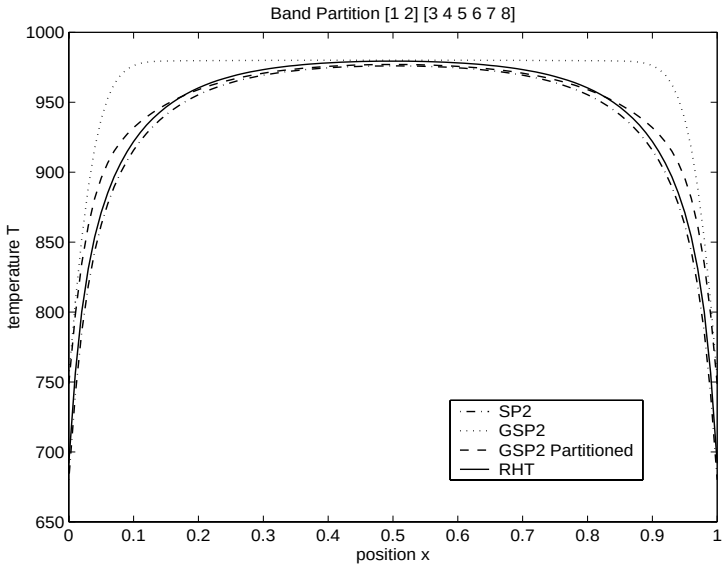


FIG. 6.7. *Numerical results for the temperature with partial averaging of eight frequency bands using two averaged bands.*

TABLE 6.2
Computational costs of the Rosseland, GSP_2*, and* SP_2 *approximations, and the full transport equation, using eight frequency bands.*

	Rosseland	GSP_2	SP_2	RHT
flops $[\times 10^6]$	3.7	4.1	36.8	417.0
time $[sec]$	52.0	54.6	220.4	6423.8

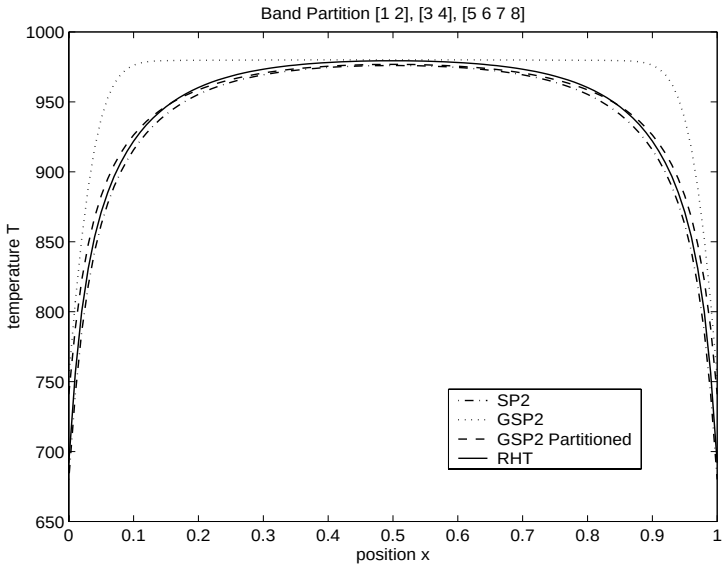


FIG. 6.8. *Numerical results for the temperature with partial averaging of eight frequency bands using three averaged bands.*

7. Conclusions. Summarizing our numerical results, we conclude that the GSP_2 approximation yields significantly more accurate temperatures in the interior of the physical system than does the Rosseland approximation. In the important outer boundary layers, where thermal stresses can cause cracking of glass, the GSP_2 solution is often extremely accurate, and when not, it seems to overpredict the steepness of the boundary layers. Thus, annealing schedules based on the GSP_2 solution should generally overpredict the time required for the glass to cool. On the other hand, annealing schedules based on the Rosseland model should generally underpredict the cooling time which may lead to cracking. The partially averaged solutions provide very accurate estimates of the temperature in the boundary layers. Of course, the optimal cooling schedule will be obtained using the full radiative transfer solution, but solving this equation is significantly more expensive than solving the grey or frequency-averaged SP_N equations.

We now wish to discuss more thoroughly a point that has already been mentioned in this paper and previously in [7]. Specifically, the asymptotic analysis in this paper and in [7] contains no boundary layer analysis and no expansion of the solution of the underlying radiative transfer problem. Instead, we perform in this paper asymptotic manipulations on the frequency-dependent SP_N equations to obtain simpler equations and boundary conditions. In all cases, we seek (systems of) second-order equations in space that are well conditioned and that behave robustly in the presence of boundary layers. Our formal manipulations (i) assume that spatial derivatives are $\mathcal{O}(1)$ and (ii) do not explicitly account for boundary layers. Also, the underlying SP_N boundary conditions are obtained not by an asymptotic analysis—which seems to be forbiddingly complex—but rather by a variational analysis. In spite of these limitations, the SP_N approximations are remarkably accurate in boundary layers [2], [7], [9]. For the problems considered in this paper, the SP_2 solutions are nearly identical to the radiative transfer solutions. Moreover, the less accurate GSP_2 and the partially averaged SP_2 approximations also perform well in the boundary layers—certainly, they

are much more accurate than the Rosseland approximation. Overall, it is unclear why this favorable result should hold. There may be processes at work, not fully understood by the authors, that allow the approach taken in this paper to yield accurate results in boundary layers, even though the underlying asymptotic scaling never explicitly takes these layers into account.

Because the GSP_2 and the partially averaged SP_2 equations capture boundary layer effects, extra care must be taken in selecting spatial grids for them. In particular, spatial grids used for GSP_2 and partially averaged SP_2 must be finer in the boundary layer region than the grids used Rosseland's approximation in order to resolve the boundary layers that are not present in the Rosseland solution.

This discussion of boundary layers leads an interesting question: if the main deficiency of the Rosseland approximation occurs in boundary layer regions, then would it be preferable to develop an approximate theory employing the Rosseland equation in the interior of the system, coupled to one-dimensional boundary layer solutions that are valid at the outer boundary (or interface) parts of the system? (For example, see [1] or [4] and the references cited therein.) The conceptual advantages of this approach are that the underlying Rosseland equation is simpler to solve than the coupled GSP_2 equations, and the one-dimensional boundary layer solutions will probably contain more accurate transport physics than the GSP_2 equations.

The methodology adopted in the present paper is to explore the gains of accuracy that are attainable from the SP_2 and GSP_2 approximation to the RHT equations. A significant advantage of this approach is that differential (diffusion) equations are obtained having a structure that is very similar to equations already implemented in industrial (diffusion) codes. Thus the implementation of the (G)SP_2 equations in these codes is relatively straightforward. (This advantage of SP_N approximations is widely recognized in the nuclear engineering community.) A second advantage is that the GSP_2 equations are only slightly more expensive to solve than the Rosseland equation (see Table 6.2), while the GSP_2 solutions are significantly more accurate than the classic Rosseland solution. A third advantage is that the treatment of boundary layers in the GSP_2 equations is "natural"; the boundary layers exist directly within the GSP_2 equations and do not have to be "matched" to the solution of a (Rosseland) diffusion problem. (This is particularly important because the underlying problem is nonlinear.)

However, we must nonetheless acknowledge the fact that when the GSP_2 solutions disagree with the exact solution, the disagreement occurs mainly in the boundary layer regions. More precisely, the GSP_2 approximation is not generally capable of the accuracy of the SP_2 equations—which are accurate in the boundary layers. It is possible that some of the physics and accuracy that are lost in collapsing the SP_2 equations down to the GSP_2 equations could be regained by including boundary layer solutions into the approximate GSP_2 solution, similar to the procedure in [1]. This would add to the complexity of the method, and to the difficulty of implementation, but it could well raise the accuracy of the overall solution. This is a significant possibility for future research on this problem.

Finally, in developing a grey approximation to the SP_N equations, we also developed (i) an alternative grey $\mathcal{O}(\varepsilon^3)$ approximation to the SP_2 equations and (ii) a grey $\mathcal{O}(\varepsilon^3)$ approximation to the SP_3 equations. However, numerical simulations showed that these approximations are less accurate than the GSP_2 approximation, so it did not seem appropriate to discuss them in detail here. As discussed in the previous paragraphs, there is room for improvement in the GSP_2 approximation developed in this paper, particularly in the boundary layer regions. For example, an approxima-

tion that is more accurate (formally $\mathcal{O}(\varepsilon^4)$ rather than $\mathcal{O}(\varepsilon^3)$) in the boundary layers would certainly be desirable. However, this advance must await fresh insights and future work.

REFERENCES

[1] G. Bal, *Transport through diffusive and nondiffusive regions, embedded objects, and clear layers*, SIAM J. Appl. Math., 62 (2002), pp. 1677–1697.

[2] P. Brantley and E. Larsen, *The simplified P_3 approximation*, Nucl. Sci. Eng., 134 (2000), p. 1.

[3] R. Howell and J. Siegel, *Thermal Radiation Heat Transfer*, 3rd ed., Taylor & Francis, London, 1992.

[4] A. Klar and C. Schmeiser, *Numerical passage from radiative heat transfer to nonlinear diffusion models*, Math. Models Methods Appl. Sci., 11 (2001), pp. 749–767.

[5] A. Klar and N. Siedow, *Boundary layers and domain decomposition for radiative heat transfer and diffusion equations: Applications to glass manufacturing processes*, European J. Appl. Math., 9 (1998), pp. 351–372.

[6] E. Larsen, G. Pomraning, and V. Badham, *Asymptotic analysis of radiative transfer problems*, J. Quant. Spectr. Radiative Transfer, 29 (1983), pp. 285–310.

[7] E. Larsen, G. Thömmes, A. Klar, M. Seaïd, and T. Götz, *Simplified p_n approximations to the equations of radiative heat transfer in glass*, J. Comput. Phys., 183 (2002), pp. 652–675.

[8] M. Modest, *Radiative Heat Transfer*, McGraw–Hill, New York, 1993.

[9] D. Tomašević and E. Larsen, *The simplified P_2 approximation*, Nucl. Sci. Eng., 122 (1996), pp. 309–325.

[10] R. Viskanta, *Radiative heat transfer*, Fortschr. Verfahrenstechnik, 22 (1984), pp. 51–81.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 583–600

SECONDARY CIRCULATION IN GRANULAR FLOW THROUGH NONAXISYMMETRIC HOPPERS*

PIERRE A. GREMAUD†, JOHN V. MATTHEWS‡, AND DAVID G. SCHAEFFER§

Abstract. Jenike's radial solution, widely used in the design of materials-handling equipment, is a similarity solution of steady-state continuum equations for the flow under gravity of granular material through an infinite, right-circular cone. In this paper we study how the geometry of the hopper influences this solution. Using perturbation theory, we compute a first-order correction to the (steady-state) velocity resulting from a small change in hopper geometry, either distortion of the cross section or tilting away from vertical. Unlike for the Jenike solution, all three components of the correction velocity are nonzero; i.e., there is secondary circulation in the perturbed flow.

Key words. granular, similarity solution, perturbation theory

AMS subject classifications. 65L80, 35J65, 76T25

DOI. 10.1137/S0036139903415124

1. Introduction.

In manufacturing industries, raw materials are stored in granular form in a silo, and when needed, they are expected to flow out of the silo under gravity through a hopper. Problems in the discharge process are frequent and expensive; see, e.g., [8]. As demonstrated by a Rand Corporation study [9], these problems are symptomatic of our poor understanding of the behavior of granular materials.[1]

Jenike's radial solution is a central component of silo design. Despite its importance, this solution is subject to many severe restrictions:

1. Granular material is modeled as a continuum, with an ad hoc constitutive law.
2. The flow is assumed to be steady.
3. The flow domain, a mathematical idealization, is an infinite cone, given in spherical polar coordinates by the formula

$$\{(r, \theta, \phi) : 0 < r < \infty,\ 0 \le \theta < \theta_w\} \qquad (\theta_w = \text{constant}).$$

4. Only similarity solutions are considered.

*Received by the editors July 7, 2003; accepted for publication (in revised form) July 22, 2003; published electronically December 31, 2003.
http://www.siam.org/journals/siap/64-2/41512.html

†Department of Mathematics and Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8205 (gremaud@math.ncsu.edu). The research of this author was partially supported by the Army Research Office through grant DAAD19-99-1-0188 and by the National Science Foundation through grant DMS-9818900.

‡Department of Mathematics, Duke University, Durham, NC 27708-0320 (jvmatthe@math.duke.edu). The research of this author was partially supported by the National Science Foundation through grant DMS-9983320.

§Department of Mathematics and Center for Nonlinear and Complex Systems, Duke University, Durham, NC 27708-0320 (dgs@math.duke.edu). The research of this author was partially supported by the National Science Foundation through grant DMS-9803305.

[1]The study compared the design output and the actual output of a total of 60 manufacturing plants in various industries, 22 that were based primarily on liquids-processing technology and 38 on solids-processing technology. On average, the liquids-processing plants produced at 84% of design capacity while the solids-processing plants produced at only 63% of design capacity. To quote Merrow [9], "In economic terms, the difference between 63% of design and 84% is very large. It implies a capital cost per unit of output about one-third higher for the solids-processing plants, on the basis of poor performance alone. In addition, poor performance is inevitably associated with higher operating and maintenance costs per unit of product." Moreover, the standard deviation of the solids-processing data set was much greater, indicative of our difficulties in predicting the behavior of granular solids.

In this paper we relax restrictions 3 and 4 partly. Specifically, we generalize the domain to an infinite *pyramidal* hopper described by the inequality

$$0 \le \theta < \theta_w + \epsilon \cos m\phi, \tag{1.1}$$

where ϵ is a small parameter and m is a positive integer. Assuming a perturbation series

$$v^{(0)} + \epsilon v^{(1)} + \cdots$$

for the flow velocity in the domain (1.1), where $v^{(0)}$ is Jenike's solution, we derive a linear PDE for the first-order correction $v^{(1)}$. The r-dependence of $v^{(1)}$ still has similarity form, and the ϕ-dependence may be handled by separation of variables. In this way we reduce solving the PDE for $v^{(1)}$ to solving a two-point boundary problem on the interval $0 < \theta < \theta_w$.

In Jenike's solution, only the radial component $v_r^{(0)}$ of the velocity is nonzero. By contrast, all three components of the correction velocity $v^{(1)}$ are nonzero. In other words, *distortion of the conical domain leads to secondary circulation.* For example, in Figure 5.1 below, the flow in the θ, ϕ-directions is shown for a circular hopper that is tilted slightly to the right, and in Figure 5.2, for a slightly distorted vertical hopper.

Circulation was previously observed by Williams and Rege in discrete element simulations of granular systems [11], [13]. While a connection between such time-dependent, discrete simulations and the steady-state continuum theory below is unclear, the similarities are uncanny. Both find a secondary circulation in essentially two-dimensional granular systems undergoing a uniform compression.

The outline of the paper is as follows. In section 2, the governing equations are recalled together with Jenike's construction of similarity solutions in conical domains. For nonaxisymmetric domains of the type (1.1), the problem is then linearized about Jenike's solution in section 3. The resulting system is discretized in section 4. Numerical results and discussion are offered in section 5.

2. The model.

2.1. Governing equations and boundary conditions. The unknowns are the 3-component velocity vector v, the 3×3 symmetric stress tensor T, and a scalar plasticity coefficient λ. (The density ρ is a constant.) In total, there are $3 + 6 + 1 = 10$ unknown functions. In writing the equations for these variables, we need the strain rate tensor $V = -1/2(\nabla v + \nabla v^T)$ and the deviatoric part of the stress tensor $\operatorname{dev} T = T - \frac{1}{3} (\operatorname{tr} T)\, \mathrm{I}$. Note the sign convention: V measures the *compression* rate of the material; analogously, positive eigenvalues of T correspond to *compressive* stresses. This sign convention reflects the fact that granular materials disintegrate under tensile stresses.

Following [12], we require that these variables satisfy

$$\nabla \cdot T = \rho\, g, \tag{2.1}$$

$$V = \lambda \operatorname{dev} T, \tag{2.2}$$

$$|\operatorname{dev} T|^2 = 2s^2 (\operatorname{tr} T/3)^2, \tag{2.3}$$

where g is the (vector) acceleration of gravity, $|\cdot|$ denotes the Frobenius norm

$$|T|^2 = \sum_{i,j=1}^{3} T_{ij}^2 = \operatorname{tr} T^2$$

(the latter equality only for symmetric tensors), and $s = \sin\delta$, with δ being the angle of internal friction of the material under consideration (see [10]). Equation (2.1) expresses force balance; i.e., Newton's second law with inertia neglected because the flow is assumed slow; it is equivalent to three scalar equations. Equations (2.2) and (2.3) are constitutive laws, the alignment condition and the von Mises yield condition, respectively; they are equivalent to six and to one scalar equations, respectively. Thus (2.1)–(2.3) is a determined system, ten equations for ten unknowns. Since (2.3) contains no derivatives, this system has a differential-algebraic character. Taking the trace of (2.2), we see that $\operatorname{div} v = -\operatorname{tr} V = 0$; thus, incompressibility is part of the constitutive assumptions. Incidentally, for a solution to be physical, the function λ in (2.2) must satisfy $\lambda \geq 0$ everywhere; otherwise, friction would be adding energy to the system rather than dissipating it. In fact, we want λ to be strictly positive since one of the assumptions underlying the derivation of (2.1)–(2.3) is that material is actually deforming.

We seek solutions of (2.1)–(2.3) in a pyramidal domain, expressed in spherical polar coordinates as

$$\Omega = \{(r, \theta, \phi) : 0 \leq \theta < \mathcal{C}(\phi)\}, \tag{2.4}$$

where $\mathcal{C}$ is a given smooth 2π-periodic function. Such a domain represents a mathematical idealization of a converging hopper, in general, a nonaxisymmetric one.

On the boundary $\partial\Omega = \{(r, \mathcal{C}(\phi), \phi)\}$, wall impenetrability imposes one boundary condition on the velocity; i.e.,

$$v_N = 0, \tag{2.5}$$

where v_N is the normal velocity. Two additional boundary conditions come from Coulomb's law of sliding friction. The surface traction τ—i.e., the force exerted by the wall on the material—is given by

$$\tau_i = \sum_{j=1}^{3} T_{ij} N_j,$$

where N is the unit interior normal to $\partial\Omega$. If the vector τ has normal component τ_N and tangential component $\tau_T = \tau - \tau_N N$, then we require that

$$\tau_T = -\mu_w \, \tau_N (v/|v|), \tag{2.6}$$

where μ_w is the coefficient of friction between the wall and the material. Note the following: (i) If T is positive definite (i.e., if all stresses are compressive), then $\tau_N > 0$. (ii) While τ_N is a scalar, τ_T is effectively a two-component vector; thus, (2.6) is equivalent to two scalar equations. (iii) Because of (2.5), the velocity v is tangential to $\partial\Omega$; we are assuming that $v \neq 0$ at the boundary.

2.2. Jenike's similarity solution. Suppose that the domain (2.4) is axisymmetric; i.e., suppose

$$\Omega = \{(r, \theta, \phi) : 0 \leq \theta < \theta_w\}, \tag{2.7}$$

where θ_w is a constant. In this case Jenike [7] found that (2.1)–(2.3) have solutions that are independent of ϕ and have a similarity dependence on r,

$$v^{(0)}(r, \theta) = r^{-2} \, \hat{v}^{(0)}(\theta), \qquad T^{(0)}(r, \theta) = r \, \hat{T}^{(0)}(\theta). \tag{2.8}$$

(Here and below, a hat above a variable indicates a function that depends on θ alone.) Moreover, only the radial component of velocity is nonzero; i.e., $v_\theta^{(0)} = v_\phi^{(0)} = 0$. Similarly, $T_{r\phi}^{(0)} = T_{\theta\phi}^{(0)} = 0$. Indeed, all components of T can be expressed in terms of two scalar variables, the so-called Sokolovskii variables [10]: the mean stress $p^{(0)} = \operatorname{tr} T^{(0)}/3$ and an angle ψ; specifically,

$$\operatorname{dev} T^{(0)} = s\, p^{(0)} \begin{bmatrix} -\frac{2}{\sqrt{3}}\cos 2\psi & -\sin 2\psi & 0 \\ -\sin 2\psi & \frac{1}{\sqrt{3}}\cos 2\psi & 0 \\ 0 & 0 & \frac{1}{\sqrt{3}}\cos 2\psi \end{bmatrix}, \tag{2.9}$$

where $p^{(0)} = r\hat{p}^{(0)}$ and the function ψ, like $\hat{p}^{(0)}$, depends only on θ.

The boundary conditions (2.5), (2.6) may be written more explicitly when Ω is axisymmetric. Equation (2.5) reduces to

$$v_\theta = 0. \tag{2.10}$$

Let us decompose the vector equation (2.6) into a direction and a magnitude. Regarding the direction, the vectors τ_T and v are parallel if

$$T_{r\theta} v_\phi - T_{\theta\phi} v_r = 0. \tag{2.11}$$

Jenike's solution satisfies both (2.10) and (2.11) trivially. The two sides of (2.6) have equal magnitude if

$$T_{r\theta} + \mu_w T_{\theta\theta} = 0. \tag{2.12}$$

We briefly summarize the construction of Jenike's solution, referring to [12] for more details. The ansatz (2.9) arranges that (2.3) holds automatically. On substitution into (2.1), we obtain a first-order 2×2 system of ODEs for $\hat{p}^{(0)}$ and ψ. This system has a regular singular point at $\theta = 0$, and one boundary condition comes from requiring that the solution be regular there; the other boundary condition comes from (2.12). Thus the stresses are determined as the solution of a two-point boundary-value problem. (In axial symmetry, the stress equations decouple from the velocity.) Once the stresses are known, (2.2) reduces to a linear first-order ODE for $\hat{v}_r^{(0)}$. The velocity is determined only up to a multiplicative constant, but the normalization of the velocity will scale out of the calculations below.

Incidentally, for Jenike's solution the plasticity coefficient λ in (2.2), which cancels out in the derivation of the equation for $\hat{v}_r^{(0)}$, has the form

$$\lambda^{(0)}(r, \theta) = r^{-4} \hat{\lambda}^{(0)}(\theta).$$

Using (2.2), the function $\hat{\lambda}^{(0)}$ may be determined from $\hat{v}_r^{(0)}$.

3. Linearized analysis for a nearly axisymmetric domain.

3.1. Derivation of linearized differential equations. Equations (2.1)–(2.3), a 10×10 nonlinear DAE system that is elliptic in the sense of Agmon, Douglis, and Nirenberg [1], present formidable mathematical and numerical challenges. In this paper, we consider a simplified problem that prepares the way for computations with the full problem on a general domain.

Suppose the function $\mathcal{C}$ specifying the boundary of Ω in (2.4) has the expansion

$$\mathcal{C}(\phi) = \theta_w + \epsilon \cos(m\,\phi) + \mathcal{O}(\epsilon^2), \tag{3.1}$$

where m is a positive integer. For example, a slightly tilted (circular) cone admits such a representation with $m = 1$, where ϵ measures the angle of tilt; likewise for a (vertical) pyramidal hopper having a slightly elliptical cross section, with $m = 2$.

An expansion of the solution

$$v = v^{(0)} + \epsilon v^{(1)} + \mathcal{O}(\epsilon^2), \qquad T = T^{(0)} + \epsilon T^{(1)} + \mathcal{O}(\epsilon^2) \tag{3.2}$$

is sought, where $v^{(0)}, T^{(0)}$ are equal to Jenike's radial solution; see (2.8). Substituting (3.2) into (2.1)–(2.3), we derive the equations for the first-order perturbation

$$\nabla \cdot T^{(1)} = 0, \tag{3.3}$$

$$V^{(1)} = \lambda^{(1)}\,\mathrm{dev}T^{(0)} + \lambda^{(0)}\mathrm{dev}T^{(1)}, \tag{3.4}$$

$$\mathrm{tr}(\mathrm{dev}T^{(0)}\,\mathrm{dev}T^{(1)}) = 2\,s^2\,p^{(0)}p^{(1)}, \tag{3.5}$$

where $p^{(i)} = \mathrm{tr}\,T^{(i)}/3,\ i = 0, 1$, are the mean stresses.

The correction velocity $v^{(1)}$ has the same r-dependence as the Jenike solution (although all three components of $v^{(1)}$ may be nonzero), and its ϕ-dependence can be obtained through separation of variables. Indeed, suppose each component of $v^{(1)}$ has the form

$$v_j^{(1)} = r^{-2}\,\hat{v}_j^{(1)}(\theta)\,\mathrm{trig}\,m\phi, \tag{3.6}$$

where $\mathrm{trig}\,m\phi$ denotes either $\cos m\phi$ or $\sin m\phi$. In order to satisfy the appropriately modified version of the boundary condition (2.10) on the perturbed domain, $v_\theta^{(1)}$ will have to be in phase with (3.1); i.e., we need

$$v_\theta^{(1)} = r^{-2}\,\hat{v}_\theta^{(1)}(\theta)\cos m\phi.$$

It is readily seen that if

$$v_r^{(1)} = r^{-2}\,\hat{v}_r^{(1)}(\theta)\cos m\phi \qquad \text{and} \qquad v_\phi^{(1)} = r^{-2}\,\hat{v}_\phi^{(1)}(\theta)\sin m\phi,$$

then all terms in

$$\nabla \cdot v^{(1)} = \partial_r v_r^{(1)} + 2r^{-1}v_r^{(1)} + r^{-1}\partial_\theta v_\theta^{(1)} + r^{-1}\cot\theta v_\theta^{(1)} + r^{-1}\csc\theta\partial_\phi v_\phi^{(1)} \tag{3.7}$$

are proportional to $r^{-3}\cos m\phi$; i.e., variables separate in the equation $\nabla \cdot v = 0$.

Tables 3.1–3.3 help systematize the elimination of ϕ-dependence in (3.3)–(3.5) with separation of variables. The appropriate r- and ϕ-dependencies for the scalar $p^{(1)}$, for the vector $v^{(1)}$, and for the tensor $T^{(1)}$ are indicated in Table 3.1. (Note that symmetric 3×3 tensors are represented as vectors in $\mathbf{R}^6$, the components being enumerated in the order shown.) In Table 3.2 we record, for the reader's convenience, the expressions in spherical coordinates for four differential operators that occur in these equations.

The main point, which makes separation of variables work in this problem, is that the θ-dependent part of each of these linear operators is given by

$$\widehat{\nabla p} = (g_1\partial_\theta + g_0)\hat{p}, \tag{3.8a}$$

$$\widehat{\nabla \cdot v} = (d_1^T\partial_\theta + d_0^T)\hat{v}, \tag{3.8b}$$

$$\hat{V} = -(G_1\partial_\theta + G_0)\hat{v}, \tag{3.8c}$$

$$\widehat{\nabla \cdot T} = (D_1\partial_\theta + D_0)\hat{T}, \tag{3.8d}$$

TABLE 3.1
The r- and ϕ-dependence of scalars, vectors, and tensors in separation of variables.

Scalars :	$p = r\,\hat{p}(\theta)\cos m\phi$		
Vectors :	$v = \frac{1}{r^2}\begin{bmatrix} \hat{v}_r(\theta)\cos m\phi \\ \hat{v}_\theta(\theta)\cos m\phi \\ \hat{v}_\phi(\theta)\sin m\phi \end{bmatrix}$	*Tensors* :	$T = r\begin{bmatrix} \hat{T}_{rr}(\theta)\cos m\phi \\ \hat{T}_{r\theta}(\theta)\cos m\phi \\ \hat{T}_{\theta\theta}(\theta)\cos m\phi \\ \hat{T}_{r\phi}(\theta)\sin m\phi \\ \hat{T}_{\theta\phi}(\theta)\sin m\phi \\ \hat{T}_{\phi\phi}(\theta)\cos m\phi \end{bmatrix}$

TABLE 3.2
Differential operators in spherical coordinates.

$$\nabla p = [\,\partial_r p,\ r^{-1}\partial_\theta p,\ r^{-1}\csc\theta\partial_\phi p\,]^T$$

$$\nabla\cdot v = \partial_r v_r + 2r^{-1}v_r + r^{-1}\partial_\theta v_\theta + r^{-1}\cot\theta v_\theta + r^{-1}\csc\theta\partial_\phi v_\phi$$

$$V = \begin{bmatrix} V_{rr} \\ V_{r\theta} \\ V_{\theta\theta} \\ V_{r\phi} \\ V_{\theta\phi} \\ V_{\phi\phi} \end{bmatrix} = -\begin{bmatrix} \partial_r v_r \\ \frac{1}{2}\left(r^{-1}\partial_\theta v_r - r^{-1}v_\theta + \partial_r v_\theta\right) \\ r^{-1}\left(v_r + \partial_\theta v_\theta\right) \\ \frac{1}{2}\left(r^{-1}\csc\theta\,\partial_\phi v_r - r^{-1}v_\phi + \partial_r v_\phi\right) \\ \frac{1}{2}r^{-1}\left(\partial_\theta v_\phi - \cot\theta\, v_\phi + \csc\theta\partial_\phi v_\theta\right) \\ r^{-1}\left(v_r + \cot\theta\, v_\theta + \csc\theta\,\partial_\phi v_\phi\right) \end{bmatrix}$$

$$\nabla\cdot T = \begin{bmatrix} \partial_r T_{rr} + r^{-1}\csc\theta\,\partial_\phi T_{r\phi} + r^{-1}\partial_\theta T_{r\theta} + r^{-1}(2T_{rr} - T_{\phi\phi} - T_{\theta\theta} + T_{r\theta}\cot\theta) \\ \partial_r T_{r\theta} + r^{-1}\csc\theta\,\partial_\phi T_{\theta\phi} + r^{-1}\partial_\theta T_{\theta\theta} + r^{-1}\left(3T_{r\theta} + (T_{\theta\theta} - T_{\phi\phi})\cot\theta\right) \\ \partial_r T_{r\phi} + r^{-1}\csc\theta\,\partial_\phi T_{\phi\phi} + r^{-1}\partial_\theta T_{\theta\phi} + r^{-1}(3T_{r\phi} + 2T_{\theta\phi}\cot\theta) \end{bmatrix}$$

where $g_1, g_0, \ldots, D_0$ are the matrices given in Table 3.3.

The calculation needed to verify (3.8b) was described above; the other equations may be verified similarly. Incidentally, (3.8a) may be derived by substituting $T = pI$ in (3.8d), and (3.8b) may be derived by taking the trace of (3.8c).

With this notation, (3.3)–(3.5) reduces to a system of ODEs in θ,

$$(D_1\partial_\theta + D_0)\hat{T}^{(1)} = 0, \tag{3.9}$$

$$-(G_1\partial_\theta + G_0)\hat{v}^{(1)} = \hat{\lambda}^{(1)}\,\mathrm{dev}\hat{T}^{(0)} + \hat{\lambda}^{(0)}\mathrm{dev}\hat{T}^{(1)}, \tag{3.10}$$

$$\mathrm{tr}(\mathrm{dev}\hat{T}^{(0)}\,\mathrm{dev}\hat{T}^{(1)}) = 2\,s^2\,\hat{p}^{(0)}\hat{p}^{(1)}. \tag{3.11}$$

Recalling the representation of symmetric tensors as 6-component vectors, we observe that the left-hand side of (3.11) may be rewritten as an inner product

$$\mathrm{tr}\left(\mathrm{dev}\hat{T}^{(0)}\,\mathrm{dev}\hat{T}^{(1)}\right) = \mathrm{dev}\hat{T}^{(0)T}\mathcal{M}\,\mathrm{dev}\hat{T}^{(1)},$$

where $\mathcal{M}$ is the 6×6 matrix

$$\mathcal{M} = \mathrm{diag}(1,2,1,2,2,1);$$

thus we may rewrite (3.11) as

$$\mathrm{dev}\hat{T}^{(0)T}\mathcal{M}\,\mathrm{dev}\hat{T}^{(1)} = 2\,s^2\,\hat{p}^{(0)}\hat{p}^{(1)}. \tag{3.12}$$

TABLE 3.3
Matrices in (3.8).

$$g_1 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T \qquad g_0 = \begin{bmatrix} 1 & 0 & \frac{m}{\sin\theta} \end{bmatrix}^T$$

$$d_1 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T \qquad d_0 = \begin{bmatrix} 0 & \cot\theta & \frac{m}{\sin\theta} \end{bmatrix}^T$$

$$G_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1/2 \\ 0 & 0 & 0 \end{bmatrix} \qquad G_0 = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -3/2 & 0 \\ 1 & 0 & 0 \\ -\frac{m}{2\sin\theta} & 0 & -3/2 \\ 0 & -\frac{m}{2\sin\theta} & -\frac{\cot\theta}{2} \\ 1 & \cot\theta & \frac{m}{\sin\theta} \end{bmatrix}$$

$$D_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \qquad D_0 = \begin{bmatrix} 3 & \cot\theta & -1 & \frac{m}{\sin\theta} & 0 & -1 \\ 0 & 4 & \cot\theta & 0 & \frac{m}{\sin\theta} & -\cot\theta \\ 0 & 0 & 0 & 4 & 2\cot\theta & -\frac{m}{\sin\theta} \end{bmatrix}$$

Let us show that the deviatoric stresses in (3.9), (3.10), (3.12) can be eliminated from these equations to obtain

$$(D_1\partial_\theta + D_0)\left(-\frac{1}{\hat\lambda^{(0)}}(G_1\partial_\theta + G_0)\hat v^{(1)} - \frac{\hat\lambda^{(1)}}{(\hat\lambda^{(0)})^2}\hat V^{(0)}\right) + (g_1\partial_\theta + g_0)\hat p^{(1)} = 0, \tag{3.13}$$

$$(d_1^T\partial_\theta + d_0^T)\hat v^{(1)} = 0, \tag{3.14}$$

where in (3.13)

$$\hat\lambda^{(1)} = -\frac{1}{2s^2}\frac{1}{(\hat p^{(0)})^2\hat\lambda^{(0)}}\,\hat V^{(0)T}\mathcal{M}(G_1\partial_\theta + G_0)\hat v^{(1)} - \frac{\hat\lambda^{(0)}}{\hat p^{(0)}}\,\hat p^{(1)}. \tag{3.15}$$

Equation (3.14) follows on taking the trace of (3.10). Next, we rewrite (3.10) as

$$\mathrm{dev}\hat T^{(1)} = -\frac{1}{\hat\lambda^{(0)}}(G_1\partial_\theta + G_0)\hat v^{(1)} - \frac{\hat\lambda^{(1)}}{(\hat\lambda^{(0)})^2}\hat V^{(0)}, \tag{3.16}$$

where we have eliminated $\mathrm{dev}\hat T^{(0)}$ using the relation $\hat V^{(0)} = \hat\lambda^{(0)}\mathrm{dev}\hat T^{(0)}$—effectively, (2.2) for Jenike's solution. Recalling that $\hat T^{(1)} = \mathrm{dev}\hat T^{(1)} + \hat p^{(1)}I$, we substitute (3.16) into (3.9) to derive (3.13). Similarly, (3.15) follows on substituting (3.16) into (3.12) and rearranging.

As a final simplification, we substitute (3.15) into (3.13), obtaining the linear, homogeneous system of ODEs

$$-(A_2\partial_{\theta\theta} + A_1\partial_\theta + A_0)\hat v^{(1)} + (b_1\partial_\theta + b_0)\hat p^{(1)} = 0, \tag{3.17}$$

$$(d_1^T\partial_\theta + d_0^T)\hat v^{(1)} = 0, \tag{3.18}$$

where, with the definition

$$P = I - \frac{1}{2s^2(\hat p^{(0)})^2(\hat\lambda^{(0)})^2}\hat V^{(0)}\hat V^{(0)T}\mathcal{M},$$

the coefficient matrices are given by

$$
\begin{aligned}
A_2 &= \frac{1}{\hat{\lambda}^{(0)}} D_1 P G_1, \\
A_1 &= \frac{1}{\hat{\lambda}^{(0)}} (D_0 P G_1 + D_1 P G_0) + D_1 \partial_\theta \left(\frac{1}{\hat{\lambda}^{(0)}} P G_1 \right), \\
A_0 &= \frac{1}{\hat{\lambda}^{(0)}} D_0 P G_0 + D_1 \partial_\theta \left(\frac{1}{\hat{\lambda}^{(0)}} P G_0 \right), \\
b_1 &= g_1 + D_1 \frac{\hat{V}^{(0)}}{\hat{p}^{(0)} \hat{\lambda}^{(0)}}, \\
b_0 &= g_0 + (D_1 \partial_\theta + D_0) \frac{\hat{V}^{(0)}}{\hat{p}^{(0)} \hat{\lambda}^{(0)}}.
\end{aligned}
$$

These matrices depend on θ and in fact several are singular as $\theta \to 0$. In Corollary 4.2 below, we show that this system has a six-dimensional solution space.

The combination $\hat{V}^{(0)}/(\hat{p}^{(0)}\hat{\lambda}^{(0)})$, which occurs in various places in the above formulas, admits a convenient representation; i.e., combining (2.2) and (2.9), we deduce that

$$
\frac{1}{\hat{p}^{(0)} \hat{\lambda}^{(0)}} \hat{V}^{(0)} = s \begin{bmatrix} -\frac{2}{\sqrt{3}} \cos 2\psi \\ -\sin 2\psi \\ \frac{1}{\sqrt{3}} \cos 2\psi \\ 0 \\ 0 \\ \frac{1}{\sqrt{3}} \cos 2\psi \end{bmatrix}. \tag{3.19}
$$

The following supplementary information will be needed in section 4.

LEMMA 3.1. *Under the reflection $\theta \mapsto -\theta$, the functions in separation of variables have the parities*

$$\hat{p}(-\theta) = (-1)^m \hat{p}(\theta), \tag{3.20a}$$

$$\hat{v}_r^{(1)}(-\theta) = (-1)^m \hat{v}_r^{(1)}(\theta), \tag{3.20b}$$

$$\hat{v}_\theta^{(1)}(-\theta) = (-1)^{m+1} \hat{v}_\theta^{(1)}(\theta), \tag{3.20c}$$

$$\hat{v}_\phi^{(1)}(-\theta) = (-1)^{m+1} \hat{v}_\phi^{(1)}(\theta). \tag{3.20d}$$

Proof. The reflection $\theta \mapsto -\theta$ and the rotation $\phi \mapsto \phi + \pi$ are different representations of the same mapping. Therefore, since p is a scalar,

$$\hat{p}(-\theta) \cos m\phi = \hat{p}(\theta) \cos m(\phi + \pi) = (-1)^m \hat{p}(\theta) \cos m\phi,$$

which proves (3.20a). Equation (3.20b) follows from the same argument since v_r transforms as a scalar under changes in the angular coordinates. Rather than analyze the parities of v_θ and v_ϕ, we prefer an indirect argument. Since $\nabla \cdot v^{(1)}$ is a scalar, $\nabla \cdot v^{(1)}$ has parity $(-1)^m$ under the reflection $\theta \mapsto -\theta$, and on inspecting (3.7), we deduce (3.20c), (3.20d). ∎

Incidentally, although we shall not need that information below, we remark that under this reflection $\hat{T}_{rr}$, $\hat{T}_{\theta\theta}$, $\hat{T}_{\theta\phi}$, and $\hat{T}_{\phi\phi}$ have parity $(-1)^m$, while $\hat{T}_{r\theta}$ and $\hat{T}_{\phi r}$ have parity $(-1)^{m+1}$.

TABLE 3.4
Leading orders in the expansions at $\theta = 0$ of the coefficient matrices of (3.17)–(3.18).

$$
\begin{array}{lcl}
A_2(\theta) & = & \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{5}{6} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} + \mathcal{O}(\theta) \\
A_1(\theta) & = & \theta^{-1} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{5}{6} & \frac{m}{3} \\ 0 & -\frac{m}{3} & \frac{1}{2} \end{bmatrix} + \mathcal{O}(1) \\
A_0(\theta) & = & \theta^{-2} \begin{bmatrix} -\frac{m^2}{2} & 0 & 0 \\ 0 & -\frac{m^2}{2} - \frac{5}{6} & -\frac{4m}{3} \\ 0 & -\frac{4m}{3} & -\frac{5m^2}{6} - \frac{1}{2} \end{bmatrix} + \mathcal{O}(\theta^{-1}) \\
b_1(\theta) & = & (1 + s/\sqrt{3}) \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T + \mathcal{O}(\theta) \\
b_0(\theta) & = & \theta^{-1}(1 + s/\sqrt{3}) \begin{bmatrix} 0 & 0 & -m \end{bmatrix}^T + \mathcal{O}(1) \\
d_1(\theta) & = & \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T \qquad \text{(exactly)} \\
d_0(\theta) & = & \theta^{-1} \begin{bmatrix} 0 & 1 & m \end{bmatrix}^T + \mathcal{O}(1)
\end{array}
$$

3.2. Boundary conditions at the centerline. Equations (3.17), (3.18) have a regular singular point at $\theta = 0$. The leading orders of the coefficient matrices in these equations are given in Table 3.4. This information may be determined *without knowing the Jenike solution explicitly* since, using the fact that $\psi(0) = 0$, we deduce from (3.19) that

$$\frac{\hat{V}^{(0)}}{\hat{p}^{(0)}\hat{\lambda}^{(0)}}(0) = \frac{s}{\sqrt{3}} \begin{bmatrix} -2 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}^T.$$

According to the method of Frobenius [3], equations (3.17), (3.18) admit solutions of the form

$$\hat{v}^{(1)}(\theta) = \theta^{\nu} F(\theta), \qquad \hat{p}^{(1)}(\theta) = \theta^{\nu-1} f(\theta), \tag{3.21}$$

where $F(\theta)$ and $f(\theta)$ are analytic near $\theta = 0$. Suppose the exponent ν is real; if $1 \le \nu$, such a solution is continuous; if $\nu < 0$, it is singular; and if $0 \le \nu < 1$, it is continuous, *provided* $f(0) = 0$.

PROPOSITION 3.2. *There are exactly three linearly independent solutions of* (3.17), (3.18) *of the form* (3.21) *that are continuous at* $\theta = 0$.

Proof. Substitution of (3.21) into (3.17), (3.18) gives an indicial equation with roots

$$\nu = \pm(m+1), \pm m, \pm(m-1). \tag{3.22}$$

If $m \ge 2$, three of the roots of (3.22) are negative and three are positive. The three solutions associated with the positive indices are linearly independent and continuous. (We remark that since the positive roots differ by integers, in principle these solutions

might contain log terms. This would not affect their continuity. Moreover, using Maple we have verified that no such log terms arise. See section 5.2 for more details about the Maple code.)

If $m = 1$, there are four nonnegative roots of (3.22), including zero which is a double root. Because of a log term, the double root zero contributes at most one continuous solution. Again using Maple we have eliminated the various alternative possibilities to show that zero and the two positive roots (3.22) contribute exactly three linearly independent continuous solutions and these do not contain any log terms. □

Incidentally, since the roots of the indicial equation are integers and since no log terms arise, the continuous solutions of the lemma are actually *analytic* near $\theta = 0$.

Note that there are six roots (3.22) of the indicial equation. Therefore (3.17), (3.18) has a six-dimensional solution space. (We also prove this by a different argument in Corollary 4.2.) Thus, according to the proposition, the condition that solutions be regular at $\theta = 0$ is equivalent to three boundary conditions. Therefore, regularity at $\theta = 0$ plus the three boundary conditions (2.5), (2.6) will provide a complete set of boundary conditions.

3.3. Boundary conditions at the hopper wall. We derive the perturbed version of (2.5) in some detail; similar issues arise for (2.6), and we treat the latter equation more succinctly. The calculations are greatly simplified by the fact that we may neglect any quantity that is $\mathcal{O}(\epsilon^2)$. To exploit this simplification efficiently, we temporarily use the notation $F \sim G$ to mean that $F = G + \mathcal{O}(\epsilon^2)$.

Including a prefactor of r^2 to remove all r-dependence from the equation, we may rewrite (2.5) as

$$r^2 v(r, \theta_w + \epsilon \cos m\phi, \phi) \cdot N = 0. \tag{3.23}$$

Because of the perturbation, (3.23) differs from (2.10) in three respects:

- the velocity v contains an additional term, $v \sim v^{(0)} + \epsilon v^{(1)}$;
- the velocity is evaluated at a location shifted by $\epsilon \cos m\phi$;
- the direction of the normal N is changed.

Regarding the first two points, we observe that

$$r^2 v(r, \theta_w + \epsilon \cos m\phi, \phi) \sim \hat{v}^{(0)}(\theta_w) + \epsilon \cos m\phi \, \partial_\theta \hat{v}^{(0)}(\theta_w) + \epsilon \operatorname{trig} m\phi \, \hat{v}^{(1)}(\theta_w),$$

where $\operatorname{trig} m\phi$ equals $\cos m\phi$ or $\sin m\phi$, depending on the component of $\hat{v}^{(1)}$. Regarding the third point, $\partial\Omega$ is the zero set of the function $\theta - \theta_w - \epsilon \cos m\phi$. Taking the gradient of this function, we conclude that the (inward) normal is

$$N \sim \begin{bmatrix} 0 & -1 & -\epsilon \frac{\sin m\phi}{\sin \theta_w} \end{bmatrix}^T.$$

Modulo an $\mathcal{O}(\epsilon^2)$-error, N has unit length. Substituting the previous two equations into (3.23), we deduce that

$$-r^2 v(r, \theta_w + \epsilon \cos m\phi, \phi) \cdot N \sim \hat{v}_\theta^{(0)}(\theta_w) + \epsilon \cos m\phi \left(\partial_\theta \hat{v}_\theta^{(0)}(\theta_w) + \hat{v}_\theta^{(1)}(\theta_w) \right) + \epsilon \frac{\sin m\phi}{\sin \theta_w} \hat{v}_\phi^{(0)}(\theta_w).$$

However, since $v_\theta^{(0)}$ and $v_\phi^{(0)}$ vanish identically for Jenike's solution, the velocity boundary condition for the perturbed problem reduces to

$$\hat{v}_\theta^{(1)}(\theta_w) = 0. \tag{3.24}$$

We turn to the stress boundary condition (2.6). As regards the scalar τ_N in (2.6), we observe that, since $T^{(0)}_{r\phi}$ and $T^{(0)}_{\theta\phi}$ vanish for Jenike's solution,

$$\tau_N = \sum_{i,j=1}^{3} T_{ij} N_i N_j \sim T^{(0)}_{\theta\theta} + \epsilon T^{(1)}_{\theta\theta}. \tag{3.25}$$

The vectors τ_T and v_T in (2.6) lie in a two-dimensional subspace tangent to $\partial\Omega$. Note that the unperturbed tangent space is spanned by the r and ϕ coordinate directions. Even allowing for the perturbation, the two sides of (2.6) will be equal iff their r- and ϕ-components are equal; in symbols, iff

$$\begin{bmatrix} \tau_{Tr} \\ \tau_{T\phi} \end{bmatrix} = -\frac{\mu_w \tau_N}{|v|} \begin{bmatrix} v_r \\ v_\phi \end{bmatrix}.$$

This equality will hold iff (i) the two sides of the equation are parallel vectors and (ii) the first components of the two sides are equal; again, in symbols, iff

$$\tau_{Tr} v_\phi - \tau_{T\phi} v_r = 0 \qquad \text{and} \tag{3.26}$$

$$\tau_{Tr} + \mu_w \tau_N (v_r/|v|) = 0. \tag{3.27}$$

Regarding v, it is clear that

$$\begin{bmatrix} v_r \\ v_\phi \end{bmatrix} \sim \begin{bmatrix} v^{(0)}_r \\ 0 \end{bmatrix} + \epsilon \begin{bmatrix} v^{(1)}_r \\ v^{(1)}_\phi \end{bmatrix}. \tag{3.28}$$

Regarding $\tau_T = \tau - \tau_N N$, we claim that

$$\begin{bmatrix} \tau_{Tr} \\ \tau_{T\phi} \end{bmatrix} \sim -\begin{bmatrix} T^{(0)}_{r\theta} \\ 0 \end{bmatrix} - \epsilon \begin{bmatrix} T^{(1)}_{r\theta} \\ T^{(1)}_{\theta\phi} \end{bmatrix}. \tag{3.29}$$

Verifying this claim is straightforward except that, in analyzing the second component, one must invoke the fact that Jenike's solution satisfies $T^{(0)}_{\theta\theta} = T^{(0)}_{\phi\phi}$. On substituting (3.28) and (3.29) into (3.26), we obtain the equation

$$\epsilon \left(T^{(0)}_{r\theta} v^{(1)}_\phi - v^{(0)}_r T^{(1)}_{\theta\phi} \right) = 0 \qquad \text{at } \theta = \theta_w + \epsilon \cos m\phi.$$

The difference between evaluating this expression at $\theta = \theta_w$ and at the perturbed location is $\mathcal{O}(\epsilon^2)$. Removing the r-dependence (proportional to r) and the ϕ-dependence (proportional to $\sin m\phi$) from this equation, we obtain the first stress boundary condition for the perturbed problem:

$$\left(\hat{T}^{(0)}_{r\theta} \hat{v}^{(1)}_\phi - \hat{v}^{(0)}_r \hat{T}^{(1)}_{\theta\phi} \right) = 0 \qquad \text{at } \theta = \theta_w. \tag{3.30}$$

Regarding (3.27), we claim that

$$|v| = \sqrt{v_r^2 + v_\theta^2 + v_\phi^2} \sim |v_r|.$$

Indeed, it is clear from (3.28) that the contribution of v_ϕ to $|v|$ is $\mathcal{O}(\epsilon^2)$, and by (3.24) the contribution of v_θ to $|v|$ is $\mathcal{O}(\epsilon^4)$. Thus, $v_r/|v| \sim -1$. Substituting (3.25) and (3.29) into (3.27), we obtain the condition

$$(T^{(0)}_{r\theta} + \epsilon T^{(1)}_{r\theta}) + \mu_w (T^{(0)}_{\theta\theta} + \epsilon T^{(1)}_{\theta\theta}) \sim 0 \qquad \text{at } \theta = \theta_w + \epsilon \cos m\phi.$$

By (2.12), $T^{(0)}_{r\theta} + \mu_w T^{(0)}_{\theta\theta}$ vanishes at $\theta = \theta_w$, but at the perturbed location these terms make an $\mathcal{O}(\epsilon)$-contribution. Allowing for this contribution and eliminating the r- and ϕ-dependence, we derive the second stress boundary condition for the perturbed problem:

$$\hat{T}^{(1)}_{r\theta} + \mu_w \hat{T}^{(1)}_{\theta\theta} = -\partial_\theta \left(\hat{T}^{(0)}_{r\theta} + \mu_w \hat{T}^{(0)}_{\theta\theta} \right) \qquad \text{at } \theta = \theta_w. \tag{3.31}$$

We have put the inhomogeneous term, which does not involve the perturbation $T^{(1)}$, on the right side of the equation. (By contrast, (3.30) and (3.24) are homogeneous.)

It is noteworthy that the perturbed boundary conditions (3.24), (3.30), (3.31) resemble (2.10), (2.11), (2.12) rather closely.

4. Numerical approximation of the two-point boundary-value problem. The coefficients in (3.17), (3.18) depend on the zeroth-order solution discussed in section 2.2. This solution can be found numerically without difficulty; see, e.g., [6], where a shooting method is used, or [10]. We will consider the zeroth-order solution as given, and we will focus on the corrections $\hat{v}^{(1)}$ and $\hat{p}^{(1)}$.

To simplify the notation before discretization, we set

$$w = \hat{v}^{(1)}, \quad z = \frac{d}{d\theta}\hat{v}^{(1)}, \quad \text{and} \quad q = p^{(1)}$$

and rewrite equations (3.17), (3.18) as a first-order system

$$\begin{bmatrix} \mathrm{I} & 0 & 0 \\ 0 & -A_2 & b_1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w' \\ z' \\ q' \end{bmatrix} + \begin{bmatrix} 0 & -\mathrm{I} & 0 \\ -A_0 & -A_1 & b_0 \\ d_0^T & d_1^T & 0 \end{bmatrix} \begin{bmatrix} w \\ z \\ q \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{4.1}$$

where the coefficient matrices are the same as above. System (4.1) is completed by the three boundary conditions (3.24), (3.30), (3.31).

The above system (4.1) is differential-algebraic; in the next lemma, we show it has index one. (The meaning of this term is defined in the proof, or see [4].) The approximation of solutions of the initial-value problem for such low-index DAEs is relatively well understood; see, for instance, [4] for convergence results. Moreover, some results for the initial-value problem may be extended to boundary-value problems; see [5]. These considerations provide a theoretical justification for our using the midpoint rule to solve (4.1) numerically.

LEMMA 4.1. *Assuming downward flow, i.e., $v_r(\theta) < 0$ for any θ, the first-order system is differential-algebraic of index one.*

Proof. We need to show that by differentiating some of the components of (4.1) at most once, the algebraic character of the system can be eliminated, leaving a purely differential equation. Let us differentiate only the last component of (4.1),

$$d_0^T w + d_1^T z = 0. \tag{4.2}$$

The resulting system may be written as

$$\begin{bmatrix} I & 0 & 0 \\ 0 & -A_2 & b_1 \\ 0 & d_1^T & 0 \end{bmatrix} \begin{bmatrix} w' \\ z' \\ q' \end{bmatrix} + \begin{bmatrix} \text{linear} \\ \text{zeroth-order} \\ \text{terms} \end{bmatrix} = 0. \tag{4.3}$$

We claim the coefficient matrix in (4.3) is nonsingular. Then, multiplying (4.3) by the inverse of this matrix, we obtain a purely differential equation.

To prove the claim, it suffices to show that

$$B = \begin{bmatrix} +\hat{\lambda}^{(0)} A_2 & b_1 \\ d_1^T & 0 \end{bmatrix} \tag{4.4}$$

is nonsingular, where, without changing invertibility, we have inserted a factor of $-\hat{\lambda}^{(0)}$ in the upper left, which simplifies the calculation. Let us introduce the notation W for the column vector on the right-hand side of (3.19), so that $\hat{V}^{(0)}/(\hat{p}^{(0)}\hat{\lambda}^{(0)}) = s\,W$. Then from the definitions following (3.17), (3.18), we have $b_1 = g_1 + sD_1W$. Similarly, regarding A_2, since $\mathcal{M}G_1 = D_1^T$, we have $\hat{\lambda}^{(0)}A_2 = D_1G_1 - \frac{1}{2}(D_1W)(D_1W)^T$. But

$$D_1W = \begin{bmatrix} -\sin 2\psi & \frac{1}{\sqrt{3}}\cos 2\psi & 0 \end{bmatrix}^T.$$

Hence matrix (4.4) equals

$$B = \begin{bmatrix} \frac{1}{2} - \frac{1}{2}\sin^2 2\psi & * & * & -s\sin 2\psi \\ \frac{1}{2\sqrt{3}}\cos 2\psi \sin 2\psi & * & * & 1 + \frac{s}{\sqrt{3}}\cos 2\psi \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

where $*$ indicates elements that do not affect the invertibility of B. It is readily calculated that

$$\det B = -\frac{1}{4}\left(\cos^2 2\psi + \frac{s}{\sqrt{3}}\cos 2\psi\right).$$

As shown on p. 43 of [12], the assumption that $v_r^{(0)} < 0$ implies that $|\psi(\theta)| < \pi/4$, and the claim follows. ∎

COROLLARY 4.2. *The solution space of* (4.1) *has dimension six.*

Proof. The solution space of (4.3), which is seven-dimensional, may be parameterized by initial values $\begin{bmatrix} w(\theta_0) & z(\theta_0) & q(\theta_0) \end{bmatrix}^T$. Since (4.3) was obtained from (4.1) by differentiating (4.2), we conclude that for a solution of (4.3),

$$d_0^T w(\theta) + d_1^T z(\theta) \equiv 0 \quad \text{iff} \quad d_0^T w(\theta_0) + d_1^T z(\theta_0) = 0.$$

Thus the solution space of (4.1) may be identified with the set of solutions of (4.3) whose initial conditions satisfy the scalar equation (4.2). ∎

The boundary-value problem (4.1), (3.24), (3.30), (3.31) is discretized using a symmetric implicit Runge–Kutta method [2], [4]. Since the solutions are expected to behave smoothly with respect to θ, the simplest of those methods, namely the midpoint rule, is chosen. In spite of being only second-order accurate, this choice is shown to be adequate below. The interval $(0, \theta_w)$ is divided into N subintervals of size $\Delta\theta = \theta_w/N$, defining a uniform mesh with nodes $\theta_i = i\,\Delta\theta$, $i = 0, 1, \ldots, N$. At each grid point θ_i there are seven unknowns,

$$U^i = [w_1^i \;\; w_2^i \;\; w_3^i \;\; z_1^i \;\; z_2^i \;\; z_3^i \;\; q^i]^T.$$

Since there are $N+1$ grid points, there are $7(N+1)$ unknowns in total. The midpoint

TABLE 4.1
Numerical boundary conditions at $\theta = 0$.

$m = 1$	$w_1^0 = 0$	$w_2^0 + w_3^0 = 0$	$z_3^0 = 0$	$q^0 = 0$
$m \geq 2$	$w_1^0 = 0$	$w_2^0 = 0$	$w_3^0 = 0$	$q^0 = 0$

rule for the ODE (4.1) is applied on each interval $[\theta_{i-1}, \theta_i]$, $i = 1, \ldots, N$, leading to $7N$ equations for the $7(N+1)$ unknowns.

Seven additional equations are needed to close the system, and these are provided by the boundary conditions. At $\theta = \theta_w$, the three conditions (3.24), (3.30), (3.31) are imposed, and at $\theta = 0$, the four numerical boundary conditions listed in Table 4.1 are imposed. The latter boundary conditions may be justified as follows. According to (3.21), (3.22), as $\theta \to 0$,

$$w \sim \theta^\nu, \qquad q \sim \theta^{\nu-1},$$

where $\nu \geq m - 1$. Thus $w(0) = 0$ if $m \geq 2$, and $q(0) = 0$ if $m \geq 3$. In fact, if $m = 2$, direct calculation of the Frobenius solution (3.21) shows that $q(0) = 0$ remains true in this case, too. If $m = 1$, we refer to Lemma 3.1: by parity, w_1, q, and $z_3 = dw_3/d\theta$ all vanish at $\theta = 0$. The fourth boundary condition in Table 4.1 follows from the last equation in (4.1) in the limit $\theta \to 0$.

The resulting $7(N+1) \times 7(N+1)$ system has the following structure:

$$\begin{bmatrix} S_1 & R_1 & & & & \\ & S_2 & R_2 & & & \\ & & \ddots & \ddots & & \\ & & & S_N & R_N & \\ B_0 & & & & B_w \end{bmatrix} \begin{bmatrix} U^0 \\ U^1 \\ \vdots \\ U^{N-1} \\ U^N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ Q \end{bmatrix}. \tag{4.5}$$

The last row of the above system corresponds to the implementation of the boundary conditions; the 7×7 matrices B_0 and B_w contain the coefficients entering in the formula from Table 4.1 and (3.24), (3.30), (3.31), respectively, while Q corresponds to the nonhomogeneous part of the boundary condition (3.31).

5. Numerical results.

5.1. Secondary circulation. We claim that, for solutions of (4.1), secondary circulation—i.e., flow tangential to the spherical cap $\{r = \text{const}\}$—may be described in terms of the stream function

$$\Psi = \frac{1}{mr} \sin\theta \sin m\phi \, w_2(\theta).$$

In other words, we must show that

$$v_\theta^{(1)} = \frac{1}{r \sin\theta} \partial_\phi \Psi, \tag{5.1a}$$

$$v_\phi^{(1)} = -\frac{1}{r} \partial_\theta \Psi. \tag{5.1b}$$

Since $v_\theta^{(1)} = r^{-2} w_2(\theta) \cos m\phi$, equation (5.1a) follows by direct differentiation. On

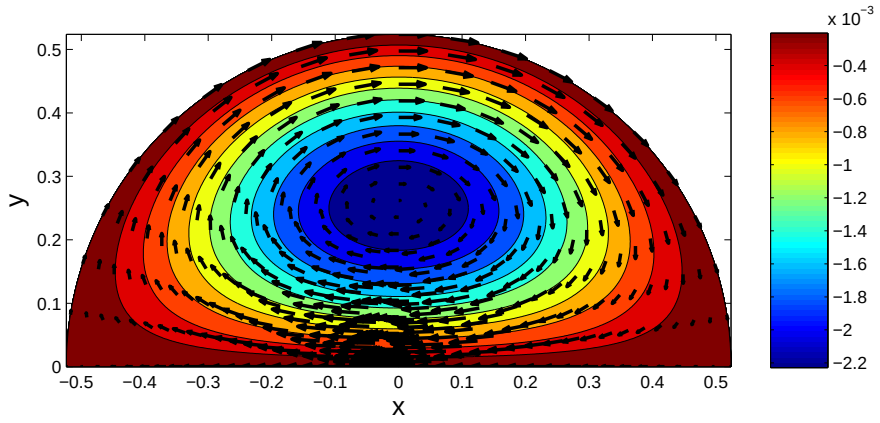


FIG. 5.1. *Stream function showing secondary flow in a tilted hopper ($m = 1$, $\theta_w = 30°$, $\delta = 30°$, angle of wall friction $= 14°$), i.e., $\mu_w = \tan 14°$. By symmetry, only half of the hopper is represented.*

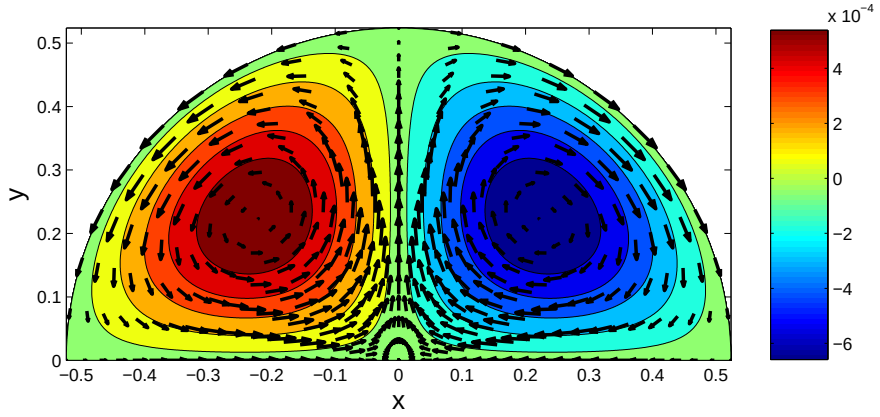


FIG. 5.2. *Stream function showing secondary flow in an "elliptical" hopper ($m = 2$, $\theta_w = 30°$, $\delta = 30°$, angle of wall friction $= 14°$).*

the other hand, since $v_r^{(1)} = r^{-2} w_1(\theta) \cos m\phi$, we have $(\partial_r + 2r^{-1}) v_r = 0$, so by (3.7)

$$(\partial_\theta + \cot\theta) w_2(\theta) \cos m\phi + \csc\theta \, w_3(\theta) \, \partial_\phi(\sin m\phi) = 0,$$

from which (5.1b) follows. Figures 5.1 and 5.2 show plots of the level lines of Ψ, which equal the projection of the streamlines onto a spherical cap $\{r = \text{const}\}$. Figure 5.1 corresponds to a tilted hopper ($m = 1$), while Figure 5.2 corresponds to an "elliptical" hopper ($m = 2$). The grains do not move along radial lines but follow more complicated and fully three-dimensional trajectories.

The sign of the main circulation changes when μ_w increases. The corresponding transition is independent of the value of m, but is a property of the radial solution itself. Specifically, the circulation vanishes when the boundary condition for the correction terms (3.31) is homogeneous, i.e., $\partial_\theta \hat{T}_{r\theta}^{(0)} + \mu_w \partial_\theta \hat{T}_{\theta\theta}^{(0)} = 0$ at $\theta = \theta_w$. The range of θ_w in Figure 5.3 is limited by the mass-flow limit—exceeding this limit leads to flows with rigid regions, to which the present model does not apply. The range of μ_w

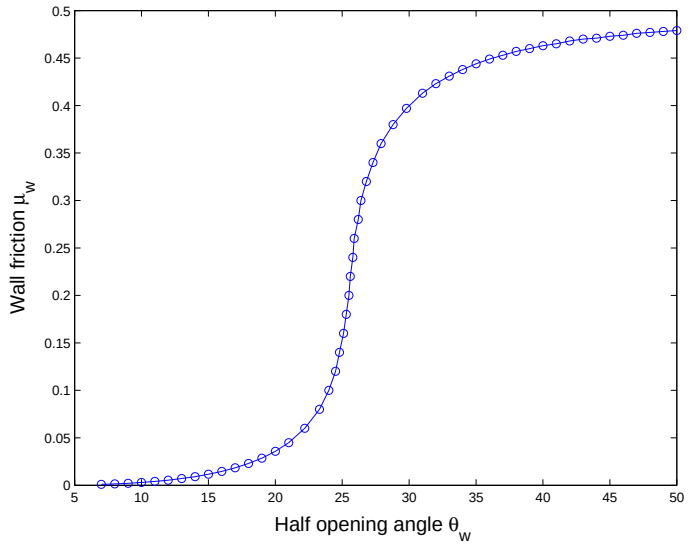


FIG. 5.3. *Critical values leading to sign changes of the circulation (internal friction $\delta = 30°$).*

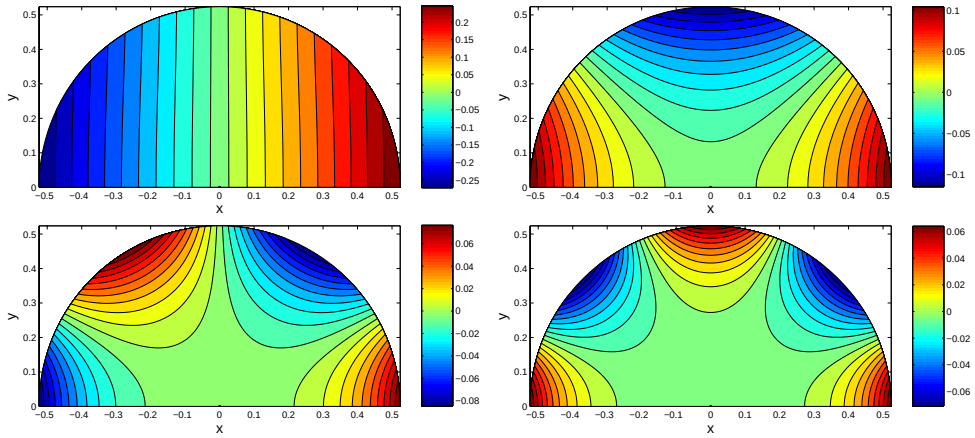


FIG. 5.4. *Influence of the geometry on the mean stress corrections ($\theta_w = 30°$, $\delta = 30°$, angle of wall friction $= 14°$); upper left: $m = 1$, upper right: $m = 2$, lower left: $m = 3$, lower right: $m = 4$.*

is limited by the condition that $\mu_w < \sin\delta = 1/2$; here the upper bound corresponds to a fully rough wall [7].

The effects of the geometry on the mean stress corrections are illustrated in Figure 5.4.

5.2. Checks on the computation. For comparison with the above numerical solution, the method of Frobenius was applied directly to the system (3.17), (3.18) using Maple. Given Jenike's radial field, a linear system for the coefficients of the series solution is readily formed and solved, yielding a solution with three free parameters, corresponding to the three linearly independent solutions in Proposition 3.2. Subsequently, the three boundary conditions (3.24), (3.30), (3.31) provide the needed relations to determine the solution to the full boundary-value problem.

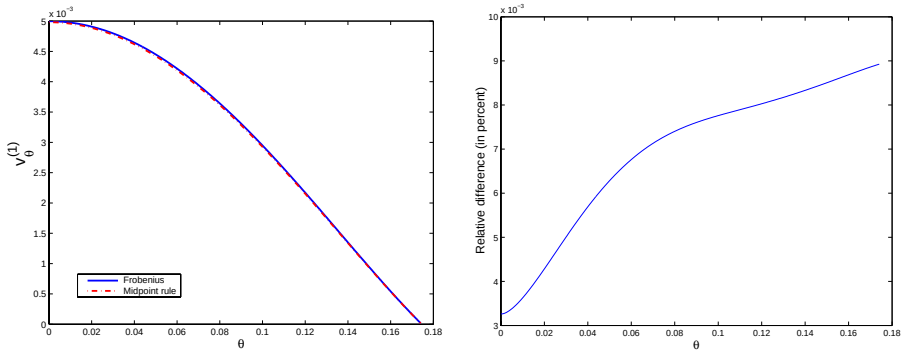


FIG. 5.5. *Comparison of* $v_\theta^{(1)}$ *from the purely numerical method of section* 4 *and from the Frobenius method of section* 5.2. *(Using* $m = 1$, $\theta_w = 10°$, $\delta = 30°$, *and* $\mu_w = 0.3$.*)*

Two methods of obtaining the radial field were employed. Under the assumption that θ_w^2 and μ_w/θ_w are both small and of the same order, a series representation of the Jenike field was computed within Maple itself. Under the less restrictive assumption that only θ_w be small (say 10°), numerical solutions were computed in MATLAB, fitted to polynomials, and then imported into Maple. In both cases, the resulting polynomials were then used to compute the first-order correction. The corrections to the stress and velocity obtained through this symbolic approach agree extremely well with the results of the purely numerical method of sections 4 and 5: for the representative values $m = 1$, $\theta_w = 10°$, $\delta = 30°$, and $\mu_w = 0.3$, the corrections obtained by the two different methods have a relative difference of less than 1%; see Figure 5.5.

Acknowledgments. The authors thank Bob Behringer, Steve Campbell, Tim Kelley, Tony Royal, and Michael Shearer for many helpful discussions.

REFERENCES

[1] S. AGMON, A. DOUGLIS, AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* II, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.

[2] U.M. ASCHER AND L.R. PETZOLD, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, SIAM, Philadelphia, 1998.

[3] C.M. BENDER AND S.A. ORSZAG, *Advanced Mathematical Methods for Scientists and Engineers*, International Series in Pure and Applied Mathematics, McGraw–Hill, New York, 1978.

[4] K.E. BRENAN, S.L. CAMPBELL, AND L.R. PETZOLD, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Classics Appl. Math. 14, SIAM, Philadelphia, 1996.

[5] K.D. CLARK AND L.R. PETZOLD, *Numerical solution of boundary value problems in differential-algebraic systems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 915–936.

[6] P.A. GREMAUD, J.V. MATTHEWS, AND M. SHEARER, *Similarity solutions for granular materials in hoppers*, in Nonlinear PDE's, Dynamics, and Continuum Physics, Contemp. Math., 255, J. Bona, K. Saxton and R. Saxton, eds., AMS, Providence, RI, 2000, pp. 79–95.

[7] A.W. JENIKE, *Gravity flow of bulk solids*, Bulletin 108, Utah Eng. Expt. Station, University of Utah, Salt Lake City, 1961.

[8] T.M. KNOWLTON, J.W. CARSON, G.E. KLINZING, AND W.C. YANG, *The importance of storage, transfer and collection*, Chem. Eng. Prog., 90 (1994), pp. 44–54.

[9] E.W. MERROW, *A quantitative assessment of R&D requirements for solids processing technology*, Publication R-3216-DOE/PSSP, Rand Corporation, Santa Monica, CA, 1986.

[10] R.M. Nedderman, *Static and Kinematic of Granular Materials*, Cambridge University Press, Cambridge, UK, 1992.

[11] N. Rege, *Computational Modeling of Granular Materials*, Ph.D. thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, 1996.

[12] D.G. Schaeffer, *Instability in the evolution equations describing incompressible granular flow*, J. Differential Equations, 66 (1987), pp. 19–50.

[13] J.R. Williams and N. Rege, *The development of circulation cell structures in granular materials undergoing compression*, Powder Technol., 90 (1997), pp. 187–194.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 601–636

RESONANCES FOR MICROSTRIP TRANSMISSION LINES*

HABIB AMMARI† AND FAOUZI TRIKI†

Abstract. In this paper we rigorously derive asymptotic formulae for resonances associated with a microstrip transmission line mounted on a planar waveguide with variable electromagnetic characteristics when the width of the line goes to zero.

Key words. microstrip transmission lines, waveguides, resonances, characteristic numbers, generalized Rouché theorem, integral operators, asymptotic expansions

AMS subject classifications. 78A50, 35J05, 35C15, 35P05

DOI. 10.1137/S0036139902418390

1. Introduction.

In this paper we discuss resonance problems inherent in microstrip transmission lines mounted on planar open waveguides. Microstrip transmission lines are widely used in printed-circuit technology, microwave integrated circuits, and the antenna industry [7], [15], [6], [11]. Since microstrip transmission lines are highly resonant structures, an accurate determination of their resonances is of great importance [4], [7], [15], [6], [11]. Our aim in this work is to rigorously derive asymptotic formulae for resonances associated with the full Maxwell's equations in a microstrip transmission line mounted on a planar open waveguide when the width of the microstrip transmission line goes to zero. The waveguide is half space ($y > 0$) with the Dirichlet boundary condition on $y = 0$. The region $0 < y < h$ is considered the core of the fiber, while the remainder is considered the cladding. The electric permittivity of the waveguide is y-dependent in the core and is constant in the cladding. The magnetic permeability is constant in each part. The electromagnetic characteristics of the waveguide are then given by

$$\varepsilon(y) = \begin{cases} \varepsilon_1(y) & \text{in }]0, h[, \\ \varepsilon_2 & \text{in }]h, +\infty[\end{cases}$$

and

$$\mu(y) = \begin{cases} \mu_1 & \text{in }]0, h[, \\ \mu_2 & \text{in }]h, +\infty[, \end{cases}$$

where $\varepsilon_1(y)\mu_1 \geq \varepsilon_2\mu_2$ and $\mu_1 \neq \mu_2$.

This resonant problem is a spectral problem nonlinear in the spectral variable, that is, the frequency. By integral equations, we reduce this problem to the existence and the distribution of the characteristic values of two families of self-adjoint integral operators in the complex plane. Powerful techniques from the theory of meromorphic operator-valued functions and careful asymptotic analysis of integral kernels are combined for solving this problem. Our results are expected to lead to sophisticated

*Received by the editors December 27, 2002; accepted for publication (in revised form) July 18, 2003; published electronically December 31, 2003. This research was partially supported by ACI Jeunes Chercheurs (0693) from the Ministry of Education and Scientific Research, France.
http://www.siam.org/journals/siap/64-2/41839.html
†Centre de Mathématiques Appliquées, CNRS UMR 7641 and Ecole Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr, triki@cmapx.polytechnique.fr).

numerical tools for working engineers and scientists in the design and analysis of microstrip structures and circuits to be installed on waveguides. It is also hoped that this work will provide some insight into correct analysis of both the resonant and the radiation problems for microstrip transmission lines.

The paper is organized as follows. In section 2, we model the search of the resonances of the microstrip transmission line as two nonlinear spectral problems (2.6) and (2.7). Then, in section 3, we reformulate (2.6) and (2.7) as two systems of integral equations depending on the small parameter 2α which is the width of the transmission line. We transform these systems into the determination of the characteristic values of two integral operator-valued functions $\mathcal{A}_\alpha$ and $\mathcal{B}_\alpha$ in the complex plane. The main ingredient for doing this is an inverse transform formula stated in Lemma 3.1. A similar formula was first derived by Magnanini and Santosa [8]. A generalization of the Rouché theorem to operator-valued functions shows the existence of resonances close to the resonant frequencies of the waveguide, considered as a reference structure. The idea of reducing the resonant problem to the study of characteristic values of some integral operators has been introduced by Russian authors; see [12] and the references listed there. Section 4 is devoted to the rigorous derivation of asymptotic expansions of the resonances as α goes to zero. In Appendix A, we give a proof of Lemma 3.1. Appendix B contains statements of the main results from the work of Gohberg and Sigal [2] that are used here.

Our asymptotic formulae can be interpreted as follows. As will be shown in this paper, the introduction of the microstrip transmission line perturbs the guided frequencies and transforms the guided modes into radiative modes with a complex z-wave number γ ($\Im(\gamma) > 0$). In other words, the complex part of the perturbations in γ that are due to the transmission line corresponds to losses in the guided modes by the planar waveguide. The main problem in applications then is to reduce these losses.

It should also be remarked that the leading order resonances perturbations resulting from the presence of the microstrip transmission line are of order $O(\frac{1}{\ln \alpha})$. This finding is closely connected with the work of Ozawa [10] (see also [5]). Ozawa derives an asymptotic expression of the eigenvalues of the Laplacian in a bounded domain in $\mathbb{R}^2$ in the presence of an interior small perfectly conducting ball (with the Dirichlet boundary condition on its boundary). Ozawa's asymptotic expression bears some resemblance to ours. We finally note that our approach can be used to correctly solve the radiation problem for microstrip transmission lines. The resonant problem for microstrip gratings that are planar periodic structures can also be solved by a similar method.

2. Position of the problem. Let (e_x, e_y, e_z) be an orthonormal basis of $\mathbb{R}^3$. The microstrip structure extends in the domain $\mathbb{R}^3_+ = \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}$ and is invariant under any translation in the z-direction. It consists of a perfectly conducting ground plane $G \times \mathbb{R} = \mathbb{R} \times \{y = 0\} \times \mathbb{R}$ coated with a dielectric layer of constant thickness h. A single infinite strip line $P \times \mathbb{R} =]-\alpha, \alpha[\times \{y = h\} \times \mathbb{R}$ of width 2α is posed on the upper plane $\{y = h\}$. The exterior medium, characterized by a positive constant electric permittivity ε_2 and a positive constant magnetic permeability μ_2, fills the upper domain $\mathcal{O}_2 \times \mathbb{R} = \mathbb{R} \times]h, +\infty[\times \mathbb{R}$. The dielectric layer fills the lower domain $\mathcal{O}_1 \times \mathbb{R} = \mathbb{R} \times]0, h[\times \mathbb{R}$. We assume that the electric permittivity ε_1 of the dielectric layer is a variable function in the y-direction and its magnetic permittivity is a positive constant μ_1. Throughout this paper, we adopt the following notation: $f^\pm$ is the limit of $f(y)$ when $y \to h^\pm$; $[f] = f^+ - f^-$ on P denotes the jump between the boundary

values of f on the two sides of P.

The microstrip transmission line is illuminated by a harmonic incident plane wave, $\mathcal{E}^{inc}(x,y,z,t)$ and $\mathcal{H}^{inc}(x,y,z,t)$. We denote by ω its frequency, by $\vec{k} = k_x e_x + k_y e_y + k_z e_z$ its wave vector, and by $\vec{X} = (x,y,z)$ the space vector. The frequency $\omega = |\vec{k}|$ is a fixed positive real. Thus we have

$$\mathcal{E}^{inc}(\vec{X},t) = E_0^{inc}(\vec{X}) e^{-i\omega t + \vec{k}.\vec{X}},$$
$$\mathcal{H}^{inc}(\vec{X},t) = H_0^{inc}(\vec{X}) e^{-i\omega t + \vec{k}.\vec{X}}.$$

Let $(\mathcal{E},\mathcal{H})$ be the total electric and magnetic fields in the waveguide structure. Hence

$$\mathcal{E}(\vec{X},t) = \mathcal{E}^{inc}(\vec{X},t) + \mathcal{E}^{sca}(\vec{X},t),$$
$$\mathcal{H}(\vec{X},t) = \mathcal{H}^{inc}(\vec{X},t) + \mathcal{H}^{sca}(\vec{X},t),$$

where $\mathcal{E}^{sca}$ and $\mathcal{H}^{sca}$ are the scattered electric and magnetic fields, respectively. The vector fields $\mathcal{E}$ and $\mathcal{H}$ are the solutions of the following linear Maxwell system:

$$\begin{cases} \operatorname{curl} \mathcal{E}(\vec{X},t) = \mu(y)\, \partial_t \mathcal{H}(\vec{X},t), \\ \operatorname{curl} \mathcal{H}(\vec{X},t) = -\varepsilon(y)\, \partial_t \mathcal{E}(\vec{X},t). \end{cases}$$

Guided modes of the waveguide structure are particular solutions such that

$$\begin{cases} \mathcal{E}(x,y,z,t) = \Re\left(\mathbb{E}(x,y,z)\exp(-i\omega t)\right), \\ \mathcal{H}(x,y,z,t) = \Re\left(\mathbb{H}(x,y,z)\exp(-i\omega t)\right), \end{cases}$$

which gives that the electric field $\mathbb{E}(x,y,z)$ satisfies

$$\begin{cases} \operatorname{curl}\left(\dfrac{1}{\mu(y)}\operatorname{curl}\mathbb{E}\right) + \omega^2 \varepsilon(y)\mathbb{E} = 0 \quad \text{in } \mathbb{R}^3_+ \setminus P \times \mathbb{R}, \\ \operatorname{div}(\varepsilon(y)\mathbb{E}) = 0 \quad \text{in } \mathbb{R}^3_+ \setminus P \times \mathbb{R}, \\ \mathbb{E} \wedge e_y = 0 \quad \text{on } (G \cup P) \times \mathbb{R}. \end{cases} \tag{2.1}$$

The z-invariance of the microstrip line structure yields to a z-dependency of the form $e^{i\gamma z}$ for the field $\mathbb{E}(x,y,z)$, where γ is real, called the z-wave number:

$$\mathbb{E}(x,y,z) = e^{i\gamma z} E(x,y). \tag{2.2}$$

Hence, if $\varepsilon_1(y)$ is assumed to be constant, then the problem (2.1) may be reduced to a system of three two-dimensional Helmholtz scalar equations. We refer the reader to [1] for a proof.

LEMMA 2.1. *The components E_x, E_y, and E_z of the electric field $E(x,y)$ defined by* (2.2)*, where $\mathbb{E}$ is a solution to* (2.1)*, solve the following transmission boundary-value problems:*

$$\begin{cases} \Delta E_y + (\omega^2 \varepsilon(y)\mu(y) - \gamma^2) E_y = 0 & \text{in} \quad \mathcal{O}_1 \cup \mathcal{O}_2, \\ [\partial_y E_y] = 0 & \text{on} \quad L, \\ [\varepsilon E_y] = 0 & \text{on} \quad L \setminus P, \\ \partial_y E_y = 0 & \text{on} \quad G \cup P, \end{cases} \tag{2.3}$$

$$\begin{cases} \Delta E_x + (\omega^2\varepsilon(y)\mu(y) - \gamma^2)E_x = 0 & \text{in} \quad \mathcal{O}_1 \cup \mathcal{O}_2, \\ [E_x] = 0 & \text{on} \quad L, \\ \left[\dfrac{1}{\mu}\partial_y E_x\right] = \dfrac{1}{\mu_2}\left(1 - \dfrac{\varepsilon_2\mu_2}{\varepsilon_1(h)\mu_1}\right)\partial_x E_y^+ & \text{on} \quad L \setminus P, \\ E_x = 0 & \text{on} \quad G \cup P, \end{cases} \tag{2.4}$$

and

$$\begin{cases} \Delta E_z + (\omega^2\varepsilon(y)\mu(y) - \gamma^2)E_z = 0 & \text{in} \quad \mathcal{O}_1 \cup \mathcal{O}_2, \\ [E_z] = 0 & \text{on} \quad L, \\ \left[\dfrac{1}{\mu}\partial_y E_z\right] = \dfrac{i\gamma}{\mu_2}\left(1 - \dfrac{\varepsilon_2\mu_2}{\varepsilon_1(h)\mu_1}\right)E_y^+ & \text{on} \quad L \setminus P, \\ E_z = 0 & \text{on} \quad G \cup P. \end{cases} \tag{2.5}$$

Our aim in this work is to investigate the problem of finding $\gamma^2(\omega)$ such that at least one of the scalar Helmholtz equations (2.3)–(2.5) has a nontrivial solution subject to a radiation condition at infinity. This behavior at infinity will be made clear after (3.6).

We will consider a slightly more general problem. Throughout this paper, we assume that $\varepsilon_1(y)$ is a variable function in $]0, h[$. This would allow us to solve as well the general resonant problem in the acoustic case. We denote by Σ the set of these numbers. We remark that since (2.3) is independent of the two other Helmholtz equations, the set Σ is then the union of two families:

Family (a): the electric field $E(x, y) = (E_x, 0, E_z)$. This solution is called a longitudinal section electric mode. One of the components E_x or E_z is nontrivial and satisfies

$$\begin{cases} \Delta u + (\omega^2\varepsilon\mu - \gamma^2)u = 0 & \text{in} \quad \mathcal{O}_1 \cup \mathcal{O}_2, \\ [u] = 0 & \text{on} \quad L, \\ \left[\dfrac{1}{\mu}\partial_y u\right] = 0 & \text{on} \quad L \setminus P, \\ u = 0 & \text{on} \quad G \cup P. \end{cases} \tag{2.6}$$

Family (b): the electric field $E(x, y) = (E_x, E_y, E_z)$. This solution is called a longitudinal section magnetic mode. The component E_y is nontrivial and satisfies

$$\begin{cases} \Delta v + (\omega^2\varepsilon\mu - \gamma^2)v = 0 & \text{in} \quad \mathcal{O}_1 \cup \mathcal{O}_2, \\ [\partial_y v] = 0 & \text{on} \quad L, \\ [\varepsilon v] = 0 & \text{on} \quad L \setminus P, \\ \partial_y v = 0 & \text{on} \quad G \cup P. \end{cases} \tag{2.7}$$

Let $\Sigma_{\mathbf{a}}$ and $\Sigma_{\mathbf{b}}$ be the set of values of γ for fixed ω (or equivalently ω for fixed γ) such that (2.6), respectively (2.7), has a nontrivial solution. Then, we have $\Sigma = \Sigma_{\mathbf{a}} \cup \Sigma_{\mathbf{b}}$. The eigenfunctions associated with γ in Σ are called degenerate modes. If $\Sigma_{\mathbf{a}} \cap \Sigma_{\mathbf{b}} = \varnothing$, then E_x and E_z are uniquely determined for any $\gamma \in \Sigma_b$. Note that because the treatments of families (a) and (b) are quite different, we consider both of them in

detail. We recall the following definitions. We refer the reader to Appendix B for more details.

DEFINITION 2.2. *A complex number γ is called a resonance if and only if the system of Helmholtz equations* (2.3)–(2.5) *has a nontrivial solution.*

DEFINITION 2.3. *Let X and Y be two Banach spaces and $\mathcal{L}(X,Y)$ be the space of bounded linear operators acting from X to Y. We denote by 0_X (respectively, 0_Y) the null element of X (respectively, Y). Let $A:\mathbb{C}\to\mathcal{L}(X,Y)$ be an operator-valued function. The complex number Γ^* is a characteristic value of A if and only if the function $A(\Gamma)$ is holomorphic in a punctured neighborhood of Γ^*; there exists a function $x:\mathbb{C}\to X$ such that $x(\Gamma^*)\neq 0_X, \Gamma\mapsto x(\Gamma)$, and $\Gamma\mapsto A(\Gamma)x(\Gamma)$ are holomorphic in $\Gamma=\Gamma^*$, and $A(\Gamma^*)x(\Gamma^*)=0_Y$. The order of Γ^*, as a zero of $\Gamma\mapsto A(\Gamma)x(\Gamma)$, is called its multiplicity. The function $\Gamma\mapsto x(\Gamma)$ is a root function associated with Γ^*.*

DEFINITION 2.4. *Let D be an open connected domain in $\mathbb{C}$. The full multiplicity of the operator-valued function $\Gamma\mapsto A(\Gamma)$ in D, denoted by $\mathcal{M}(A,\partial D)$, is defined by*

$$\mathcal{M}(A,\partial D)=N(A,\partial D)-P(A,\partial D),$$

where $P(A,\partial D)$ (respectively, $N(A,\partial D)$) is the number of poles (respectively, characteristic values) of A in D, counted according to their multiplicity.

3. Integral representation formulae. We assume that $y\to\varepsilon_1(y)$ is a decreasing, positive, and piecewise $\mathcal{C}^1$ function. We introduce the following notation:

$$\begin{aligned}
&q(y)=\omega^2(\varepsilon_1(0)\mu_1-\varepsilon(y)\mu(y)),\\
&d^2(\omega)=\omega^2(\varepsilon_1(0)\mu_1-\varepsilon_2\mu_2)\geq 0,\\
&\widetilde{\varepsilon}(y)=\begin{cases}\varepsilon_1(h) & \text{in }]0,h[,\\ \varepsilon_2 & \text{in }]h,+\infty[.\end{cases}
\end{aligned}$$

Note that the function $q(y)$ is positive since $y\to\varepsilon_1(y)$ is a decreasing function and $d^2(\omega)$ is its minimum.

3.1. Family (a). In this subsection, we study the completeness of an associated one-dimensional eigenvalue problem. We derive an integral representation formula of solutions $u(x,y)$ to (2.6). Then we prove that the jump between the boundary value of $\frac{1}{\mu(y)}\partial_y u(x,y)$ on the two sides of the transmission line P is a characteristic function of an integral operator-valued function.

Let $g_a(y,\lambda)$ be defined by

$$\begin{cases}
\partial_{yy}g_a(y,\lambda)+(\lambda-q(y))g_a(y,\lambda)=0 \quad \text{in }]0,h[\cup]h,+\infty[,\\
[g_a(.,\lambda)]=0 \quad \text{on } y=h,\\
\left[\dfrac{1}{\mu}\partial_y g_a(.,\lambda)\right]=0 \quad \text{on } y=h,\\
g_a(0,\lambda)=0 \quad \text{and } \partial_y g_a(0,\lambda)=\sqrt{\lambda}.
\end{cases}\tag{3.1}$$

Setting $\phi_a(y,\lambda)$ the solution of the ODE,

$$\begin{cases}
\partial_{yy}\phi_a(y,\lambda)+(\lambda-q(y))\phi_a(y,\lambda)=0 \quad \text{in }]0,h[,\\
\phi_a(0,\lambda)=0 \quad \text{and } \partial_y\phi_a(0,\lambda)=\sqrt{\lambda},
\end{cases}\tag{3.2}$$

we then write

$$g_a(y,\lambda)=\begin{cases}\phi_a(y,\lambda) & \text{if } y\in]0,h[,\\ \phi_a(h,\lambda)\cos\Big[\sqrt{\lambda-d^2}(y-h)\Big]+\dfrac{\mu_2}{\mu_1}\dfrac{\partial_y\phi_a(h,\lambda)}{\sqrt{\lambda-d^2}}\sin\Big[\sqrt{\lambda-d^2}(y-h)\Big] & \\ \text{if } y\in]h,+\infty[.\end{cases}$$

For $\lambda\geq d^2$, $g_a(y,\lambda)$ is bounded. For $\lambda<d^2$, in view of the above expression of g_a, we impose the dispersion relation

$$\phi_a(h,\lambda)+\frac{\mu_2}{\mu_1}\frac{\partial_y\phi_a(h,\lambda)}{\sqrt{\lambda-d^2}}=0 \tag{3.3}$$

to make $g_a(y,\lambda)$ bounded in $\mathbb{R}^+$. According to [8], there will be a finite number of roots $\lambda_l^a(\omega)$ to (3.3) with associated solutions: $g_a(y,\lambda_l^a)$ for $l=1,2,\dots,m_a$. Moreover, the set of eigenfunctions $g_a(y,\lambda),\lambda\in]0,+\infty[$ is complete in $L^2(\mathbb{R}_+)$. When the magnetic permeabilities μ_1 and μ_2 are equal ($\mu_1=\mu_2$), Magnanini and Santosa [8] proved the completeness of the associated eigenvalue problem and rigorously derived an inverse transform formula. See also the work of Wilcox [13], [14], where the spectrum of the Pekeris operator is investigated. Here the following more general inverse transform formula will prove essential. We refer the reader to Appendix A for a proof.

LEMMA 3.1. *Let $f\in L^2(\mathbb{R}_+,\frac{dy}{\mu(y)})$. We have the inverse transform formula:*

$$\begin{aligned}f(x)=&\sum_{l=1}^{m_a}\frac{2\mu_1\sqrt{d^2-\lambda_l^a}\int_0^{+\infty}g_a(y,\lambda_l^a)f(y)\frac{dy}{\mu(y)}}{\frac{\mu_1}{\mu_2}\phi_a(h,\lambda_l^a)^2+2\sqrt{d^2-\lambda_l^a}\int_0^h\phi_a(y,\lambda_l^a)^2dy}g_a(x,\lambda_l^a)\\ &+\frac{1}{\pi}\int_{d^2}^{+\infty}\frac{\mu_2\sqrt{\lambda-d^2}\int_0^{+\infty}g_a(y,\lambda)f(y)\frac{dy}{\mu(y)}}{(\lambda-d^2)\phi_a(h,\lambda)^2+(\frac{\mu_2}{\mu_1})^2\partial_y\phi_a(h,\lambda)^2}g_a(x,\lambda)d\lambda.\end{aligned} \tag{3.4}$$

We now return to the Helmholtz equation (2.6). Let

$$U(x,\lambda)=\int_0^{+\infty}u(x,y)g_a(y,\lambda)\,\frac{dy}{\mu(y)}.$$

Multiplying (2.6) by $\frac{1}{\mu(y)}g_a(y,\lambda)$ and integrating with respect to the variable y over the interval $]0,+\infty[$, we obtain after some straightforward manipulations for $x\in\mathbb{R}$

$$\partial_{xx}U(x,\lambda)+(\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2)U(x,\lambda)=\phi_a(h,\lambda)\left[\frac{1}{\mu}\partial_y u\right](x,h)\chi(]-\alpha,\alpha[), \tag{3.5}$$

where $\chi(]-\alpha,\alpha[)$ denotes the characteristic function of the interval $]-\alpha,\alpha[$. The solution of (3.5), which is outgoing for $0\leq\lambda+\gamma^2<\omega^2\varepsilon_1(0)\mu_1$ and decays exponentially for $\lambda+\gamma^2>\omega^2\varepsilon_1(0)\mu_1$ as $|x|\to+\infty$, is readily given for $x\in\mathbb{R}$ by the following expression:

$$U(x,\lambda)=\phi_a(h,\lambda)\int_{-\alpha}^{\alpha}\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}}{2i\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}\left[\frac{1}{\mu}\partial_y u\right](\zeta,h)d\zeta. \tag{3.6}$$

By the inversion formula in Lemma 3.1, we have

$$\begin{aligned}u(x,y)=&\sum_{l=1}^{m_a}\frac{2\mu_1\sqrt{d^2-\lambda_l^a}U(x,\lambda_l^a)}{\frac{\mu_1}{\mu_2}\phi_a(h,\lambda_l^a)^2+2\sqrt{d^2-\lambda_l^a}\int_0^h\phi_a(y,\lambda_l^a)^2dy}g_a(y,\lambda_l^a)\\ &+\frac{1}{\pi}\int_{d^2}^{+\infty}\frac{\mu_2\sqrt{\lambda-d^2}U(x,\lambda)}{(\lambda-d^2)\phi_a(h,\lambda)^2+(\frac{\mu_2}{\mu_1})^2\partial_y\phi_a(h,\lambda)^2}g_a(y,\lambda)d\lambda\quad\forall\,(x,y)\in\mathbb{R}^2_+;\end{aligned}$$

hence, by (3.6) and by interchanging the order of integration, we obtain that the jump between the boundary values of $\frac{1}{\mu(y)}\partial_y u(x,y)$ on the two sides of the transmission line P solves the integral equation:

$$\begin{aligned}
&\int_{-\alpha}^{\alpha}\left[\sum_{l=1}^{m_a}\frac{\mu_1\sqrt{d^2-\lambda_l^a}\phi_a(h,\lambda_l^a)^2}{\frac{\mu_1}{\mu_2}\phi_a(h,\lambda_l^a)^2+2\sqrt{d^2-\lambda_l^a}\int_0^h\phi_a(y,\lambda_l^a)^2dy}\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda_l^a-\gamma^2}}}{i\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda_l^a-\gamma^2}}\right.\\
&\left.+\frac{1}{\pi}\int_{d^2}^{+\infty}\frac{\mu_2\sqrt{\lambda-d^2}\phi_a(h,\lambda)^2}{(\lambda-d^2)\phi_a(h,\lambda)^2+(\frac{\mu_2}{\mu_1})^2\partial_y\phi_a(h,\lambda)^2}\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}}{2i\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}d\lambda\right]\\
&\times\left[\frac{1}{\mu}\partial_y u\right](\zeta,h)d\zeta\\
&=0
\end{aligned}$$

for all $x\in]-\alpha,\alpha[$. By the change of variables $X=\frac{x}{\alpha}$ and $X'=\frac{\zeta}{\alpha}$ we immediately obtain

$$\begin{aligned}
&\alpha\int_{-1}^{1}\left[\sum_{l=1}^{m_a}\frac{\mu_1\sqrt{d^2-\lambda_l^a}\phi_a(h,\lambda_l^a)^2}{\frac{\mu_1}{\mu_2}\phi_a(h,\lambda_l^a)^2+2\sqrt{d^2-\lambda_l}\int_0^h\phi_a(y,\lambda_l^a)^2dy}\frac{e^{i\alpha|X'-X|\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda_l^a-\gamma^2}}}{i\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda_l^a-\gamma^2}}\right.\\
&\left.+\frac{1}{\pi}\int_{d^2}^{+\infty}\frac{\mu_2\sqrt{\lambda-d^2}\phi_a(h,\lambda)^2}{(\lambda-d^2)\phi_a(h,\lambda)^2+(\frac{\mu_2}{\mu_1})^2\partial_y\phi_a(h,\lambda)^2}\frac{e^{i\alpha|X'-X|\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}}{2i\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}d\lambda\right]\\
&\times\left[\frac{1}{\mu}\partial_y u\right](\alpha X',h)dX'\\
&=0
\end{aligned}\tag{3.7}$$

for all $X\in]-1,1[$.

Let ω be a nonnegative real, $0\le l_0\le m$, and $\Gamma=\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda_{l_0}^a-\gamma^2}$ a complex value, where $\lambda_{l_0}^a(\omega)$ is a fixed root of the dispersion relation (3.3). We introduce the kernel

$$\begin{aligned}
a_\alpha(\Gamma;X,X')&=\frac{\frac{1}{\ln\alpha}\mu_1\sqrt{d^2-\lambda_{l_0}^a}\phi_a(h,\lambda_{l_0}^a)^2}{\frac{\mu_1}{\mu_2}\phi_a(h,\lambda_{l_0}^a)^2+2\sqrt{d^2-\lambda_{l_0}^a}\int_0^h\phi(y,\lambda_{l_0}^a)^2dy}\frac{e^{i\alpha|X'-X|\Gamma}}{i\Gamma}\\
&+\sum_{l=1,l\neq l_0}^{m_a}\frac{\frac{1}{\ln\alpha}\mu_1\sqrt{d^2-\lambda_l^a}\phi_a(h,\lambda_l^a)^2}{\frac{\mu_1}{\mu_2}\phi_a(h,\lambda_l^a)^2+2\sqrt{d^2-\lambda_l^a}\int_0^h\phi_a(y,\lambda_l^a)^2dy}\frac{e^{i\alpha|X'-X|\sqrt{\lambda_{l_0}-\lambda_l^a+\Gamma^2}}}{i\sqrt{\lambda_{l_0}^a-\lambda_l^a+\Gamma^2}}\\
&+\frac{1}{\pi\ln\alpha}\int_{d^2}^{+\infty}\frac{\mu_2\sqrt{\lambda-d^2}\phi_a(h,\lambda)^2}{(\lambda-d^2)\phi_a(h,\lambda)^2+(\frac{\mu_2}{\mu_1})^2\partial_y\phi_a(h,\lambda)^2}\frac{e^{i\alpha|X'-X|\sqrt{\lambda_{l_0}^a-\lambda+\Gamma^2}}}{2i\sqrt{\lambda_{l_0}^a-\lambda+\Gamma^2}}d\lambda
\end{aligned}$$

and $\mathcal{A}_\alpha(\Gamma)$ to be the integral operator defined by

$$\begin{aligned}
\mathcal{A}_\alpha(\Gamma):\quad &(H^{1/2})'(-1,1)\longrightarrow\widetilde{H}^{1/2}(-1,1),\\
&\psi(X)\longrightarrow\mathcal{A}_\alpha(\Gamma)\psi(X)=\int_{-1}^{1}a_\alpha(\Gamma;X,X')\psi(X')\,dX',
\end{aligned}$$

where $(H^{1/2})'(-1,1)$ is the L^2-dual of $H^{1/2}(-1,1)$ and $\widetilde{H}^{1/2}(-1,1)$ denotes the space of functions in $H^{1/2}(-1,1)$ such that their extensions by 0 on $]-\infty,-1[\cup]1,+\infty[$ are in $H^{1/2}(\mathbb{R})$. Identity (3.7) may alternatively be written in the form

$$\mathcal{A}_\alpha(\Gamma)\left[\frac{1}{\mu}\partial_y u\right](\alpha X) = 0 \quad \forall\, X \in\,]-1,1[.$$

It is therefore obvious that the function $X' \to [\frac{1}{\mu}\partial_y u](\alpha X', h)$ is then a characteristic function of the integral operator-valued function $\mathcal{A}_\alpha$. We will give a rigorous study of the integral operator-valued function $\Gamma \to \mathcal{A}_\alpha(\Gamma)$, when Γ is in a small complex neighborhood of 0.

3.2. Family (b). In this subsection we derive an integral representation formula for the solution $v(x,y)$ to the Helmholtz equation (2.7). We prove that the jump of $\varepsilon(y)v(x,y)$ across the transmission line P is a characteristic function of an integral operator-valued function. Let $g_b(y,\lambda)$ denote the solution to

$$\begin{cases} \partial_{yy} g_b(y,\lambda) + (\lambda - q(y)) g_b(y,\lambda) = 0 \quad \text{in }]0,h[\cup]h,+\infty[, \\ [\widetilde{\varepsilon}(.) g_b(.,\lambda)] = 0 \quad \text{on } y = h, \\ [\partial_y g_b(.,\lambda)] = 0 \quad \text{on } y = h, \\ g_b(0,\lambda) = 1 \text{ and } \partial_y g_b(0,\lambda) = 0. \end{cases} \tag{3.8}$$

We introduce $\phi_b(y,\lambda)$ as the solution of

$$\begin{cases} \partial_{yy}\phi_b(y,\lambda) + (\lambda - q(y))\phi_b(y,\lambda) = 0 \quad \text{in }]0,h[, \\ \phi_b(0,\lambda) = 1 \text{ and } \partial_y \phi_b(0,\lambda) = 0 \end{cases} \tag{3.9}$$

to get

$$g_b(y,\lambda) = \begin{cases} \phi_b(y,\lambda) \quad \text{if } y \in\,]0,h[, \\ \dfrac{\varepsilon_1(h)}{\varepsilon_2}\phi_b(h,\lambda)\cos\Big[\sqrt{\lambda - d^2}(y-h)\Big] + \dfrac{\partial_y\phi_b(h,\lambda)}{\sqrt{\lambda-d^2}}\sin\Big[\sqrt{\lambda-d^2}(y-h)\Big] \\ \quad \text{if } y \in\,]h,+\infty[. \end{cases}$$

For $\lambda \geq d^2$, $g_b(y,\lambda)$ is bounded. For $\lambda < d^2$, we should impose as before the condition

$$\phi_b(h,\lambda) + \frac{\varepsilon_2}{\varepsilon_1(h)}\frac{\partial_y\phi_b(h,\lambda)}{\sqrt{\lambda-d^2}} = 0 \tag{3.10}$$

to make ϕ_b bounded in $\mathbb{R}^+$. We now know that there is a finite number of roots $\lambda_l^b(\omega)$ to the dispersion relation (3.10) with associated solutions $g_b(y,\lambda_l^b)$ for $l = 1,2,\ldots,m_b$. Furthermore, the associated system of eigenfunctions $g_b(y,\lambda)$ for $\lambda \in]0,+\infty[$ is complete in $L^2(\mathbb{R}_+)$. We will need the following inverse transform.

LEMMA 3.2. *Let $f \in L^2(\mathbb{R}_+, \widetilde{\varepsilon}(y)dy)$. We have the inverse transform formula*

$$\begin{aligned} f(x) = \frac{1}{\varepsilon_2}\sum_{l=1}^{m_b} & \frac{2\sqrt{d^2-\lambda_l^b}\int_0^{+\infty} g_b(y,\lambda_l^b) f(y)\,\widetilde{\varepsilon}(y)dy}{\frac{\varepsilon_2}{\varepsilon_1(h)}\phi_b(h,\lambda_l^a)^2 + 2\sqrt{d^2-\lambda_l^b}\int_0^h \phi_b(y,\lambda_l^b)^2 dy} g_b(x,\lambda_l^b) \\ & + \frac{1}{\pi}\frac{\varepsilon_2}{\varepsilon_1^2(h)}\int_{d^2}^{+\infty} \frac{\sqrt{\lambda-d^2}\int_0^{+\infty} g_b(y,\lambda) f(y)\,\widetilde{\varepsilon}(y)dy}{(\lambda-d^2)\phi_b(h,\lambda)^2 + (\frac{\varepsilon_2}{\varepsilon_1(h)})^2\partial_y\phi_b(h,\lambda)^2} g_b(x,\lambda)d\lambda. \end{aligned} \tag{3.11}$$

Let us introduce

$$V(x,\lambda)=\int_0^{+\infty} v(x,y)g_b(y,\lambda)\,\widetilde{\varepsilon}(y)dy.$$

As for establishing identity (3.5), multiplying (2.7) by $\widetilde{\varepsilon}(y)g_b(y,\lambda)$ and integrating in the variable y over the interval $]0,+\infty[$, we arrive at

$$\partial_{xx}V(x,\lambda)+(\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2)V(x,\lambda)=\partial_y\phi_b(h,\lambda)[\varepsilon v](x,h)\chi(]-\alpha,\alpha[) \tag{3.12}$$

for $x\in\mathbb{R}$. The solution of the ODE (3.12), which is outgoing for $0\le\lambda+\gamma^2<\omega^2\varepsilon_1(0)\mu_1$ and decays exponentially for $\lambda+\gamma^2>\omega^2\varepsilon_1(0)\mu_1$ as $|x|\to+\infty$, is given by

$$V(x,\lambda)=\partial_y\phi_b(h,\lambda)\int_{-\alpha}^{\alpha}\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}}{2i\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}[\varepsilon v](\zeta,h)d\zeta\quad\forall\,x\in\mathbb{R}.$$

Hence, by Lemma 3.2 and by interchanging the order of integration, we obtain from the Neumann boundary condition that is imposed on the transmission line P, $\partial_y v(x,h)=0$, for any $x\in\,]-\alpha,\alpha[$, that the following integral equation on P holds:

$$\int_{-\alpha}^{\alpha}\left[\sum_{l=1}^{m_b}\frac{\frac{1}{\varepsilon_2}\sqrt{d^2-\lambda_l^b}\partial_y\phi_b(h,\lambda_l^b)^2}{\frac{\varepsilon_2}{\varepsilon_1(h)}\phi_b(h,\lambda_l^b)^2+\sqrt{d^2-\lambda_l^b}\int_0^h\phi_b(y,\lambda_l^b)^2dy}\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda_l^b-\gamma^2}}}{i\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda_l^b-\gamma^2}}\right.$$
$$\left.+\frac{1}{\pi}\int_{d^2}^{+\infty}\frac{\frac{\varepsilon_2}{\varepsilon_1^2(h)}\sqrt{\lambda-d^2}\partial_y\phi_b(h,\lambda)^2}{(\lambda-d^2)\phi_b(h,\lambda)^2+(\frac{\varepsilon_2}{\varepsilon_1(h)})^2\partial_y\phi_b(h,\lambda)^2}\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}}{2i\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}d\lambda\right]$$
$$\times[\varepsilon v](\zeta,h)d\zeta=0$$

for all $x\in\,]-\alpha,\alpha[$. By the change of variables $X=\frac{x}{\alpha}$ and $X'=\frac{\zeta}{\alpha}$ the last integral equation becomes

$$\mathcal{B}_\alpha(\Gamma)\left([\varepsilon v](\alpha.,h)\right)(X)=0\quad\forall\,X\in\,]-1,1[,$$

where the integral operator-valued function $\mathcal{B}_\alpha(\Gamma)$ is defined by

$$\mathcal{B}_\alpha(\Gamma):\widetilde{H}^{1/2}(-1,1)\longrightarrow H^{-1/2}(-1,1),$$
$$\psi(X)\longrightarrow\mathcal{B}_\alpha(\Gamma)\psi(X)=\int_{-1}^{1}b_\alpha(\Gamma;X,X')\psi(X')dX'.$$

Here $\Gamma=\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda_{l_0}^b-\gamma^2}$, where ω is a positive real, $0\le l_0\le m$, $\lambda_{l_0}^b(\omega)$ is a fixed root of (3.10), and the the kernel b_α is given by

$$b_\alpha(\Gamma;X,X')=\frac{\frac{\alpha^2}{\varepsilon_2}\sqrt{d^2-\lambda_{l_0}^b}\partial_y\phi_b(h,\lambda_{l_0}^b)^2}{\frac{\varepsilon_2}{\varepsilon_1(h)}\phi_b(h,\lambda_{l_0}^b)^2+\sqrt{d^2-\lambda_{l_0}^b}\int_0^h\phi_b(y,\lambda_{l_0}^b)^2dy}\frac{e^{i\alpha|X'-X|\Gamma}}{i\Gamma}$$
$$+\sum_{l=1,l\neq l_0}^{m_b}\frac{\frac{\alpha^2}{\varepsilon_2}\sqrt{d^2-\lambda_l^b}\partial_y\phi_b(h,\lambda_l^b)^2}{\frac{\varepsilon_2}{\varepsilon_1(h)}\phi_b(h,\lambda_l^b)^2+\sqrt{d^2-\lambda_l^b}\int_0^h\phi_b(y,\lambda_l^b)^2dy}\frac{e^{i\alpha|X'-X|\sqrt{\lambda_{l_0}^b-\lambda_l^b+\Gamma^2}}}{i\sqrt{\lambda_{l_0}^b-\lambda_l^b+\Gamma^2}}$$
$$+\frac{\alpha^2}{\pi}\int_{d^2}^{+\infty}\frac{\frac{\varepsilon_2}{\varepsilon_1^2(h)}\sqrt{\lambda-d^2}\partial_y\phi_b(h,\lambda)^2}{(\lambda-d^2)\phi_b(h,\lambda)^2+(\frac{\varepsilon_2}{\varepsilon_1(h)})^2\partial_y\phi_b(h,\lambda)^2}\frac{e^{i|x-\zeta|\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}}{2i\sqrt{\omega^2\varepsilon_1(0)\mu_1-\lambda-\gamma^2}}d\lambda$$

for all $X,X'\in]-1,1[$.

4. Characteristic problem and asymptotic expansions. In this section, we first prove that the resonances are exactly the characteristic values of the operator-valued functions $\mathcal{A}_\alpha(\Gamma)$ and $\mathcal{B}_\alpha(\Gamma)$. Next, we derive an asymptotic formula for these functions. By the generalized Rouché theorem [2] (see Theorem B.4 in Appendix B for a precise statement of this theorem), we derive the leading order terms in the asymptotic expansions of the resonances and their associated guided modes.

4.1. The operator-valued function $\mathcal{A}_\alpha(\Gamma)$. Let ω be a fixed positive real, $0 \leq l_0 \leq m_a$, and $\Gamma = \sqrt{\omega^2\varepsilon_1(0)\mu_1 - \lambda_{l_0}^a - \gamma^2}$ a complex variable, where $\lambda_{l_0}^a(\omega)$ is a root of the dispersion relation (3.3) that is supposed to be different from $d^2(\omega)$. Then 0 is a resonance of the reference waveguide. In this section, we prove that in a fixed neighborhood of 0 lying in the set

$$\mathcal{V}_0 = \left\{ \Gamma \in \mathbb{C}, |\Gamma| < \min_{\lambda \in \{(\lambda_l^a)_{0\leq l\leq m_a, l\neq l_0}, d^2\}} |\lambda_{l_0}^a - \lambda| \right\},$$

there exists a unique resonance of the microstrip line for small values of α, defined as the unique characteristic value of the operator-valued function $\Gamma \mapsto \mathcal{A}_\alpha(\Gamma)$ in $\mathcal{V}_0$. The following results hold.

THEOREM 4.1. *Let $D_{\delta_0}(0) = \{\Gamma \in \mathbb{C}, |\Gamma| < \delta_0\}$, for $\delta_0 > 0$, be a complex neighborhood around 0 in $\mathcal{V}_0$. δ_0 is chosen such that 0 is the unique pole of $\mathcal{A}_\alpha(\Gamma)$ in $D_{\delta_0}(0)$. Then there exists a constant $\alpha_0 > 0$ such that, for $|\alpha| \leq \alpha_0$, we have the following:*

(a) *The operator-valued function $\mathcal{A}_\alpha(\Gamma)$ is finitely meromorphic and of Fredholm type at every point of the domain $D_{\delta_0}(0)$.*

(b) *The following asymptotic formula holds:*

$$\mathcal{A}_\alpha(\Gamma) = A_0^\alpha + \frac{1}{\ln\alpha}\Big(A_1(\Gamma) + \frac{A_{-1}}{\Gamma}\Big) + O(\alpha), \tag{4.1}$$

where $O(\alpha)$ is uniform in the set $\big\{\phi \in (H^{1/2})'(-1,1), ||\phi||_{(H^{1/2})'(-1,1)} \leq 1\big\}$ and for $\Gamma \in \partial D_{\delta_0}(0) = \{\Gamma \in \mathbb{C}, |\Gamma| = \delta_0\}$. The operators $A_0^\alpha, A_1(\Gamma)$, and A_{-1} are defined by

$$\begin{aligned}
A_0^\alpha &: (H^{1/2})'(-1,1) \longrightarrow \widetilde{H}^{1/2}(-1,1),\\
A_0^\alpha\phi(X) &= -\frac{a_0\mu_2}{\pi\ln\alpha}\left(\frac{\mu_1}{\mu_2}\right)^2 \int_{-1}^{1} \ln(\alpha|X'-X|)\phi(X')dX',\\
A_1(\Gamma) &: (H^{1/2})'(-1,1) \longrightarrow \widetilde{H}^{1/2}(-1,1),\\
A_1(\Gamma)\phi(X) &= C(\Gamma)\int_{-1}^{1}\phi(X')dX',\\
A_{-1} &: (H^{1/2})'(-1,1) \longrightarrow \widetilde{H}^{1/2}(-1,1),\\
A_{-1}\phi(X) &= 2\mu_1 c_{l_0}^a \int_{-1}^{1}\phi(X')dX',
\end{aligned}$$

where $\Gamma \to C(\Gamma)$ is a holomorphic function in the disc $D_{\delta_0}(0)$,

$$c_l^a = \frac{1}{i}\frac{\sqrt{d^2-\lambda_l^a}\,\phi_a(h,\lambda_l^a)^2}{\frac{\mu_1}{\mu_2}\phi_a(h,\lambda_l^a)^2 + 2\sqrt{d^2-\lambda_l^a}\int_0^h \phi(y,\lambda_l^a)^2 dy}, \quad 1 \leq l \leq m_a, \tag{4.2}$$

and

(4.3)

$$a_0 = a_0(\eta) = (1-\eta)\sum_{k=0}^{+\infty}\left(\sum_{p=0}^{k} C_k^p(-2)^p \frac{C_{2(p+1)}^{(p+1)}}{2^{2(p+1)}}\right)\eta^k \text{ for } \eta = \frac{(\frac{\mu_1}{\mu_2})^2 - 1}{(\frac{\mu_1}{\mu_2})^2 + 1}.$$

In order to prove this theorem it is convenient to first establish two technical auxiliary results. These results are stated in the following lemmas.

LEMMA 4.2. *For every $y \in [0,h]$ we have*

$$\phi_a(y,\lambda) = \sin(\sqrt{\lambda}y) - \frac{\cos(\sqrt{\lambda}y)}{2\sqrt{\lambda}}\int_0^y q(\tau)d\tau + O\left(\frac{1}{\lambda}\right), \tag{4.4}$$

$$\partial_y\phi_a(y,\lambda) = \sqrt{\lambda}\cos(\sqrt{\lambda}y) + \frac{1}{2}\sin(\sqrt{\lambda}y)\int_0^y q(\tau)d\tau + O\left(\frac{1}{\sqrt{\lambda}}\right), \tag{4.5}$$

where $O(\frac{1}{\lambda})$ and $O(\frac{1}{\sqrt{\lambda}})$ are uniform in $y \in [0,h]$.

Proof. Let $y \in]0,h]$. Multiplying the ODE (3.2) by $\sin\sqrt{\lambda}(y-\tau)$ and integrating by parts over $]0,y[$ with respect to the variable τ, we obtain that

$$\int_0^y q(\tau)\sin(\sqrt{\lambda}(y-\tau))\phi_a(\tau,\lambda)d\tau + \sqrt{\lambda}(\sin(\sqrt{\lambda}y) - \phi_a(y,\lambda)) = 0.$$

Therefore

$$\phi_a(y,\lambda) = \sin(\sqrt{\lambda}y) + \frac{1}{\sqrt{\lambda}}\int_0^y q(\tau)\sin(\sqrt{\lambda}(y-\tau))\phi_a(\tau,\lambda)d\tau. \tag{4.6}$$

Let $\theta(\lambda) = \max 0 \le y \le h|\phi_a(y,\lambda)|$. Since $q(\tau) \ge 0$, for $\tau \in (0,h)$, it immediately follows from the last equation that

$$\theta(\lambda) \le 1 + \frac{\theta(\lambda)}{\sqrt{\lambda}}\int_0^h q(\tau)d\tau,$$

which gives that

$$\theta(\lambda) \le \frac{1}{1 - \frac{1}{\sqrt{\lambda}}\int_0^h q(\tau)d\tau} \le \frac{1}{2} \quad \text{as } \lambda \to +\infty,$$

and so

$$\phi_a(y,\lambda) = \sin(\sqrt{\lambda}y) + O\left(\frac{1}{\sqrt{\lambda}}\right),$$

where $O(\frac{1}{\sqrt{\lambda}})$ is uniform in $y \in [0,h]$. Substituting now the last expansion into (4.6) and using the piecewise differentiability of the function $\varepsilon_1(\tau)$, we get

$$\phi_a(y,\lambda) = \sin(\sqrt{\lambda}y) - \frac{\cos(\sqrt{\lambda}y)}{2\sqrt{\lambda}}\int_0^y q(\tau)d\tau + O\left(\frac{1}{\lambda}\right).$$

Multiplying the ODE (3.2) by $\cos(\sqrt{\lambda}(y-\tau))$ and integrating by parts over $]0,h[$, we find that

$$\partial_y\phi_a(y,\lambda) = \sqrt{\lambda}\cos\sqrt{\lambda}y + \frac{1}{\sqrt{\lambda}}\int_0^y q(\tau)\cos(\sqrt{\lambda}(y-\tau))\phi_a(\tau,\lambda)d\tau. \tag{4.7}$$

Therefore, substituting (4.4) into (4.7) and using one more time the piecewise differentiability of $\varepsilon_1(\tau)$ leads to the promised result. □

Now, multiplying (3.2) by $\partial_y\phi_a(y,\lambda)$ and integrating by parts over $]0,h[$, we obtain that

$$\partial_y\phi_a^2(h,\lambda)+\lambda\phi_a^2(h,\lambda)=\lambda+2\int_0^h q(y)\phi_a(y,\lambda)\partial_y\phi_a(y,\lambda)dy.$$

As a consequence, we calculate

$$\begin{aligned}\partial_y\phi_a^2(h,\lambda)+\left(\frac{\mu_1}{\mu_2}\right)^2(\lambda-d^2)\phi_a^2(h,\lambda)&=\lambda\left[1+\left(\left(\frac{\mu_1}{\mu_2}\right)^2-1\right)\phi_a^2(h,\lambda)\right]\\&\quad+2\int_0^h q(y)\phi_a(y,\lambda)\partial_y\phi_a(y,\lambda)dy-\left(\frac{\mu_1}{\mu_2}\right)^2d^2\phi_a^2(h,\lambda).\end{aligned}$$

Since $y\to q(y)$ is $\mathcal{C}^1$-piecewise, the function of λ given by $\displaystyle\int_0^h q(y)\phi_a(y,\lambda)\partial_y\phi_a(y,\lambda)dy$ is $O(1)$ as $\lambda\to+\infty$. A direct application of Lemma 4.2 yields

$$\frac{\phi_a^2(h,\lambda)}{(\frac{\mu_1}{\mu_2})^2(\lambda-d^2)\phi_a(h,\lambda)^2+\partial_y\phi_a(h,\lambda)^2}=\frac{\phi_a^2(h,\lambda)}{\lambda[1+((\frac{\mu_1}{\mu_2})^2-1)\phi_a^2(h,\lambda)]}+R_{2,a}(\lambda),$$

where

(4.8)

$$\begin{aligned}R_{2,a}(\lambda)&=\phi_a^2(h,\lambda)\frac{-2\int_0^h q(y)\phi_a(y,\lambda)\partial_y\phi_a(y,\lambda)dy+(\frac{\mu_1}{\mu_2})^2d^2\phi_a^2(h,\lambda)}{\lambda[1+((\frac{\mu_1}{\mu_2})^2-1)\phi_a^2(h,\lambda)](({\frac{\mu_1}{\mu_2}})^2(\lambda-d^2)\phi_a(h,\lambda)^2+\partial_y\phi_a(h,\lambda)^2)}\\&=O\left(\frac{1}{\lambda^2}\right)\quad\text{as }\lambda\to+\infty.\end{aligned}$$

From Lemma 4.2 it also follows that

$$\frac{\phi_a^2(h,\lambda)}{\lambda[1+((\frac{\mu_1}{\mu_2})^2-1)\phi_a^2(h,\lambda)]}=\frac{\sin^2(h\sqrt{\lambda})}{\lambda[1+((\frac{\mu_1}{\mu_2})^2-1)\sin^2(h\sqrt{\lambda})]}+R_{3,a}(\lambda),$$

where

(4.9)

$$\begin{aligned}R_{3,a}(\lambda)&=\frac{(1-(\frac{\mu_1}{\mu_2})^2)\phi_a^2(h,\lambda)(\phi_a^2(h,\lambda)-\sin^2(h\sqrt{\lambda}))}{\lambda[1+((\frac{\mu_1}{\mu_2})^2-1)\sin^2(h\sqrt{\lambda})][1+((\frac{\mu_1}{\mu_2})^2-1)\phi_a^2(h,\lambda)]}\\&\quad+\frac{\phi_a^2(h,\lambda)-\sin^2(h\sqrt{\lambda})}{\lambda[1+((\frac{\mu_1}{\mu_2})^2-1)\sin^2(h\sqrt{\lambda})]}\\&=O\left(\frac{1}{\lambda^{3/2}}\right)\quad\text{as }\lambda\to+\infty.\end{aligned}$$

Combining identities (4.8) and (4.9), we conclude that

$$\begin{aligned}&\frac{\sqrt{\lambda-d^2}\phi_a^2(h,\lambda)}{(\frac{\mu_1}{\mu_2})^2(\lambda-d^2)\phi_a(h,\lambda)^2+\partial_y\phi_a(h,\lambda)^2}\frac{e^{-\alpha|X'-X|\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}}}{\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}}\\&=\frac{\sin^2(h\sqrt{\lambda})}{1+((\frac{\mu_1}{\mu_2})^2-1)\sin^2(h\sqrt{\lambda})}\frac{e^{-\alpha|X'-X|\sqrt{\lambda}}}{\lambda}+R_a^\alpha(\lambda,\Gamma,X'-X),\end{aligned}$$

where

$$R_a^\alpha(\lambda,\Gamma,X'-X)=R_{1,a}^\alpha(\lambda,\Gamma,X'-X)\frac{\sin^2(h\sqrt{\lambda})}{\lambda[1+((\frac{\mu_1}{\mu_2})^2-1)\sin^2(h\sqrt{\lambda})]}$$
$$+\left(R_{2,a}(\lambda)+R_{3,a}(\lambda)\right)e^{-\alpha|X'-X|\sqrt{\lambda}}+R_{1,a}^\alpha(\lambda,\Gamma,X'-X)\left(R_{2,a}(\lambda)+R_{3,a}(\lambda)\right).$$

Here the remainder $R_{1,a}^\alpha$ is given by

$$\begin{aligned}R_{1,a}^\alpha(\lambda,\Gamma,X'-X) &= \frac{(\lambda_{l_0}^a+\Gamma^2-d^2)e^{-\alpha|X'-X|\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}}}{\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}\Big(\sqrt{\lambda-d^2}+\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}\Big)}\\ &+ e^{-\alpha|X'-X|\sqrt{\lambda}}\Big(e^{\alpha|X'-X|\frac{\lambda_{l_0}^a+\Gamma^2}{\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}+\sqrt{\lambda}}}-1\Big).\end{aligned}\tag{4.10}$$

Moreover, $R_a^\alpha(\lambda,\Gamma,X'-X)=O(\frac{1}{\lambda^{3/2}})$ as $\lambda\to+\infty$, where $O(\frac{1}{\lambda^{3/2}})$ is uniform in $\alpha\in D_{\alpha_0}(0)$, $\Gamma\in D_{\delta_0}(0)$, and $(X-X')\in[-2,2]$. On the other hand, we have $R_a^0(\lambda,X'-X)=R_a^0(\lambda,\Gamma)$ and $R_a^\alpha(\lambda,\Gamma,X'-X)-R_a^0(\lambda,\Gamma)=O(\frac{1}{\lambda})$. Consequently, the following estimate holds:

$$\int_{d^2}^{+\infty}\frac{\sqrt{\lambda-d^2}\phi_a(h,\lambda)^2}{(\frac{\mu_1}{\mu_2})^2(\lambda-d^2)\phi_a(h,\lambda)^2+\partial_y\phi_a(h,\lambda)^2}\frac{e^{-\alpha|X'-X|\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}}}{\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}}d\lambda\tag{4.11}$$
$$=\int_{d^2}^{+\infty}\frac{\sin^2(h\sqrt{\lambda})}{1+((\frac{\mu_1}{\mu_2})^2-1)\sin^2(h\sqrt{\lambda})}\frac{e^{-\alpha|X'-X|\sqrt{\lambda}}}{\lambda}d\lambda+\int_{d^2}^{+\infty}R_a^0(\lambda,\Gamma)d\lambda+O(\alpha\ln\alpha),$$

where the remainder $O(\alpha\ln\alpha)$ is holomorphic in the disc $D_{\delta_0}(0)$ and uniform in $(X'-X)\in[-2,2]$.

Remark 4.1. Following the lines of the work of Magnanini and Santosa [8] that is based on the Levitan–Levinson method as described in [3], it is possible to derive an expression for the Green function $H(X,X',Y,Y')$ of the following transmission problem:

$$\begin{cases}\Delta_{X,Y}H(X,X',Y,Y')=\delta(X-X')\delta(Y-Y') & \text{in } O_1\cup O_2,\\ [\mu H]=0 \quad \text{on } L,\\ [\partial_y H]=0 \quad \text{on } L,\\ H=0 \quad \text{on } G.\end{cases}$$

It is quite easy to see from the previous calculations that $H(\alpha X,\alpha X',h^-,h^-)$ is in fact given by

$$H(\alpha X,\alpha X',h^-,h^-)=-\frac{1}{2\pi}\int_0^{+\infty}\frac{\sin^2(h\sqrt{\lambda})}{1+((\frac{\mu_1}{\mu_2})^2-1)\sin^2(h\sqrt{\lambda})}\frac{e^{-\alpha|X'-X|\sqrt{\lambda}}}{\lambda}d\lambda.$$

We should now rigorously derive an asymptotic expansion of $H(\alpha X,\alpha X',h^-,h^-)$ as α approaches 0. To do so, we introduce the parameter

$$\eta=\frac{(\frac{\mu_1}{\mu_2})^2-1}{(\frac{\mu_1}{\mu_2})^2+1}$$

to write

$$\frac{\sin^2(h\sqrt{\lambda})}{1+((\frac{\mu_1}{\mu_2})^2-1)\sin^2(h\sqrt{\lambda})}=\frac{(1-\eta)\sin^2(h\sqrt{\lambda})}{1+\eta\cos(2h\sqrt{\lambda})}$$

and use the following result.

LEMMA 4.3. *Let the function F_η be defined by*

$$F_\eta(X)=\frac{1}{2\pi}\int_0^{+\infty}\frac{\sin^2(h\sqrt{\lambda})}{\cos^2 h\sqrt{\lambda}+\frac{1+\eta}{1-\eta}\sin^2(h\sqrt{\lambda})}\frac{e^{-\alpha X\sqrt{\lambda}}}{\lambda}d\lambda \quad \textit{for } X>0.$$

Under the assumption that $|\eta|<1$ we have

$$F_\eta(X)=-\frac{a_0}{2\pi}\ln\left(\frac{X^2}{X^2+4h^2}\right)+\frac{1}{\pi(1-\eta)}\sum_{k=1}^{+\infty}\left(\sum_{p=1}^{k}C_k^p(-2)^p\sum_{q=1}^{p+1}w_q^{p+1}\ln\left(\frac{X^2}{h^2}+4q^2\right)\right)\eta^k.$$

Furthermore, when X approaches 0^+, we have

$$F_\eta(X)=-\frac{a_0}{\pi}\ln X+a_1+\sum_{n=1}^{+\infty}a_{2n}X^{2n},$$

where

$$\begin{cases}
a_0(\eta)=(1-\eta)\displaystyle\sum_{k=0}^{+\infty}\left(\sum_{p=0}^{k}C_k^p(-2)^p w_0^{p+1}\right)\eta^k,\\
a_1(\eta)=\dfrac{\ln h(1-\eta)}{\pi}\displaystyle\sum_{k=0}^{+\infty}\left(\sum_{p=0}^{k}C_k^p(-2)^p w_0^{p+1}\right)\eta^k\\
\qquad+\dfrac{1-\eta}{\pi}\displaystyle\sum_{k=0}^{+\infty}\left(\sum_{p=0}^{k}C_k^p(-2)^p\left(\sum_{q=1}^{p+1}w_q^{p+1}\ln 4q^2\right)-d_p\right)\eta^k,\\
a_{2n}(\eta)=\dfrac{1-\eta}{\pi}\displaystyle\sum_{k=0}^{+\infty}\left(\sum_{p=0}^{k}C_k^p(-2)^p\sum_{q=1}^{p+1}\frac{w_q^{p+1}}{(2qh)^{2n}}\right)\eta^k, \quad 1\le n,\\
w_0^k=\dfrac{C_{2k}^k}{2^{2k}}, \quad 1\le k,\\
w_p^k=\dfrac{C_{2k}^{k+p}}{2^{2k}p}, \quad 1\le p\le k.
\end{cases}$$

Proof. Since $|\eta|<1$, by using the trigonometric identity $\cos(2h\sqrt{\lambda})=-2\sin^2(h\sqrt{\lambda})+1$, and after some easy manipulations, we obtain the expansion

$$F_\eta(X)=\frac{1-\eta}{\pi}\sum_{k=0}^{+\infty}\left(\sum_{p=0}^{k}C_k^p(-2)^p D_p(X)\right)\eta^k,$$

where

$$D_p(X)=\frac{1}{2}\int_0^{+\infty}\sin^{2(p+1)}(h\sqrt{\lambda})e^{-X\sqrt{\lambda}}\frac{d\lambda}{\lambda}.$$

By applying the substitution $h^2\lambda = \nu^2$ we readily get

$$D_p(X) = \int_0^{+\infty} \sin^{2(p+1)}(\nu) e^{-\nu \frac{X}{h}} \frac{d\nu}{\nu}$$

for $X > 0$, and so

$$\partial_X D_p(X) = -\frac{1}{h} \int_0^{+\infty} \sin^{2(p+1)}(\nu) e^{-\nu \frac{X}{h}} d\nu \quad \forall\, X > 0.$$

Let us now introduce

$$H_p(X) = \int_0^{+\infty} \sin^{2p}(\nu) e^{-\nu X} d\nu.$$

Straightforward integrations by parts yield

$$H_p(X) = \frac{2p(2p-1)}{X^2 + 4p^2} H_{p-1}(X), \quad p \geq 1, \quad H_0(X) = \frac{1}{X},$$

and consequently

$$H_p(X) = \frac{2p!}{X \prod_{q=1}^{p} (X^2 + 4q^2)}, \quad p \geq 1,$$

from which it immediately follows that

$$\partial_X D_p(X) = -\frac{1}{h} H_{p+1}\left(\frac{X}{h}\right), \quad p \geq 0.$$

Thus

$$D_p(X) = -w_0^{p+1} \ln\left(\frac{X}{h}\right) + \sum_{q=1}^{p+1} w_q^{p+1} \ln\left(\frac{X^2}{h^2} + 4q^2\right) + d_p, \quad p \geq 0,$$

where

$$d_p = \int_0^{+\infty} \sin^{2(p+1)}(\nu) e^{-\nu} \frac{d\nu}{\nu} - \sum_{q=1}^{p+1} w_q^{p+1} \ln(1 + 4q^2).$$

Based on the first identity stated at the beginning of the proof it is now quite easy to obtain the desired result. □

Remark 4.2. By Remark 4.1 and the latter lemma, we easily verify that in the particular case $\mu_1 = \mu_2$ (i.e., $\eta = 0$) we have

$$\begin{aligned} H(X, 0, h^-, h^-) = -\frac{1}{2\pi} \int_0^{+\infty} \sin^2(h\sqrt{\lambda}) \frac{e^{-X\sqrt{\lambda}}}{\lambda} d\lambda = -F_0(X) \\ = \frac{1}{4\pi} \ln\left(\frac{X^2}{X^2 + 4h^2}\right), \end{aligned}$$

which is in accordance with the following well-known expression of H,

$$H(X, X', Y, Y') = \frac{1}{4\pi} \ln\left(\frac{(X - X')^2 + (Y - Y')^2}{(X - X')^2 + (Y + Y')^2}\right).$$

We are now ready to proceed with the following proof.

Proof of Theorem 4.1. We start with deriving an asymptotic formula for $a_\alpha(\Gamma; X, X')$ when α approaches 0. With the definition of c_l^a we rewrite

$$a_\alpha(\Gamma; X, X') = \frac{2\mu_1 c_{l_0}^a}{\ln\alpha}\frac{e^{i\alpha|X'-X|\Gamma}}{\Gamma} + \frac{2\mu_1}{\ln\alpha}\sum_{l=1,l\neq l_0}^{m_a} c_l^a \frac{e^{i\alpha|X'-X|\sqrt{\lambda_{l_0}^a-\lambda_l^a+\Gamma^2}}}{\sqrt{\lambda_{l_0}^a-\lambda_l^a+\Gamma^2}}$$
$$- \frac{\mu_2}{2\pi\ln\alpha}\int_{d^2}^{+\infty} \frac{\sqrt{\lambda-d^2}\phi_a(h,\lambda)^2}{(\lambda-d^2)\phi_a(h,\lambda)^2 + (\frac{\mu_2}{\mu_1})^2\partial_y\phi_a(h,\lambda)^2}\frac{e^{-\alpha|X'-X|\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}}}{\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}}d\lambda.$$

Note that the first term in the expression of $a_\alpha(\Gamma; X, X')$ is holomorphic in $\alpha \in D_{\alpha_0}(0)$. Therefore, we only need to handle the last two terms. For the estimation of the second term we use that

$$\frac{\sqrt{\lambda-d^2}}{\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}}e^{-\alpha|X'-X|\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}} = e^{-\alpha|X'-X|\sqrt{\lambda}} + R_{1,a}^\alpha(\lambda,\Gamma,X'-X),$$

where $R_{1,a}^\alpha(\lambda,\Gamma,X'-X)$ is defined by (4.10) and satisfies the estimate

$$R_{1,a}^\alpha(\lambda,\Gamma,X'-X) = O\left(\frac{1}{\sqrt{\lambda}}\right) \tag{4.12}$$

as $\lambda \to +\infty$. The remainder $O(\frac{1}{\sqrt{\lambda}})$ is uniform in $\alpha \in D_{\alpha_0}(0)$, $\Gamma \in D_{\delta_0}(0)$, and $(X-X') \in [-2,2]$. The third term is now easy to estimate. From Lemmas 4.2 and 4.3 we know that

(4.13)
$$-\frac{\mu_2}{2\pi\ln\alpha}\int_{d^2}^{+\infty} \frac{\sqrt{\lambda-d^2}\phi_a(h,\lambda)^2}{(\frac{\mu_1}{\mu_2})^2(\lambda-d^2)\phi_a(h,\lambda)^2 + \partial_y\phi_a(h,\lambda)^2}\frac{e^{-\alpha|X'-X|\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}}}{\sqrt{\lambda-\lambda_{l_0}^a-\Gamma^2}}d\lambda$$
$$= -\frac{a_0\mu_2}{\pi\ln\alpha}\ln(\alpha|X'-X|)$$
$$+ \frac{1}{\ln\alpha}\left(-\mu_2 a_1 + \frac{\mu_2}{2\pi}\int_0^{d^2}\frac{\sin^2(h\sqrt{\lambda})}{1+((\frac{\mu_1}{\mu_2})^2-1)\sin^2(h\sqrt{\lambda})}\frac{d\lambda}{\lambda} - \frac{\mu_2}{2\pi}\int_{d^2}^{+\infty} R_a^0(\lambda,\Gamma)d\lambda\right)$$
$$+ O(\alpha),$$

where $O(\alpha)$ is holomorphic in $D_{\delta_0}(0)$ and uniform in $(X'-X) \in [-2,2]$. Upon insertion of the last expansion into the expression of $a_\alpha(\Gamma; X, X')$ we finally obtain the asymptotic formula

$$a_\alpha(\Gamma; X, X') = -\frac{a_0\mu_2}{\pi\ln\alpha}\left(\frac{\mu_1}{\mu_2}\right)^2\ln(\alpha|X'-X|) + \frac{1}{\ln\alpha}\left(\frac{2\mu_1 c_{l_0}^a}{\Gamma} + C(\Gamma)\right) + O(\alpha),$$

where

$$C(\Gamma) = \left(\frac{\mu_1}{\mu_2}\right)^2\left[-\mu_2 a_1 + \frac{\mu_2}{2\pi}\int_0^{d^2}\frac{\sin^2(h\sqrt{\lambda})}{1+((\frac{\mu_1}{\mu_2})^2-1)\sin^2(h\sqrt{\lambda})}\frac{d\lambda}{\lambda}\right.$$
$$\left. - \frac{\mu_2}{2\pi}\int_{d^2}^{+\infty} R_a^0(\lambda,\Gamma)d\lambda\right] + \sum_{l=1,l\neq l_0}^{m_a}\frac{2\mu_1 c_l^a}{\sqrt{\lambda_{l_0}^a-\lambda_l^a+\Gamma^2}},$$

and $O(\alpha)$ is holomorphic in $\Gamma \in D_{\delta_0}(0)$ and uniform in $(X' - X) \in [-2, 2]$. This completes the proof of the second part of Theorem 4.1.

By construction, we know that $\Gamma \to \mathcal{A}_\alpha(\Gamma)$ is a meromorphic operator-valued function in $D_{\delta_0}(0)$ and has 0 as a unique pole. Moreover, for every $\Gamma \in \overline{D_{\delta_0}(0)}/\{0\}$, $\mathcal{A}_\alpha(\Gamma)$ is an invertible operator. From the asymptotic expansion (4.1), it immediately follows that $\mathcal{A}_\alpha(\Gamma)$ is a finitely meromorphic operator that is of Fredholm type at $\Gamma = 0$ for small values of α. Then, it is in all the domain $D_{\delta_0}(0)$. □

We now study the existence and the distribution of the characteristic values of the operator-valued function $\mathcal{A}_\alpha(\Gamma)$.

The following theorem asserts that there exits a unique resonance Γ_α lying in a small neighborhood of 0.

THEOREM 4.4. *There exists a constant $\alpha_0 > 0$ such that for $|\alpha| < \alpha_0$ we have*

$$\mathcal{M}(\mathcal{A}_\alpha(\Gamma), \partial D_{\delta_0}(0)) = 0.$$

Furthermore, there exists a unique characteristic value Γ_α of the operator-valued function $\Gamma \mapsto \mathcal{A}_\alpha(\Gamma)$ in $D_{\delta_0}(0)$.

Proof. We begin the proof by establishing an explicit formula for $(A_0^\alpha)^{-1}$. Let $\phi \in (H^{1/2})'(]-1,1[)$ and $\psi \in \widetilde{H}^{1/2}(]-1,1[)$ be such that $A_0^\alpha \phi = \psi$, or equivalently

$$\big(1/\phi\big) + \frac{1}{\ln \alpha} \mathcal{L}_0 \phi(X) = -\frac{\pi}{\mu_2 a_0} \left(\frac{\mu_2}{\mu_1}\right)^2 \psi(X), \tag{4.14}$$

where

$$\big(1/\phi\big) = \int_{-1}^{1} \phi(x') dX',$$

and

$$\mathcal{L}_0 \phi(X) = \int_{-1}^{1} \ln|X' - X| \phi(X') dX'.$$

Since the operator $\mathcal{L}_0 : (H^{1/2})'(]-1,1[) \to \widetilde{H}^{1/2}(]-1,1[)$ is invertible, we have

$$\big(1/\phi\big) \mathcal{L}_0^{-1} 1(X) + \frac{1}{\ln \alpha} \phi(X) = \frac{\pi}{\mu_1 a_0} \mathcal{L}_0^{-1} \psi(X).$$

A simple integration over $]-1,1[$ gives

$$\big(1/\phi\big) = -\frac{\pi}{\mu_2 a_0} \left(\frac{\mu_2}{\mu_1}\right)^2 \frac{\big(\mathcal{L}_0^{-1}\psi(X)/1\big)}{\big(\mathcal{L}_0^{-1}1(x)/1\big) + \frac{1}{\ln \alpha}}.$$

By inserting this into (4.14) we immediately obtain

$$\phi(X) = (A_0^\alpha)^{-1}\psi(X) = -\frac{\pi \ln \alpha}{\mu_2 a_0} \left(\frac{\mu_2}{\mu_1}\right)^2 \left[\mathcal{L}_0^{-1}\psi(X) - \frac{\big(\mathcal{L}_0^{-1}\psi(X)/1\big)}{\big(\mathcal{L}_0^{-1}1(x)/1\big) + \frac{1}{\ln \alpha}} \mathcal{L}_0^{-1} 1(X)\right].$$

By the Hilbert inversion formula [9] we may calculate $\big(\mathcal{L}_0^{-1}1(x)/1\big) = \frac{2}{\ln 2}$. We remark that $(A_0^\alpha)^{-1} A_1(\Gamma) = O(1)$ and $(A_0^\alpha)^{-1} A_{-1} = O(1)$. Thus, the following estimate

$$(A_0^\alpha)^{-1}(\mathcal{A}_\alpha(\Gamma) - A_0^\alpha) = O\left(\frac{1}{\ln \alpha}\right)$$

holds uniformly in $\Gamma \in \partial D_{\delta_0}(0)$. From this we deduce that there exists a constant α_0 such that for $|\alpha| \leq \alpha_0$ we have

$$|(A_0^\alpha)^{-1}(\mathcal{A}_\alpha(\Gamma) - A_0^\alpha)|_{\mathcal{L}((H^{1/2})'(]-1,1[),(H^{1/2})'(]-1,1[))} < 1 \quad \forall\, \Gamma \in \partial D_{\delta_0}(0).$$

Due to the generalized Rouché theorem [2] (see Theorem B.2 in Appendix B) we now obtain the following assertion for all $|\alpha| \leq \alpha_0$:

$$\mathcal{M}(\mathcal{A}_\alpha(\Gamma), \partial D_{\delta_0}(0)) = \mathcal{M}(\mathcal{A}_0^\alpha, \partial D_{\delta_0}(0)) = 0.$$

Since 0 is the unique pole of $\mathcal{A}_\alpha(\Gamma)$ in $D_{\delta_0}(0)$, it immediately follows from the latter assertion that there exists a unique characteristic value Γ_α in $D_{\delta_0}(0)$. ∎

We now want to derive an asymptotic expression for Γ_α. Let the operators $\widetilde{A}_k$ be defined by

$$A_1(\Gamma)\Gamma + A_{-1} = \sum_{k=-1}^{+\infty} \widetilde{A}_k \Gamma^{k+1}.$$

In the following theorems we summarize our main findings in this section.

THEOREM 4.5. *There exists a positive constant α_0 such that for $|\alpha| \leq \alpha_0$ the following holds:*

$$\Gamma_\alpha = \sum_{p=1}^{+\infty} \frac{1}{(-\ln \alpha)^p} \operatorname{tr}\left[(\mathcal{A}_0^\alpha)^{-p} \sum_{j=1}^{+\infty} \sum_{\substack{k_1+\cdots+k_j=p-1-j \\ k_s \geq -1}} \widetilde{A}_{k_1} \ldots \widetilde{A}_{k_j}\right] + O(\alpha). \tag{4.15}$$

Furthermore, the leading order term in the asymptotic expansion of Γ_α is given by

$$\Gamma_\alpha = \frac{2\pi c_{l_0}^a \mu_2}{a_0 \mu_1} (\ln \alpha)^{-1} + O((\ln \alpha)^{-2}), \tag{4.16}$$

where $c_{l_0}^a$ and a_0 are defined by (4.2) *and* (4.3)*, respectively.*

Proof. From the generalized Rouché theorem [2] (see Theorem B.4 in Appendix B) we immediately obtain the asymptotic expansion (4.15) for the resonance Γ_α. The leading order term in this expression, given by (4.16), can be easily derived from the asymptotic expansion of $\mathcal{A}_\alpha(\Gamma)$ stated in Theorem 4.1. ∎

The following theorem holds.

THEOREM 4.6. *There exists a constant $\alpha_0 > 0$ such that for $|\alpha| \leq \alpha_0$ we have*

$$\Gamma_\alpha = \frac{1}{2i\pi} \sum_{p=1}^{+\infty} \frac{1}{p} \operatorname{tr}\left[(\mathcal{A}_0^\alpha)^{-p} \int_{\partial D_{\delta_0}} (\mathcal{A}_0^\alpha - \mathcal{A}_\alpha(\Gamma))^p d\Gamma\right].$$

Proof. Recalling that $\mathcal{A}_\alpha(\Gamma)$ is a finitely meromorphic operator-valued function that is of Fredholm type in $D_{\delta_0}(0)$, 0 is the unique pole of $\mathcal{A}_\alpha(\Gamma)$ in $D_{\delta_0}(0)$, and Γ_α is the unique characteristic value of $\mathcal{A}_\alpha(\Gamma)$ in $D_{\delta_0}(0)$, the generalized Rouché theorem implies

$$\Gamma_\alpha = \frac{1}{2i\pi} \operatorname{tr} \int_{\partial D_{\delta_0}} \Gamma(\mathcal{A}_\alpha(\Gamma))^{-1} \frac{\partial}{\partial \Gamma} \mathcal{A}_\alpha(\Gamma) d\Gamma.$$

Since, for $|\alpha| \leq \alpha_0$, we have

$$(\mathcal{A}_\alpha(\Gamma))^{-1} = \sum_{p=0}^{+\infty} (\mathcal{A}_0^\alpha)^{-p} \Big[(\mathcal{A}_0^\alpha - \mathcal{A}_\alpha(\Gamma))(\mathcal{A}_0^\alpha)^{-1}\Big]^p,$$

the expression of Γ_α can alternatively be written as follows:

$$\Gamma_\alpha = \frac{1}{2i\pi} \sum_{p=0}^{+\infty} \text{ tr } \left[(\mathcal{A}_0^\alpha)^{-(p+1)} \int_{\partial D_{\delta_0}} \Gamma (\mathcal{A}_0^\alpha - \mathcal{A}_\alpha(\Gamma))^p \frac{\partial}{\partial \Gamma} \mathcal{A}_\alpha(\Gamma) d\Gamma \right].$$

By noticing that

$$\frac{1}{p+1} \sum_{s=0}^{p} (\mathcal{A}_0^\alpha - \mathcal{A}_\alpha(\Gamma))^s \frac{\partial}{\partial \Gamma} \mathcal{A}_\alpha(\Gamma) (\mathcal{A}_0^\alpha - \mathcal{A}_\alpha(\Gamma))^{p-s} = -\frac{1}{p+1} \frac{\partial}{\partial \Gamma} (\mathcal{A}_0^\alpha - \mathcal{A}_\alpha(\Gamma))^{p+1}$$

and by integrating by parts, we arrive at the desired result. □

It immediately follows from Theorem 4.5 that

$$\frac{1}{2i\pi} \text{ tr } \left[(\mathcal{A}_0^\alpha)^{-(p+1)} \int_{\partial D_{\delta_0}} (\mathcal{A}_0^\alpha - \mathcal{A}_\alpha(\Gamma))^p d\Gamma \right]$$
$$= \left(\frac{-1}{\ln \alpha} \right)^p \text{ tr } \left[(\mathcal{A}_0^\alpha)^{-p} \sum_{j=1}^{+\infty} \sum_{\substack{k_1 + \cdots + k_j = p-1-j \\ k_s \geq -1}} \widetilde{A}_{k_1} \ldots \widetilde{A}_{k_j} \right] + O(\alpha). \quad \square$$

We now construct an asymptotic expansion for the characteristic function corresponding to the characteristic value Γ_α. Let ϕ_α^a be the normalized characteristic function corresponding to the characteristic value Γ_α in $L^2(]-1,1[)$:

$$\phi_\alpha^a(X) = \frac{[\frac{1}{\mu} \partial_y u_\alpha](\alpha X)}{|[\frac{1}{\mu} \partial_y u_\alpha](\alpha X)|_{L^2(]-1,1[)}}.$$

Here we have used the extra-regularity of the characteristic function ϕ_α^a. A combination of the fact that the integral operator with kernel $\ln|X - X'|$ is invertible from the Hölder space

$$\left\{ \varphi \in \mathcal{C}^0(]-1,1[), \sqrt{1-X^2}\varphi(X) \in \mathcal{C}^{0,\delta}([-1,1]) \right\}$$

onto the Hölder space

$$\Big\{ \psi \in \mathcal{C}^{0,\delta}([-1,1]) \cap \mathcal{C}^1(]-1,1[), \sqrt{1-X^2}\psi'(X) \in \mathcal{C}^{0,\delta}(]-1,1[) \cap \mathcal{C}^{0,\nu}([-1,1]),$$
$$\sqrt{1-X^2}\psi'(X)|_{X=\pm 1} = 0 \Big\},$$

where $\mathcal{C}^{0,\delta}$ and $\mathcal{C}^{0,\nu}$ are the Hölder spaces with indices $0 < \nu < \delta < \frac{1}{2}$ (see [1] for a proof), together with the classical Sobolev imbedding theorems, ensures that the characteristic function ϕ_α^a lies in fact in $L^2(]-1,1[)$.

Let P_α^a denote the orthogonal projection on $\text{Ker}(\mathcal{A}_\alpha(\Gamma_\alpha))$:

$$P_\alpha^a = \big(./\phi_\alpha^a(X)\big)_{L^2(]-1,1[)} \phi_\alpha^a(X).$$

The following theorem holds.

THEOREM 4.7. *There exists a positive constant α_0 such that for $|\alpha| \leq \alpha_0$,*

$$P_\alpha^a = P_0^a + O(\alpha),$$

where $P_0^a = \frac{1}{2}\big(./1\big)_{L^2(]-1,1[)}$ *and the remainder* $O(\alpha)$ *is the operator norm.*

Proof. The generalized Rouché theorem implies that

$$P_\alpha^a - P_0^a = \frac{1}{2i\pi}\int_{\partial D_{\delta_0}} (\mathcal{A}_\alpha(\Gamma))^{-1}\frac{\partial}{\partial\Gamma}\mathcal{A}_\alpha(\Gamma)d\Gamma.$$

On the other hand, we have for $|\alpha| \leq \alpha_0$

$$(\mathcal{A}_\alpha(\Gamma))^{-1} = \sum_{p=0}^{+\infty}(\mathcal{A}_0^\alpha)^{-p}[(\mathcal{A}_0^\alpha - \mathcal{A}_\alpha(\Gamma))(\mathcal{A}_0^\alpha)^{-1}]^p,$$

and therefore

$$P_\alpha^a - P_0^a = \sum_{p=0}^{+\infty}(\mathcal{A}_0^\alpha)^{-1}\frac{1}{2i\pi}\int_{\partial D_{\delta_0}}\Big((\mathcal{A}_0^\alpha - \mathcal{A}_\alpha(\Gamma))(\mathcal{A}_0^\alpha)^{-1}\Big)^p\frac{\partial}{\partial\Gamma}\mathcal{A}_\alpha(\Gamma)d\Gamma.$$

Recalling from Theorem 4.5 that

$$\mathcal{A}_0^\alpha - \mathcal{A}_\alpha(\Gamma) = -\frac{1}{\ln\alpha}\Big(A_1(\Gamma) + \frac{A_{-1}}{\Gamma}\Big) + O(\alpha),$$

insertion of the above identity into the expression of P_α^a immediately gives

$$\begin{aligned}
&P_\alpha^a - P_0^a\\
&= \sum_{p=0}^{+\infty}\left(\frac{1}{\ln\alpha}\right)^{p+1}(\mathcal{A}_0^\alpha)^{-1}\frac{1}{2i\pi}\\
&\times\int_{\partial D_{\delta_0}}\left(-\left(A_1(\Gamma)+\frac{A_{-1}}{\Gamma}\right)(\mathcal{A}_0^\alpha)^{-1}\right)^p\frac{\partial}{\partial\Gamma}\left(A_1(\Gamma)+\frac{A_{-1}}{\Gamma}\right)d\Gamma\\
&+ O(\alpha).
\end{aligned}$$

Note that $\mathcal{A}_0^\alpha A_1(\Gamma) = A_1(\Gamma)\mathcal{A}_0^\alpha$ and $\mathcal{A}_0^\alpha A_{-1} = A_{-1}\mathcal{A}_0^\alpha$. By writing now that

$$\left(A_1(\Gamma)+\frac{A_{-1}}{\Gamma}\right)^p\frac{\partial}{\partial\Gamma}\left(A_1(\Gamma)+\frac{A_{-1}}{\Gamma}\right) = \frac{1}{p+1}\frac{\partial}{\partial\Gamma}\left(A_1(\Gamma)+\frac{A_{-1}}{\Gamma}\right)^{p+1},$$

and by integrating by parts, we finally obtain

$$\frac{1}{2i\pi}\int_{\partial D_{\delta_0}}\left(A_1(\Gamma)+\frac{A_{-1}}{\Gamma}\right)^p\frac{\partial}{\partial\Gamma}\left(A_1(\Gamma)+\frac{A_{-1}}{\Gamma}\right)d\Gamma = 0,$$

which is the desired result. □

Direct application of Theorem 4.7 yields

$$\big(1/\phi_\alpha^a(X)\big)^2_{L^2(]-1,1[)} = 2 + O(\alpha).$$

Corollary 4.8. *There exists a constant* α_0 *such that for* $|\alpha| \leq \alpha_0$ *we have*

$$\phi_\alpha^a(X) = \frac{1}{\sqrt{2}} + O(\alpha),$$

where the remainder $O(\alpha)$ *is uniform in* $X \in [-1,1]$.

4.2. The operator-valued function $\mathcal{B}_\alpha(\Gamma)$. Let ω be a fixed positive real, $0 \le l_0 \le m_b$, and $\Gamma = \sqrt{\omega^2\varepsilon_1(0)\mu_1 - \lambda_{l_0}^b - \gamma^2}$ a complex variable, where $\lambda_{l_0}^b(\omega)$ is a root of (3.3) different from $d^2(\omega)$. Then 0 is a resonance of the reference waveguide. In this section, we prove that in a fixed complex neighborhood of 0 lying in the set

$$\mathcal{V}_0 = \left\{ \Gamma \in \mathbb{C}, |\Gamma| < \min_{\lambda \in \{(\lambda_l^b)_{0\le l\le m_b, l\ne l_0}, d^2\}} |\lambda_{l_0}^b - \lambda| \right\},$$

there exists a unique resonance of the microstrip transmission line for small values of α, defined as the unique characteristic value of the function $\mathcal{B}_\alpha(\Gamma)$ in that neighborhood.

Our main results in this section are summarized in the following theorem.

THEOREM 4.9. *Let $D_{\delta_0}(0) = \{\Gamma \in \mathbb{C}, |\Gamma| < \delta_0\}$, $\delta_0 > 0$, a neighborhood of 0 lying in $\mathcal{V}_0$. There exists a positive constant α_0 such that, for $|\alpha| \le \alpha_0$, we have the following:*

(a) *The operator-valued function $\mathcal{B}_\alpha(\Gamma)$ is finitely meromorphic and of Fredolm type at every point of the domain $D_{\delta_0}(0)$.*

(b) *The following asymptotic formula holds:*

$$\mathcal{B}_\alpha(\Gamma) = B_0 + B_1\alpha^2 \ln\alpha + \Big(B_2(\Gamma) + \frac{B_{-1}}{\Gamma}\Big)\alpha^2 + O(\alpha^3 \ln\alpha),$$

where the remainder $O(\alpha^3 \ln\alpha)$ is uniform in $\{\phi \in \widetilde{H}^{1/2}(-1,1), |\phi|_{\widetilde{H}^{1/2}(-1,1)} \le 1\}$ and in $\Gamma \in \partial D_{\delta_0}(0) = \{\Gamma \in \mathbb{C}, |\Gamma| = \delta_0\}$. The operators $B_0^\alpha, B_1(\Gamma)$, and B_{-1} are defined by

$$\begin{aligned}
B_0^\alpha &: \widetilde{H}^{1/2}(-1,1) \longrightarrow H^{-1/2}(-1,1),\\
B_0^\alpha\phi(X) &= -\frac{a_0\varepsilon_2}{\pi\varepsilon_1^2(h)} \int_{-1}^{1} \frac{1}{(X'-X)^2}\phi(X')dX',\\
b_1(\Gamma) &: \widetilde{H}^{1/2}(-1,1) \longrightarrow H^{-1/2}(-1,1),\\
B_1(\Gamma)\phi(X) &= b_1 \int_{-1}^{1} \phi(X')dX',\\
B_2(\Gamma) &: \widetilde{H}^{1/2}(-1,1) \longrightarrow H^{-1/2}(-1,1),\\
B_2(\Gamma)\phi(X) &= \int_{-1}^{1} b_2(\Gamma; X', X)\phi(X')dX',\\
B_{-1} &: \widetilde{H}^{1/2}(-1,1) \longrightarrow H^{-1/2}(-1,1),\\
B_{-1}\phi(X) &= \frac{2c_{l_0}^b}{\varepsilon_2} \int_{-1}^{1} \phi(X')dX'.
\end{aligned}$$

Here $\Gamma \to B_2(\Gamma)$ is an holomorphic operator-valued function in the disc $D_{\delta_0}(0)$.

In order to prove this theorem we will need two technical results. By the same arguments as we went through earlier in the proof of Lemma 4.2, the following result that will prove useful later holds.

LEMMA 4.10. *For every $y \in [0,h]$ we have*

$$\phi_b(y,\lambda) = \cos(\sqrt{\lambda}y) + \frac{\int_0^y q(\tau)d\tau}{2}\frac{\sin(\sqrt{\lambda}y)}{\sqrt{\lambda}} + O\left(\frac{1}{\lambda}\right),$$

$$\partial_y\phi_b(y,\lambda) = -\sqrt{\lambda}\sin(\sqrt{\lambda}y) + \frac{\int_0^y q(\tau)d\tau}{2}\cos(\sqrt{\lambda}y) + \left(\frac{3q(y) - \int_0^y q(\tau)\int_0^\tau q(s)dsd\tau}{4}\right)\frac{\sin(\sqrt{\lambda}y)}{\sqrt{\lambda}} - q(y)\frac{\sin^3(\sqrt{\lambda}y)}{\sqrt{\lambda}} + O\left(\frac{1}{\lambda}\right).$$

Here $O(\frac{1}{\lambda})$ and $O(\frac{1}{\sqrt{\lambda}})$ are uniform in $y \in [0,h]$.

Next, multiplying (3.10) by $\partial_y\phi_b(y,\lambda)$ and integrating over $]0,h[$ yields

$$\partial_y\phi_b^2(h,\lambda) + \phi_b^2(h,\lambda) = \lambda + 2\int_0^h q(y)\phi_b(y,\lambda)\partial_y\phi_b(y,\lambda)dy.$$

Hence

$$(\lambda - d^2)\phi_b^2(h,\lambda) + \left(\frac{\varepsilon_2}{\varepsilon_1(h)}\right)^2\partial_y\phi_b^2(h,\lambda) = \lambda\left[1 + \left(\left(\frac{\varepsilon_2}{\varepsilon_1(h)}\right)^2 - 1\right)\frac{\partial_y\phi_b^2(h,\lambda)}{\lambda}\right] + 2\int_0^h q(y)\phi_b(y,\lambda)\partial_y\phi_b(y,\lambda)dy - d^2\phi_b^2(h,\lambda).$$

Since $y \to (\varepsilon_1(0)\mu_1 - \varepsilon_1(y)\mu_1)$ is $\mathcal{C}^1$-piecewise, the λ-function

$$\omega^2\int_0^h(\varepsilon_1(0)\mu_1 - \varepsilon_1(y)\mu_1(y))\phi_b(y,\lambda)\partial_y\phi_b(y,\lambda)dy$$

is $O(1)$ as $\lambda \to +\infty$. Lemma 4.10 yields

$$\frac{\partial_y\phi_b^2(h,\lambda)}{(\lambda - d^2)\phi_b^2(h,\lambda) + (\frac{\varepsilon_2}{\varepsilon_1(h)})^2\partial_y\phi_b^2(h,\lambda)} = \frac{\partial_y\phi_b^2(h,\lambda)}{\lambda\left[1 + ((\frac{\varepsilon_2}{\varepsilon_1(h)})^2 - 1)\frac{\partial_y\phi_b^2(h,\lambda)}{\lambda}\right]} + R_{3,b}(\lambda),$$

where

$$R_{3,b}(\lambda) = \partial_y\phi_b^2(h,\lambda)\frac{2\omega^2\int_0^h(\varepsilon_1(0)\mu_1 - \varepsilon_1(y)\mu_1(y))\phi_b(y,\lambda)\partial_y\phi_b(y,\lambda)dy - d^2\phi_b^2(h,\lambda)}{\lambda\left[1 + ((\frac{\varepsilon_2}{\varepsilon_1(h)})^2 - 1)\frac{\partial_y\phi_b^2(h,\lambda)}{\lambda}\right]((\lambda - d^2)\phi_b^2(h,\lambda) + (\frac{\varepsilon_2}{\varepsilon_1(h)})^2\partial_y\phi_b^2(h,\lambda))},$$

and $R_{3,b}(\lambda) = O(\frac{1}{\lambda^2})$ as $\lambda \to +\infty$. We also have

$$\frac{\partial_y\phi_b^2(h,\lambda)}{\lambda\left[1 + ((\frac{\varepsilon_2}{\varepsilon_1(h)})^2 - 1)\frac{\partial_y\phi_b^2(h,\lambda)}{\lambda}\right]} = \frac{\sin^2\sqrt{\lambda}h}{1 + ((\frac{\varepsilon_2}{\varepsilon_1(h)})^2 - 1)\sin^2\sqrt{\lambda}h} + \beta_1\frac{\sin\sqrt{\lambda}h\cos\sqrt{\lambda}h}{\sqrt{\lambda}\left[1 + ((\frac{\varepsilon_2}{\varepsilon_1(h)})^2 - 1)\sin^2\sqrt{\lambda}h\right]} + \beta_2\frac{1}{\lambda\left[1 + ((\frac{\varepsilon_2}{\varepsilon_1(h)})^2 - 1)\sin^2\sqrt{\lambda}h\right]} + \beta_3\frac{\sin^2\sqrt{\lambda}h}{\lambda\left[1 + ((\frac{\varepsilon_2}{\varepsilon_1(h)})^2 - 1)\sin^2\sqrt{\lambda}h\right]} + \beta_4\frac{\sin^4\sqrt{\lambda}h}{\lambda\left[1 + ((\frac{\varepsilon_2}{\varepsilon_1(h)})^2 - 1)\sin^2\sqrt{\lambda}h\right]} + R_{4,b}(\lambda),$$

where

$$\begin{cases} \beta_1 = -\displaystyle\int_0^h q(\tau)d\tau, \quad \beta_2 = \left(\frac{\int_0^h q(\tau)d\tau}{2}\right)^2, \\ \beta_3 = \dfrac{1}{2}\displaystyle\int_0^h q(\tau)\int_0^\tau q(s)dsd\tau - \left(\frac{\int_0^h q(\tau)d\tau}{2}\right)^2 - \frac{3}{2}q(h), \\ \beta_4 = 2q(h), \end{cases}$$

and $R_{4,b}(\lambda) = O(\frac{1}{\lambda^{3/2}})$ as $\lambda \to +\infty$.

Finally, we may conclude that

$$\begin{aligned} &\frac{\sqrt{\lambda - d^2}\partial_y\phi_b^2(h,\lambda)}{(\lambda-d^2)\phi_b^2(h,\lambda) + (\frac{\varepsilon_2}{\varepsilon_1(h)})^2\partial_y\phi_b^2(h,\lambda)} \frac{e^{-\alpha|X'-X|\sqrt{\lambda-\lambda_{l_0}^b-\Gamma^2}}}{\sqrt{\lambda-\lambda_{l_0}^b-\Gamma^2}} \\ &= \frac{\sin^2\sqrt{\lambda}h}{1+((\frac{\varepsilon_2}{\varepsilon_1(h)})^2-1)\sin^2\sqrt{\lambda}h} e^{-\alpha|X'-X|\sqrt{\lambda}} \\ &+ \frac{e^{-\alpha|X'-X|\sqrt{\lambda}}}{\sqrt{\lambda}}\Bigg[\beta_1\frac{\sin\sqrt{\lambda}h\cos\sqrt{\lambda}h}{1+((\frac{\varepsilon_2}{\varepsilon_1(h)})^2-1)\sin^2\sqrt{\lambda}h} \\ &\qquad + \alpha|X-X'|\frac{(\lambda_{l_0}^b+\Gamma^2)}{2}\frac{\sin^2\sqrt{\lambda}h}{1+((\frac{\varepsilon_2}{\varepsilon_1(h)})^2-1)\sin^2\sqrt{\lambda}h}\Bigg] \\ &+ \frac{e^{-\alpha|X'-X|\sqrt{\lambda}}}{\lambda}\Bigg[\beta_2\frac{1}{1+((\frac{\varepsilon_2}{\varepsilon_1(h)})^2-1)\sin^2\sqrt{\lambda}h} + \beta_3\frac{\sin^2\sqrt{\lambda}h}{1+((\frac{\varepsilon_2}{\varepsilon_1(h)})^2-1)\sin^2\sqrt{\lambda}h} \\ &\qquad + \beta_4\frac{\sin^4\sqrt{\lambda}h}{1+((\frac{\varepsilon_2}{\varepsilon_1(h)})^2-1)\sin^2\sqrt{\lambda}h} \\ &\qquad + \alpha|X-X'|\beta_1\frac{(\lambda_{l_0}^b+\Gamma^2)}{2}\frac{\sin\sqrt{\lambda}h\cos\sqrt{\lambda}h}{1+((\frac{\varepsilon_2}{\varepsilon_1(h)})^2-1)\sin^2\sqrt{\lambda}h}\Bigg] \\ &+ \alpha^2 R_{1,b}^\alpha(\lambda,\Gamma,X'-X)\frac{\sin^2\sqrt{\lambda}h}{1+((\frac{\varepsilon_2}{\varepsilon_1(h)})^2-1)\sin^2\sqrt{\lambda}h} + R_{5,b}^\alpha(\lambda,\Gamma,X'-X), \end{aligned}$$

where $R_{1,b}^\alpha(\lambda,\Gamma,X'-X) = O(\frac{1}{\lambda})$, and $R_{5,b}^\alpha(\lambda,\Gamma,X'-X)$ is uniform in $\alpha \in D_{\alpha_0}(0)$, $\Gamma \in D_{\delta_0}(0)$, and $(X-X') \in [-2,2]$.

Let us now introduce the following functions. For $X > 0$ let

$$E_\eta^0(X) = \int_{d^2}^{+\infty} \frac{\sin^2\sqrt{\lambda}h}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h} e^{-X\sqrt{\lambda}}d\lambda,$$

$$E_\eta^1(X) = \int_{d^2}^{+\infty} \frac{\sin\sqrt{\lambda}h\cos\sqrt{\lambda}h}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h} \frac{e^{-X\sqrt{\lambda}}}{\sqrt{\lambda}}d\lambda,$$

$$E_\eta^2(X) = \int_{d^2}^{+\infty} \frac{1}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h} \frac{e^{-X\sqrt{\lambda}}}{\lambda}d\lambda,$$

$$E^3_\eta(X) = \int_{d^2}^{+\infty} \frac{\sin^2\sqrt{\lambda}h}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h} \frac{e^{-X\sqrt{\lambda}}}{\lambda} d\lambda,$$

$$E^4_\eta(X) = \int_{d^2}^{+\infty} \frac{\sin^4\sqrt{\lambda}h}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h} \frac{e^{-X\sqrt{\lambda}}}{\lambda} d\lambda,$$

$$E^5_\eta(X) = \int_{d^2}^{+\infty} \frac{\sin^2\sqrt{\lambda}h}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h} \frac{e^{-X\sqrt{\lambda}}}{\sqrt{\lambda}} d\lambda,$$

$$E^6_\eta(X) = \int_{d^2}^{+\infty} \frac{\sin\sqrt{\lambda}h\cos\sqrt{\lambda}h}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h} \frac{e^{-X\sqrt{\lambda}}}{\lambda} d\lambda.$$

By the same arguments as we went through in the proof of Lemma 4.3 we may quite easily show that the following holds.

LEMMA 4.11. *Suppose that* $|\eta| < 1$. *Then we have*

$$E^0_\eta(X) = \frac{e_0}{X^2} + f_0 + O(X),$$

$$E^1_\eta(X) = f_1 + O(X),$$

$$E^2_\eta(X) = e_1 \ln X + f_2 + O(X),$$

$$E^3_\eta(X) = e_2 \ln X + f_3 + O(X),$$

$$E^4_\eta(X) = e_3 \ln X + f_4 + O(X),$$

$$E^5_\eta(X) = \frac{e_4}{X} + f_5 + O(X),$$

$$E^6_\eta(X) = O(1),$$

where

$$e_0(\eta) = 2a_0(\eta),$$

$$f_0(\eta) = (1-\eta)\sum_{k=0}^{+\infty}\sum_{p=0}^{k} C^p_k(-2)^p \sum_{q=1}^{p+1} \frac{w_q^{p+1}}{q^2}\eta^k - \int_0^{d^2} \frac{\sin^2\sqrt{\lambda}h}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h} d\lambda,$$

$$f_1(\eta) = (1-\eta)\sum_{k=0}^{+\infty}\sum_{p=0}^{k} C^p_k(-2)^p \frac{w_0^{p+1} - \sin^{2(p+1)} hd}{h(p+1)}\eta^k,$$

$$e_1(\eta) = (1-\eta)\left(2 - \sum_{k=0}^{+\infty}\sum_{p=0}^{k} C_k^p (-2)^{p+1} w_0^{p+1}\eta^k\right),$$

$$f_2(\eta) = (1-\eta)\left(\int_{d^2}^{+\infty} \frac{e^{-\sqrt{\lambda}}}{\lambda} d\lambda + 2\sum_{n=1}^{+\infty}\frac{(-d)^n}{n!n} + \sum_{k=0}^{+\infty}\sum_{p=0}^{k} C_k^p (-2)^{p+1}\left(\int_0^d \frac{\sin^{2p} h\nu}{\nu} d\nu - w_0^p \ln h - \sum_{q=1}^{p} w_q^p \ln 4q^2\right)\eta^k\right),$$

$$e_2(\eta) = -2a_0(\eta),$$

$$f_3(\eta) = 2a_1(\eta) - \int_0^{d^2} \frac{\sin^2\sqrt{\lambda}h}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h}\frac{d\lambda}{\lambda},$$

$$e_3(\eta) = (1-\eta)\sum_{k=0}^{+\infty}\sum_{p=0}^{k} C_k^{p+1}(-2)^{p+2} w_0^{p+2}\eta^k,$$

$$f_4(\eta) = -(1-\eta)\sum_{k=0}^{+\infty}\sum_{p=0}^{k} C_k^{p+1}(-2)^{p+2} w_0^{p+2}\eta^k \ln h - (1-\eta)\sum_{k=0}^{+\infty}\sum_{p=0}^{k} C_k^{p+1}(-2)^{p+2}\left(\sum_{q=1}^{p+2} w_q^{p+2}\ln 4q^2 + d_{p+1}\right)\eta^k - \int_{d^2}^{+\infty} \frac{\sin^4\sqrt{\lambda}h}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h}\frac{d\lambda}{\lambda},$$

$$e_4(\eta) = -h(1-\eta)\sum_{k=0}^{+\infty}\sum_{p=0}^{k} C_k^{p}(-2)^{p+1} w_0^{p+1}\eta^k,$$

$$f_5(\eta) = -\int_{d^2}^{+\infty} \frac{\sin^2\sqrt{\lambda}h}{\cos^2\sqrt{\lambda}h + \frac{1+\eta}{1-\eta}\sin^2\sqrt{\lambda}h}\frac{d\lambda}{\sqrt{\lambda}}.$$

Here the terms $O(X)$ and $O(1)$ are uniform in $X \in [0,2]$ and holomorphic with respect to $\Gamma \in \overline{D_{\delta_0}(0)}$.

We are now ready to proceed with the following proof.

Proof of Theorem 4.9. We begin by deriving an asymptotic formula for $b_\alpha(\Gamma; X, X')$ when α approaches 0. Let us define

$$c_l^b = \frac{1}{2i}\frac{\sqrt{d^2-\lambda_l^b}\,\partial_y\phi_b^2(h,\lambda_l^b)}{\frac{\varepsilon_2}{\varepsilon_1(h)}\phi_b^2(h,\lambda_l^b) + \sqrt{d^2-\lambda_l^b}\int_0^h \phi^2(y,\lambda_l^b)dy}, \qquad 1 \le l \le m_b.$$

With this definition we may rewrite the expression of the kernel $b_\alpha(\Gamma; X', X)$ as follows:

$$b_\alpha(\Gamma; X, X') = \alpha^2 \frac{2}{\varepsilon_2} c^b_{l_0} \frac{e^{i\alpha|X'-X|\Gamma}}{\Gamma} + \alpha^2 \frac{2}{\varepsilon_2} \sum_{l=1, l\neq l_0}^{m_b} c^b_l \frac{e^{i\alpha|X'-X|\sqrt{\lambda^b_{l_0}-\lambda^b_l+\Gamma^2}}}{\sqrt{\lambda^b_{l_0}-\lambda^b_l+\Gamma^2}}$$
$$-\frac{\alpha^2\varepsilon_2}{2\pi\varepsilon_1^2(h)} \int_{d^2}^{+\infty} \frac{\sqrt{\lambda-d^2}\partial_y\phi_b^2(h,\lambda)}{(\lambda-d^2)\phi_b^2(h,\lambda)+(\frac{\varepsilon_2}{\varepsilon_1(h)})^2\partial_y\phi_b^2(h,\lambda)} \frac{e^{-\alpha|X'-X|\sqrt{\lambda-\lambda^b_{l_0}-\Gamma^2}}}{\sqrt{\lambda-\lambda^b_{l_0}-\Gamma^2}} d\lambda.$$

We start with estimating the third term in the last expression. We have

$$\frac{\sqrt{\lambda-d^2}}{\sqrt{\lambda-\lambda^b_{l_0}-\Gamma^2}} e^{-\alpha|X'-X|\sqrt{\lambda-\lambda^b_{l_0}-\Gamma^2}} = e^{-\alpha|X'-X|\sqrt{\lambda}}$$
$$+\alpha|X-X'|\frac{(\lambda^b_{l_0}+\Gamma^2)}{2}\frac{e^{-\alpha|X'-X|\sqrt{\lambda}}}{\sqrt{\lambda}} + \frac{(\lambda^b_{l_0}+\Gamma^2-d^2)}{2}\frac{e^{-\alpha|X'-X|\sqrt{\lambda}}}{\lambda}$$
$$+\alpha^2 R^\alpha_{1,b}(\lambda,\Gamma,X'-X) + R^\alpha_{2,a}(\lambda,\Gamma,X'-X),$$

where $R^\alpha_{1,b}(\lambda,\Gamma,X'-X) = O(\frac{1}{\lambda})$, and $R^\alpha_{2,b}(\lambda,\Gamma,X'-X) = O(\frac{1}{\lambda^{3/2}})$ when $\lambda \to +\infty$. The remainders $O(\frac{1}{\lambda^s})$, for $s = 1, 3/2$, are uniform in $\alpha \in D_{\alpha_0}(0)$, $\Gamma \in D_{\delta_0}(0)$, and $(X - X') \in [-2, 2]$. A forward application of the last lemma yields

$$-\frac{\alpha^2}{2\pi}\int_{d^2}^{+\infty} \frac{\sqrt{\lambda-d^2}\partial_y\phi_b^2(h,\lambda)}{(\lambda-d^2)\phi_b^2(h,\lambda)+(\frac{\varepsilon_2}{\varepsilon_1(h)})^2\partial_y\phi_b^2(h,\lambda)} \frac{e^{-\alpha|X'-X|\sqrt{\lambda-\lambda^b_{l_0}-\Gamma^2}}}{\sqrt{\lambda-\lambda^b_{l_0}-\Gamma^2}} d\lambda$$
$$= -\frac{a_0}{\pi}\frac{1}{(X'-X)^2} - \frac{1}{2\pi}(\beta_2 e_1+\beta_3 e_2+\beta_4 e_3)\alpha^2\ln\alpha$$
$$-\frac{1}{2\pi}\Big(f_0+\beta_1 f_1+\beta_2 f_2+\beta_3 f_3+\beta_4 f_4+\frac{(\lambda^b_{l_0}+\Gamma^2)e_4}{2}$$
$$+\ (\beta_2 e_1+\beta_3 e_2+\beta_4 e_3)\ln|X'-X|\Big)\alpha^2 + O(\alpha^3\ln\alpha),$$

and, therefore, we finally have

$$b_\alpha(\Gamma; X, X') = -\frac{a_0\varepsilon_2}{\pi\varepsilon_1^2(h)}\frac{1}{(X'-X)^2} + b_1\alpha^2\ln\alpha$$
$$+\ \Big(b_2(\Gamma; X, X')\alpha^2 + \frac{2c^b_{l_0}}{\varepsilon_1(h)}\frac{1}{\Gamma}\Big)\alpha^2 + O(\alpha^3\ln\alpha),$$

where

$$\begin{cases} b_1 = -\dfrac{\varepsilon_2}{2\pi\varepsilon_1^2(h)}(\beta_2 e_1+\beta_3 e_2+\beta_4 e_3), \\ b_2(\Gamma; X, X') = -\dfrac{\varepsilon_2}{2\pi\varepsilon_1^2(h)}\Big(f_0+\beta_1 f_1+\beta_2 f_2+\beta_3 f_3+\beta_4 f_4\dfrac{(\lambda^b_{l_0}+\Gamma^2)e_4}{2} \\ \qquad\qquad +(\beta_2 e_1+\beta_3 e_2+\beta_4 e_3)\ln|X'-X|\Big) \\ \qquad +\frac{2}{\varepsilon_2}\sum_{l=1,l\neq l_0}^{m_b}\frac{c^b_l}{\sqrt{\lambda^a_{l_0}-\lambda^b_l+\Gamma^2}}. \end{cases}$$

Here the term $O(\alpha^3 \ln \alpha)$ is uniform in $X' - X \in [-2, 2]$ and holomorphic in $\Gamma \in \overline{D_{\delta_0}(0)}$, and so the proof of the second part of Theorem 4.9 is now over. By construction, we know that $\Gamma \to \mathcal{B}_\alpha(\Gamma)$ is meromorphic operator-valued in $D_{\delta_0}(0)$ and has a unique pole that is 0. Moreover, for every $\Gamma \in \overline{D_{\delta_0}(0)}/\{0\}$, $\mathcal{B}_\alpha(\Gamma)$ is an invertible operator. From the asymptotic expansion in Theorem 4.9, it immediately follows that $\mathcal{B}_\alpha(\Gamma)$ is a finitely meromorphic and of Fredholm type at $\Gamma = 0$ for small values of α. Then, it is the case in all the domain $D_{\delta_0}(0)$. □

We now study the existence and the localization of the characteristic values of the operator-valued function $\mathcal{B}_\alpha(\Gamma)$.

THEOREM 4.12. *There exists a constant $\alpha_0 > 0$ such that for $|\alpha| < \alpha_0$ we have*

$$\mathcal{M}(\mathcal{B}_\alpha(\Gamma), \partial D_{\delta_0}(0)) = 0,$$

and, therefore, there exists a unique characteristic value Γ_α^b of $\mathcal{B}_\alpha(\Gamma)$ in $D_{\delta_0}(0)$ which is the unique resonance in this domain.

Proof. Since B_0 is an invertible operator it follows from Theorem 4.9 that $(B_0)^{-1}(\mathcal{B}_\alpha(\Gamma) - B_0) = O(\alpha^2 \ln \alpha)$ uniformly in $\Gamma \in \partial D_{\delta_0}(0)$. Therefore, there exists a constant α_0 such that for $|\alpha| \leq \alpha_0$ we have

$$|(B_0)^{-1}(\mathcal{B}_\alpha(\Gamma) - B_0)|_{\mathcal{L}(\widetilde{H}^{1/2}(]-1,1[), \widetilde{H}^{1/2}(]-1,1[))} < 1 \quad \forall\, \Gamma \in \partial D_{\delta_0}(0).$$

By the generalized Rouché theorem (see Theorem B.2 in Appendix B), we immediately obtain for all $|\alpha| \leq \alpha_0$ that

$$\mathcal{M}(\mathcal{B}_\alpha(\Gamma), \partial D_{\delta_0}(0)) = \mathcal{M}(\mathcal{B}_0, \partial D_{\delta_0}(0)) = 0.$$

Since 0 is the unique pole of $\mathcal{B}_\alpha(\Gamma)$ in $D_{\delta_0}(0)$ there exists a unique characteristic value Γ_α^b in $D_{\delta_0}(0)$. Theorem 4.12 is then proved. □

We now derive an asymptotic expansion for Γ_α^b. Let us define

$$\mathcal{L}_b = -\frac{\pi \varepsilon_1^2(h)}{a_0 \varepsilon_2} B_0.$$

THEOREM 4.13. *There exists a positive constant α_0 such that, for $|\alpha| \leq \alpha_0$, we have*

$$\Gamma_\alpha^b = \frac{1}{2i\pi} \sum_{p=1}^{+\infty} \frac{1}{p} \mathrm{tr} \left[(\mathcal{B}_0)^{-p} \int_{\partial D_{\delta_0}} (\mathcal{B}_0 - \mathcal{B}_\alpha(\Gamma))^p d\Gamma \right],$$

which gives

$$\Gamma_\alpha^b = 2\pi \frac{c_{l_0}^b \varepsilon_2}{a_0 \varepsilon_1(h)} \big(\mathcal{L}_b^{-1} 1/1\big) \alpha^2 + O(\alpha^4 \ln \alpha).$$

Proof. Recalling that $\mathcal{B}_\alpha(\Gamma)$ is a finitely meromorphic operator-valued function and of Fredholm type in $D_{\delta_0}(0)$, 0 is the unique pole of $\mathcal{B}_\alpha(\Gamma)$ in $D_{\delta_0}(0)$, and Γ_α is the unique characteristic value of $\mathcal{B}_\alpha(\Gamma)$ in $D_{\delta_0}(0)$, the generalized Rouché theorem (see Theorem B.4 in Appendix B) immediately implies that

$$\Gamma_\alpha^b = \frac{1}{2i\pi} \mathrm{tr} \int_{\partial D_{\delta_0}} \Gamma (\mathcal{B}_\alpha(\Gamma))^{-1} \frac{\partial}{\partial \Gamma} \mathcal{B}_\alpha(\Gamma) d\Gamma.$$

Since, for $|\alpha| \leq \alpha_0$, we have

$$(\mathcal{B}_\alpha(\Gamma))^{-1} = \sum_{p=0}^{+\infty} (\mathcal{B}_0)^{-p} [(\mathcal{B}_0 - \mathcal{B}_\alpha(\Gamma))(\mathcal{B}_0)^{-1}]^p,$$

the expression of Γ_α^b may be rewritten as follows:

$$\Gamma_\alpha^b = \frac{1}{2i\pi} \sum_{p=0}^{+\infty} \mathrm{tr} \left[(\mathcal{B}_0)^{-(p+1)} \int_{\partial D_{\delta_0}} \Gamma (\mathcal{B}_0 - \mathcal{B}_\alpha(\Gamma))^p \frac{\partial}{\partial \Gamma} \mathcal{B}_\alpha(\Gamma) d\Gamma \right].$$

Now by writing

$$\frac{1}{p+1} \sum_{s=0}^{p} (\mathcal{B}_0 - \mathcal{B}_\alpha(\Gamma))^s \frac{\partial}{\partial \Gamma} \mathcal{B}_\alpha(\Gamma) (\mathcal{B}_0 - \mathcal{B}_\alpha(\Gamma))^{p-s} = -\frac{1}{p+1} \frac{\partial}{\Gamma} (\mathcal{B}_0 - \mathcal{B}_\alpha(\Gamma))^{p+1}$$

and integrating by parts, we obtain the desired result. □

Next, we give an expression for the characteristic function corresponding to the value Γ_α^b. We will calculate the leading order term in this asymptotic expression.

Let

$$\phi_\alpha^b(X) = \frac{[\varepsilon v_\alpha](\alpha X)}{|[\varepsilon v_\alpha](\alpha X)|_{L^2(]-1,1[)}}$$

and

$$P_\alpha^b = \big(./\phi_\alpha^b(X)\big)_{L^2(]-1,1[)} \phi_\alpha^b(X)$$

denote the normalized characteristic function corresponding to the value Γ_α^b in $L^2(]-1,1[)$ and the orthogonal projection on $\mathrm{Ker}(\mathcal{B}_\alpha(\Gamma_\alpha^b))$, respectively.

THEOREM 4.14. *There exists a positive constant α_0 such that for $|\alpha| \leq \alpha_0$*

$$P_\alpha^b = P_0^b + 2\pi \frac{c_{l_0}^b \varepsilon_1^2(h)}{a_0 \varepsilon_2^2} \big(./1\big)_{L^2(]-1,1[)} \mathcal{L}_b^{-1} 1(X) \alpha^2 + O(\alpha^4 \ln \alpha),$$

where $P_0^b = \frac{1}{2}\big(./1\big)_{L^2(]-1,1[)}$.

Proof. The generalized Rouché theorem implies that

$$P_\alpha^b - P_0^b = \frac{1}{2i\pi} \int_{\partial D_{\delta_0}} (\mathcal{B}_\alpha(\Gamma))^{-1} \partial_\Gamma \mathcal{B}_\alpha(\Gamma)\, d\Gamma.$$

On the other hand, we have for $|\alpha| \leq \alpha_0$

$$(\mathcal{B}_\alpha(\Gamma))^{-1} = \sum_{p=0}^{+\infty} (\mathcal{B}_0)^{-p} [(\mathcal{B}_0 - \mathcal{B}_\alpha(\Gamma))(\mathcal{B}_0)^{-1}]^p.$$

Therefore,

$$P_\alpha^b - P_0^b = \sum_{p=0}^{+\infty} (\mathcal{B}_0)^{-1} \frac{1}{2i\pi} \int_{\partial D_{\delta_0}} \big((\mathcal{B}_0 - \mathcal{B}_\alpha(\Gamma))(\mathcal{B}_0)^{-1}\big)^p \partial_\Gamma \mathcal{B}_\alpha(\Gamma)\, d\Gamma.$$

Theorem 4.9 yields

$$\mathcal{B}_0 - \mathcal{B}_\alpha(\Gamma) = -B_1\alpha^2\ln\alpha - \left(B_2(\Gamma) + \frac{B_{-1}}{\Gamma}\right)\alpha^2 + O(\alpha^3\ln\alpha).$$

Inserting the above identity into the expression of P_α^a gives

$$P_\alpha^b - P_0^b = -B_0^{-1}B_{-1} + O(\alpha^4\ln\alpha).$$

Hence

$$P_\alpha^b - P_0^b = 2\pi\frac{c_{l_0}^b\varepsilon_1^2(h)}{a_0\varepsilon_2^2}\big(./1\big)_{L^2(]-1,1[)}\mathcal{L}_b^{-1}1(X)\alpha^2 + O(\alpha^4\ln\alpha),$$

and the proof of the theorem is then complete. □

COROLLARY 4.15. *There exists a constant α_0 such that for $|\alpha| \le \alpha_0$*

$$\phi_\alpha^a(X) = \frac{1}{\sqrt{2}} + \sqrt{2}\pi\frac{c_{l_0}^b\varepsilon_1^2(h)}{a_0\varepsilon_2^2}\left(\mathcal{L}_b^{-1}1(X) - \frac{1}{2}\big(\mathcal{L}_b^{-1}1(X)/1\big)_{L^2(]-1,1[)}\right)\alpha^2 + O(\alpha^4\ln\alpha),$$

where the remainder $O(\alpha^4\ln\alpha)$ is uniform in $X \in [-1,1]$.

Proof. Theorem 4.14 gives

$$\big(1/\phi_\alpha^a(X)\big)_{L^2(]-1,1[)} = \sqrt{2}\left(1 + \pi\frac{c_{l_0}^b\varepsilon_1^2(h)}{a_0\varepsilon_2^2}\big(\mathcal{L}_b^{-1}1(X)/1\big)_{L^2(]-1,1[)}\right)\alpha^2 + O(\alpha),$$

which is exactly the identity stated in the corollary. □

Appendix A. Proof of Lemma 3.1. We only give the proof of Lemma 3.1. The proof for Lemma 3.2 follows exactly the same lines. To simplify notations, we will drop the subscript a.

Let $t > 0$ and $\lambda_{k,t}$, $k = 1, 2, \dots$, be the eigenvalues of the following Sturm–Liouville problem in the interval $]0,t[$ with homogeneous Dirichlet boundary condition at $y = t$:

$$\text{(A.1)}\qquad \begin{cases} \partial_{yy}g(y,\lambda;t) + (\lambda - q(y))g(y,\lambda;t) = 0 \quad \text{in }]0,h[\cup]h,t[,\\ [g(.,\lambda;t)] = 0 \quad \text{on } y = h,\\ \left[\dfrac{1}{\mu}\partial_y g(.,\lambda;t)\right] = 0 \quad \text{on } y = h,\\ g(0,\lambda;t) = 0 \quad \text{and} \quad g(t,\lambda;t) = 0. \end{cases}$$

From the properties of the solution g_a to (3.1), we have that $\lambda_{k,t}$ are the roots of

$$\text{(A.2)}\qquad g(t,\lambda_{k,t};t) = 0,$$

or equivalently

$$\text{(A.3)}\qquad \phi(h,\lambda_{k,t})\sqrt{\lambda_{k,t} - d^2} + \frac{\mu_2}{\mu_1}\partial_y\phi(h,\lambda_{k,t})\tanh\left(\sqrt{\lambda_{k,t} - d^2}(t-h)\right) = 0,$$

where ϕ is defined by (3.2).

We denote by $\hat{g}_{k,t}, k = 1, 2, \dots$, the corresponding eigenfunctions

$$\hat{g}_{k,t}(y) = g(y, \lambda_{k,t}), \quad y \in]0, +\infty[, \tag{A.4}$$

and introduce the weight

$$\frac{1}{r_{k,t}} = \int_0^t \hat{g}_{k,t}^2(y) \frac{dy}{\mu(y)}. \tag{A.5}$$

Let $f(y) \in L^2(0, +\infty; \frac{dy}{\mu(y)})$. Standard Sturm–Liouville theory allows us to write the Parseval identity for f:

$$\int_0^t f^2(y) \frac{dy}{\mu(y)} = \sum_{k=1}^{+\infty} \frac{1}{r_{k,t}} \left(\int_0^t f(y) \hat{g}_{k,t}(y) \frac{dy}{\mu(y)} \right)^2. \tag{A.6}$$

Let

$$\rho_t(\lambda) = \sum_{0 \le \lambda_{k,t} \le \lambda} \frac{1}{r_{k,t}}, \quad \lambda \in]0, +\infty[. \tag{A.7}$$

Then (A.6) takes the form

$$\int_0^t f^2(y) \frac{dy}{\mu(y)} = \int_0^{+\infty} F_t^2(\lambda) d\rho_t(\lambda), \tag{A.8}$$

where

$$F_t(\lambda) = \int_0^t f(y) g(\lambda, y) \frac{dy}{\mu(y)}.$$

We proceed by calculating the limit of $\rho_t(\lambda)$ as $t \to +\infty$.

LEMMA A.1. *Let*

$$\begin{aligned}
\rho(\lambda) &= \begin{cases} \displaystyle\sum_{0 < \lambda_k \le \lambda} \frac{1}{r_k}, & 0 < \lambda \le d^2, \\ \displaystyle\sum_{0 < \lambda_k \le d^2} \frac{1}{r_k} + \int_0^{\sqrt{\lambda - d^2}} r(\nu^2 + d^2) d\nu, & d^2 \le \lambda, \end{cases} \\
r_k &= \frac{2\mu_1 \sqrt{d^2 - \lambda_k}}{\frac{\mu_1}{\mu_2} \phi^2(h, \lambda) + 2\sqrt{d^2 - \lambda_k} \int_0^h \phi^2(h, \lambda_k) dy}, \quad 1 \le k \le m, \\
r(\lambda) &= \frac{1}{\pi} \frac{\mu_2 (\lambda - d^2)}{(\lambda - d^2)\phi^2(h, \lambda) + (\frac{\mu_2}{\mu_1})^2 \partial_y \phi^2(h, \lambda)}, \quad d^2 \le \lambda.
\end{aligned} \tag{A.9}$$

Then the function $\rho_t(\lambda)$ converges to $\rho(\lambda)$ uniformly with respect to λ in any finite interval in $]0, +\infty[$.

Proof. Isolated eigenvalues. The eigenvalues $t \to \lambda_{k,t}$ which belong to the interval $]0, d^2[$ satisfy (A.3). As $t \to +\infty$, the left-hand side of (A.3) converges uniformly in λ; hence the $\lambda_{k,t}$'s smaller than d^2 converge to the roots λ_k of (3.3) for $k = 1, \dots, m_a$. Moreover, by (A.5) we have that

$$\begin{aligned}
\frac{1}{r_{k,t}} &= \frac{1}{\mu_1} \int_0^h \phi^2(y, \lambda_{k,t}) dy \\
&+ \frac{1}{\mu_2 \sqrt{d^2 - \lambda_{k,t}}} \int_0^{(t-h)\sqrt{d^2 - \lambda_{k,t}}} \left(\phi(h, \lambda_{k,t}) \cosh \tau - \frac{\mu_2}{\mu_1} \frac{\partial_y \phi(h, \lambda_{k,t})}{\sqrt{d^2 - \lambda_{k,t}}} \sinh \tau \right)^2 d\tau,
\end{aligned}$$

so that the $r_{k,t}$'s converge to the values r_k's defined in (A.9) as $t \to +\infty$.

The continuous spectrum. For fixed $k = 1, 2, \dots$, any eigenvalue $\lambda_{k,t} > d^2$ converges to d^2 as $t \to +\infty$. Let $h_k(t) = (t-h)\sqrt{\lambda_{k,t} - d^2}$. If $\partial_y \phi(h, d^2) \neq 0$, then (A.3) implies that $h_k(t) \to k\pi$ and so $h_{k+1}(t) - h_k(t) \to \pi$ for $t \to +\infty$. We can write

$$r_{k,t}(\lambda) = \Big(\sqrt{\lambda_{k+1,t} - d^2} - \sqrt{\lambda_{k,t} - d^2}\Big)\hat{r}_{k,t}(\lambda),$$

where

$$\hat{r}_{k,t}(\lambda) = \frac{1}{h_{k+1}(t) - h_k(t)} \times$$

$$\frac{(\lambda - d^2)}{\frac{1}{\mu_1}\frac{(\lambda-d^2)}{t-h}\int_0^h \phi^2(y,\lambda)\, dy + \frac{1}{\mu_2}\frac{1}{\sqrt{\lambda - d^2}(t-h)}\int_0^{(t-h)\sqrt{d^2-\lambda_{k,t}}} \Big(\phi(h,\lambda_{k,t})\cos\tau - \frac{\mu_2}{\mu_1}\frac{\partial_y\phi(h,\lambda_{k,t})}{\sqrt{d^2-\lambda_{k,t}}}\sin\tau\Big)^2 d\tau}.$$

We can easily verify that $\hat{r}_{k,t}(\lambda)$ converges uniformly in compact subsets of $]0, +\infty[$ to $r(\lambda)$ and

$$\sum_{d^2 \le \lambda_{k,t} \le \lambda} (\hat{r}_{k,t}(\lambda_{k,t}) - r(\lambda_{k,t}))\left(\sqrt{\lambda_{k+1,t} - d^2} - \sqrt{\lambda_{k,t} - d^2}\right) \to 0, \tag{A.10}$$

as $t \to +\infty$. Let $\nu_{k,t} = \sqrt{\lambda_{k,t} - d^2}$. Then we have

$$\begin{aligned}\sum_{d^2 \le \lambda_{k,t} \le \lambda} r_{k,t} = &\sum_{d^2 \le \lambda_{k,t} \le \lambda} (\hat{r}_{k,t}(\lambda_{k,t}) - r(\lambda_{k,t}))\left(\sqrt{\lambda_{k+1,t} - d^2} - \sqrt{\lambda_{k,t} - d^2}\right) \\ &+ \sum_{0 \le \nu_{k,t} \le \sqrt{d^2-\lambda}} r(\nu_{k,t}^2 + d^2)(\nu_{k+1,t} - \nu_{k,t}).\end{aligned} \tag{A.11}$$

By (A.10) and the monotony of the sequence $\lambda_{k,t}$, we immediately obtain that

$$\lim_{t \to +\infty} \sum_{d^2 \le \lambda_{k,t} \le \lambda} r_{k,t} = \int_0^{\sqrt{\lambda - d^2}} r(\nu^2 + d^2) d\nu. \tag{A.12}$$

If $\partial_y \phi(h, d^2) = 0$, then since $\partial_y \phi(y, \lambda)$ is an analytic function of λ, by (A.3) we have that either $h_k(t) \to (k - 1/2)\pi$ or $(k + 1/2)\pi$ as $t \to +\infty$. Repeating the previous argument, we arrive at the same conclusion. ∎

A direct consequence of the last lemma gives that for any $\psi(\lambda) \in \mathcal{D}(\mathbb{R}_+)$ we have

$$\begin{aligned}&\lim_{t \to +\infty} \big(d\rho_t(\lambda)/\psi\big) = \big(d\rho(\lambda)/\psi\big) \\ &= \sum_{k=1}^m r_k \psi(\lambda_k) + \frac{1}{\pi}\int_{d^2}^{+\infty} \frac{\mu_2\sqrt{\lambda - d^2}}{(\lambda - d^2)\phi^2(h,\lambda) + (\frac{\mu_2}{\mu_1})^2 \partial_y \phi^2(h,\lambda)} \psi(\lambda) d\lambda,\end{aligned}$$

and so the inversion formula holds for any continuous function with compact support on $\mathbb{R}_+$.

Let us now complete the proof of Lemma 3.1 by proving that it holds for any $f \in L^2(\mathbb{R}_+, \frac{dy}{\mu(y)})$. This will be obtained by a limiting process from the corresponding formula, derived for smooth functions with compact support.

We know that there exists a sequence $\{f_n(y)\}_n$ of functions of class $\mathcal{C}^2(]0,h[\cup]h,+\infty[)$, which satisfy the conditions $f_n(0)=0$, are equal to zero outside the interval $[0,n]$, $\frac{\mu_1}{\mu_2}f_n'(h^+)=f_n'(h^-)$, $f_n(h^+)=f_n(h^-)$, and

$$\lim_{t\to+\infty}\int_0^{+\infty}\{f_n(y)-f(y)\}^2\frac{dy}{\mu(y)}=0. \tag{A.13}$$

Applying the Parseval identity to these functions, we obtain for any large value of t $(t\geq n)$ that

$$\int_0^n f_n^2(y)\frac{dy}{\mu(y)}=\int_0^{+\infty}\check{F}_n^2(\lambda)d\rho_t(\lambda), \tag{A.14}$$

where

$$\check{F}_n(\lambda)=\int_0^n f_n(y)g(y,\lambda)\frac{dy}{\mu(y)}. \tag{A.15}$$

Multiply (3.1) by $f_n(y)$ and integrate by parts twice to get

$$\check{F}_n(\lambda)=-\frac{1}{\lambda}\int_0^t(f_n''(y)-q(y)f_n(y))g(y,\lambda)\frac{dy}{\mu(y)}. \tag{A.16}$$

Therefore for a finite $N>0$ we have

$$\begin{aligned}
&\int_{\lambda>N}\check{F}_n^2(\lambda)d\rho_t(\lambda)\\
&\leq\frac{1}{N^2}\int_{\lambda>N}\left\{\int_0^t(f_n''(y)-q(y)f_n(y))g(y,\lambda)\frac{dy}{\mu(y)}\right\}^2d\rho_t(\lambda)\\
&<\frac{1}{N^2}\int_0^{+\infty}\left\{\int_0^t(f_n''(y)-q(y)f_n(y))g(y,\lambda)\frac{dy}{\mu(y)}\right\}^2d\rho_t(\lambda)\\
&=\frac{1}{N^2}\int_0^n(f_n''(y)-q(y)f_n(y))^2\frac{dy}{\mu(y)}.
\end{aligned}$$

From this bound and (A.14) we obtain

$$\left|\int_0^n f_n^2(y)\frac{dy}{\mu(y)}-\int_0^N\check{F}_n^2(\lambda)d\rho_t(\lambda)\right|\leq\frac{1}{N^2}\int_0^n(f_n''(y)-q(y)f_n(y))^2\frac{dy}{\mu(y)}. \tag{A.17}$$

Lemma A.1 leads to

$$\int_0^N\check{F}_n^2(\lambda)d\rho_t(\lambda)\to\int_0^N\check{F}_n^2(\lambda)d\rho(\lambda) \tag{A.18}$$

as $t\to+\infty$. Passing to the limit in (A.17) we find the inequality

$$\left|\int_0^n f_n^2(y)\frac{dy}{\mu(y)}-\int_0^N\check{F}_n^2(\lambda)d\rho(\lambda)\right|\leq\frac{1}{N^2}\int_0^n(f_n''(y)-q(y)f_n(y))^2\frac{dy}{\mu(y)}. \tag{A.19}$$

Finally, letting $N\to+\infty$ in this inequality, we obtain

$$\int_0^{+\infty}f_n^2(y)\frac{dy}{\mu(y)}=\int_0^{+\infty}\check{F}_n^2(\lambda)d\rho(\lambda). \tag{A.20}$$

Since

$$\int_0^{+\infty} \{f_n(y) - f_m(y)\}^2 \frac{dy}{\mu(y)} \to 0 \tag{A.21}$$

as $n, m \to +\infty$, it follows that

$$\lim_{n,m\to+\infty} \int_0^{+\infty} \{\check{F}_n(\lambda) - \check{F}_m(\lambda)\}^2 d\rho(\lambda) = 0. \tag{A.22}$$

We denote by $\mathcal{H}_\rho$ the space of functions which are square integrable with respect to $\rho(\lambda)$. Now, we will show that $\mathcal{H}_\rho$ is a complete space. Let $\eta_p(\lambda)$ be a sequence of functions in $\mathcal{H}_\rho$ which satisfies

$$\lim_{p,q\to+\infty} \int_0^{+\infty} \{\eta_p(\lambda) - \eta_q(\lambda)\}^2 d\rho(\lambda) = 0. \tag{A.23}$$

The last identity implies

$$\lim_{p,q\to+\infty} \{\eta_p(\lambda_k) - \eta_q(\lambda_k)\}^2 = 0, \quad 1 \le k \le m, \tag{A.24}$$

and

$$\lim_{p,q\to+\infty} \frac{1}{\pi} \int_{d^2}^{+\infty} \frac{\mu_2\sqrt{\lambda - d^2}}{(\lambda - d^2)\phi^2(h,\lambda) + \frac{\mu_1^2}{\mu_2^2}\partial_y\phi^2(h,\lambda)} \{\eta_p(\lambda) - \eta_q(\lambda)\}^2 d\lambda = 0. \tag{A.25}$$

The completeness of $\mathbb{R}_+$ and $L^2(\mathbb{R}_+)$ gives that there exists a unique function $\eta(\lambda)$ such that $\eta_p(\lambda) \to \eta(\lambda)$ in $\mathcal{H}_\rho$ as $p \to +\infty$. Then, the relation (A.22) implies the existence of a limit function $F(\lambda)$ in $\mathcal{H}_\rho$ satisfying the Parseval identity

$$\int_0^{+\infty} f^2(y) \frac{dy}{\mu(y)} = \int_0^{+\infty} F^2(\lambda) d\rho(\lambda). \tag{A.26}$$

We will now show that $F_n(\lambda) = \int_0^n f(y)g(y,\lambda)\frac{dy}{\mu(y)}$ converges to $F(\lambda)$ in $\mathcal{H}_\rho$. Let $g_n(\lambda) = f(y)$ for $0 \le y \le n$ and $g_n(y) \equiv 0$ for $y > n$. Since $g_n(y) \in L^2(0, +\infty; \frac{dy}{\mu(y)})$, there exists $G_n(\lambda)$ in $\mathcal{H}_\rho$ such that

$$\int_0^{+\infty} \{f(y) - g_n(y)\}^2(y) \frac{dy}{\mu(y)} = \int_0^{+\infty} \{F(\lambda) - G_n(\lambda)\}^2 d\rho(\lambda). \tag{A.27}$$

But

$$G_n(\lambda) = \int_0^n f(y)g(y,\lambda)\frac{dy}{\mu(y)}.$$

Hence we obtain

$$\int_0^{+\infty} \{F(\lambda) - F_n(\lambda)\}^2 d\rho(\lambda) = \int_n^{+\infty} f^2(y) \frac{dy}{\mu(y)} \to 0 \tag{A.28}$$

as $n \to +\infty$. The following results hold.

LEMMA A.2. *Let $f(y) \in L^2(0,+\infty;\frac{dy}{\mu(y)})$. There exists a nondecreasing function $\rho(\lambda)$, which does not depend on $f(x)$, and a function $F(\lambda)$ (the generalized Fourier transform of $f(y)$) such that*

$$\int_0^{+\infty} f^2(y)\frac{dy}{\mu(y)} = \int_0^{+\infty} F^2(\lambda)d\rho(\lambda). \tag{A.29}$$

The function $F(\lambda)$ is the limit in quadratic mean relative to $d\rho(\lambda)$ of the sequence of continuous functions $F_n(\lambda) = \int_0^n f(y)g(y,\lambda)\frac{dy}{\mu(y)}$:

$$\lim_{t\to+\infty}\int_0^{+\infty}\{F(\lambda) - F_n(\lambda)\}^2 d\rho(\lambda) = 0. \tag{A.30}$$

LEMMA A.3. *Let $f(y)$ be a continuous function, and suppose that the integral $\int_0^{+\infty} F(\lambda)g(y,\lambda)d\rho(\lambda)$ converges absolutely and uniformly with respect to y in any finite interval. Then*

$$f(y) = \int_0^{+\infty} F(\lambda)g(y,\lambda)d\rho(\lambda), \quad y \in]0,+\infty[. \tag{A.31}$$

Appendix B. The generalized Rouché theorem. For convenience we recall the main results of Ghoberg and Sigal in [2].

B.1. Notations and definitions. Let $\mathcal{H}$ and $\mathcal{H}'$ be two Banach spaces, and let $\mathcal{L}(\mathcal{H},\mathcal{H}')$ be the algebra of all bounded-valued functions acting from $\mathcal{H}$ into $\mathcal{H}'$.

Let λ_0 be a fixed complex value in $\mathbb{C}$. We denote by $\mathcal{A}(\lambda)$ an operator-valued function acting from $D_\varepsilon(\lambda_0)$ into $\mathcal{L}(\mathcal{H},\mathcal{H}')$, where $D_\varepsilon(\lambda_0)$ is a disc of center λ_0 and radius $\varepsilon > 0$.

λ_0 is called a *characteristic* value of $\mathcal{A}(\lambda)$ if

(i) $\mathcal{A}(\lambda)$ is holomorphic in some neighborhood of λ_0, except possibly at this point itself;

(ii) there exists a vector-valued function $\phi(\lambda)$: $D_\varepsilon(\lambda_0) \to \mathcal{H}$ holomorphic at λ_0 and that verifies $\phi(\lambda_0) \neq 0$, such that $\mathcal{A}(\lambda)\phi(\lambda)$ is a holomorphic at λ_0 and vanishes at this point. $\phi(\lambda)$ is called a *root* function of $\mathcal{A}(\lambda)$ associated with λ_0, and the vector $\phi_0 = \phi(\lambda_0)$ is called an eigenvector. The closure of the linear set of eigenvectors corresponding to λ_0 is denoted by $\mathrm{Ker}\mathcal{A}(\lambda_0)$.

Suppose that λ_0 is a characteristic value of the function $\mathcal{A}(\lambda)$ and $\phi(\lambda)$ is a root function satisfying (ii). Then there exists a number $m(\phi) \geq 1$ and a vector-valued function $\psi(\lambda) : D_\varepsilon(\lambda_0) \to \mathcal{H}$ holomorphic such that

$$\mathcal{A}(\lambda)\phi(\lambda) = (\lambda - \lambda_0)^{m(\phi)}\psi(\lambda),$$
$$\psi(\lambda_0) \neq 0.$$

The number $m(\phi)$ is called the *multiplicity* of the root function $\phi(\lambda)$. Let ϕ_0 be an eigenvector corresponding to λ_0 and let

$$\mathcal{R}(\phi_0) = \{m(\phi); \phi(\lambda) \text{ is a root function such } \phi(\lambda_0) = \phi_0\}.$$

Then by rank of ϕ_0 we mean $\mathrm{rank}(\phi_0) = \max \mathcal{R}(\phi_0)$. Suppose that $n = \dim\mathrm{Ker}\mathcal{A}(\lambda_0) < +\infty$ and that the ranks of all vectors in $\mathrm{Ker}\mathcal{A}(\lambda_0)$ are finite. A system of eigenvectors ϕ_0^j, $j = 1,\dots,n$, is called a *canonical system of eigenvectors* of $\mathcal{A}(\lambda)$ associated to λ_0

if the ranks possess the following property: $\operatorname{rank}(\phi_0^j)$ is the maximum of the ranks of all eigenvectors in some direct complement in $\dim\operatorname{Ker}\mathcal{A}(\lambda_0)$ of the linear span of the vectors $\phi_0^1, \dots, \phi_0^{j-1}$. Let $r_j = \operatorname{rank}(\phi_0^j)$. Then $(r_j)_j$ determines the function $\mathcal{A}(\lambda)$ uniquely. We call

$$N(\mathcal{A}(\lambda_0)) = \sum_{j=1}^{n} r_j$$

the *null multiplicity of the characteristic* value λ_0 of $\mathcal{A}(\lambda)$. If λ_0 is not a characteristic value of $\mathcal{A}(\lambda)$, we put $N(\mathcal{A}(\lambda_0)) = 0$.

Suppose that $\mathcal{A}^{-1}(\lambda)$ exists and is holomorphic in some neighborhood of λ_0, except possibly at this point itself. Then the number

$$M(\mathcal{A}(\lambda_0)) = N(\mathcal{A}(\lambda_0)) - N(\mathcal{A}^{-1}(\lambda_0))$$

is called the *multiplicity* of the characteristic value λ_0 of $\mathcal{A}(\lambda)$. Suppose that λ_1 is a pole of the operator-valued function. The Laurent expansion of $\mathcal{A}(\lambda)$ in λ_1 is given by

$$\mathcal{A}(\lambda) = \sum_{j \geq -s} (\lambda - \lambda_1)^j A_j.$$

If in the last expression the operators A_{-j}, $j = 1, \dots, s$, are finite-dimensional, then $\mathcal{A}(\lambda)$ is called *finitely meromorphic* at λ_1.

The operator-valued function $\mathcal{A}(\lambda)$ is said to be of Fredholm type at the point λ_1 if the operator A_0 in the last expansion is a Fredholm operator. If $\mathcal{A}(\lambda)$ is holomorphic at the point λ_0 and the operator $\mathcal{A}(\lambda_0)$ is invertible, then λ_0 is called a *regular point* of $\mathcal{A}(\lambda)$.

B.2. Main results. The point λ_0 is called a *normal point* of $\mathcal{A}(\lambda)$ if there exists a constant $0 < \varepsilon_0 \leq \varepsilon$ such that $\mathcal{A}(\lambda)$ is finitely meromorphic and of Fredholm type at λ_0 and all the points of $D_{\varepsilon_0}(\lambda_0) - \{\lambda_0\}$ are regular for $\mathcal{A}(\lambda)$, where $D_{\varepsilon_0}(\lambda_0)$ is a disc of center λ_0 and radius $\varepsilon_0 > 0$.

LEMMA B.1. *Every normal point λ_0 of $\mathcal{A}(\lambda)$ is a normal point of $\mathcal{A}^{-1}(\lambda)$.*

Let $\partial D_{\varepsilon_0}$ be the contour bounding the domain $D_{\varepsilon_0}(\lambda_0)$. An operator-valued function $\mathcal{A}(\lambda)$ which is finitely meromorphic and of Fredholm type in $D_{\varepsilon_0}(\lambda_0)$ and continuous at $\partial D_{\varepsilon_0}$ is called *normal with respect to* $\partial D_{\varepsilon_0}$ if the operator $\mathcal{A}(\lambda)$ is invertible in $\overline{D_{\varepsilon_0}(\lambda_0)}$, except for a finite number of points of $D_{\varepsilon_0}(\lambda_0)$ which are normal points of $\mathcal{A}(\lambda)$. Now, if $\mathcal{A}(\lambda)$ is normal with respect to the contour $\partial D_{\varepsilon_0}$ and λ_i, $i = 1, \dots, \sigma$, are all its characteristic values and poles lying in $D_{\varepsilon_0}(\lambda_0)$, we put

$$\mathcal{M}(\mathcal{A}(\lambda); \partial D_{\varepsilon_0}) = \sum_{i=1}^{\sigma} M(\mathcal{A}(\lambda_i)).$$

THEOREM B.2. *Suppose that the operator-valued $\mathcal{A}(\lambda)$ is normal with respect to $\partial D_{\varepsilon_0}$; then we have*

$$\mathcal{M}(\mathcal{A}(\lambda); \partial D_{\varepsilon_0}) = \frac{1}{2i\pi} \operatorname{tr} \int_{\partial D_{\varepsilon_0}} \mathcal{A}^{-1}(\lambda) \frac{d}{d\lambda} \mathcal{A}(\lambda) d\lambda.$$

By tr we mean the trace of operator which is the sum of all its nonzero characteristic values, see [2, p. 609] for an exact statement.

The operator generalization of the Rouché theorem is stated below.

THEOREM B.3. *Let $\mathcal{A}(\lambda)$ be an operator-valued function which is normal with respect to $\partial D_{\varepsilon_0}$. If an operator-valued function $S(\lambda)$ which is finitely meromorphic in $D_{\varepsilon_0}(\lambda_0)$ and continuous at $\partial D_{\varepsilon_0}$ satisfies the condition*

$$|\mathcal{A}^{-1}(\lambda)S(\lambda)|_{\mathcal{L}(\mathcal{H},\mathcal{H})} < 1, \quad \lambda \in \partial D_{\varepsilon_0},$$

then $\mathcal{A}(\lambda) + S(\lambda)$ is also normal with respect to $\partial D_{\varepsilon_0}$, and

$$\mathcal{M}(\mathcal{A}(\lambda); \partial D_{\varepsilon_0}) = \mathcal{M}(\mathcal{A}(\lambda) + S(\lambda); \partial D_{\varepsilon_0}).$$

The generalization of the Steinberg theorem is given by the following.

THEOREM B.4. *Suppose that $\mathcal{A}(\lambda)$ is an operator-valued function which is finitely meromorphic and of Fredholm type in the domain $D_{\varepsilon_0}(\lambda_0)$. If the operator $\mathcal{A}(\lambda)$ is invertible at one point of D_{ε_0}, then $\mathcal{A}(\lambda)$ has a bounded inverse for all $\lambda \in D_{\varepsilon_0}$, except possibly for certain isolated points.*

Finally, the following result is central.

THEOREM B.5. *Suppose that $\mathcal{A}(\lambda)$ is an operator-valued function which is normal with respect to $\partial D_{\varepsilon_0}$. Let $f(\lambda)$ be a scalar function which is analytic in $D_{\varepsilon_0}(\lambda_0)$ and continuous in $\overline{D_{\varepsilon_0}(\lambda_0)}$. Then*

$$\frac{1}{2i\pi}\ \mathrm{tr}\ \int_{\partial D_{\varepsilon_0}} f(\lambda)\mathcal{A}^{-1}(\lambda)\frac{d}{d\lambda}\mathcal{A}(\lambda)d\lambda = \sum_{j=1}^{\sigma} M(\mathcal{A}(\lambda_j))f(\lambda_j),$$

where λ_j, $j = 1, \ldots, \sigma$, are all the points in $D_{\varepsilon_0}(\lambda_0)$ which are either poles or characteristic values of $\mathcal{A}(\lambda)$.

REFERENCES

[1] H. AMMARI, N. BÉREUX, AND J. C. NÉDÉLEC, *Resonances for Maxwell's equations in a periodic structure*, Japan J. Indust. Appl. Math., 17 (2000), pp. 149–198.

[2] I. T. S. GOHBERG AND E. I. SIGAL, *Operator extension of the logarithmic residue theorem and Rouché's theorem*, Mat. Sb. (N.S.), 84 (1971), pp. 607–642.

[3] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, R.E. Krieger Publishing, Malabar, FL, 1984.

[4] R. E. COLLIN, *Field Theory of Guided Wave,* 2nd ed., IEEE Press, New York, 1991.

[5] R. R. GADYL'SHIN AND A. M. IL'IN, *Asymptotic behavior of the eigenvalues of the Dirichlet problem in a domain with a narrow slit*, Sb. Math., 189 (1998), pp. 503–526.

[6] K. C. GUPTA, I. J. BAHL, P. BHARTIA, AND R. GARG, *Microstrip Lines and Slotlines*, 2nd ed., Artech House, London, 1996.

[7] J. R. JAMES AND P. S. HALL, EDS., *Handbook of Microstrip Antennas*, Peter Peregrinus, London, 1988.

[8] R. MAGNANINI AND F. SANTOSA, *Wave propagation in a 2D optical wave guide*, SIAM J. Appl. Math., 61 (2000), pp. 1237–1252.

[9] N. I. MUSKHELISHVILI, *Singular Integral Equations*, Noordhoff International, Leyden, 1977.

[10] S. OZAWA, *Singular variation of domains and eigenvalues of the Laplacian*, Duke Math. J., 48 (1981), pp. 767–778.

[11] D. M. POZAR AND D. H. SCHAUBERT, EDS., *Microstrip Antennas: The Analysis and Design of Microstrip Antennas and Arrays*, John Wiley & Sons, New York, 1995.

[12] V. P. SHESTOPALOV AND Y. V. SHESTOPALOV, *Spectral Theory and Excitation of Open Structures*, IEEE Press, London, 1996.

[13] C. WILCOX, *Spectral analysis of the Pekeris operator in the theory of acoustic wave propagation*, Arch. Ration. Mech. Anal., 60 (1975), pp. 259–300.

[14] C. WILCOX, *Sound Propagation in Stratified Fluids*, Appl. Math. Sci., 50, Springer-Verlag, New York, 1984.

[15] K. L. WONG, *Design of Nonplanar Microstrip Antennas and Transmission Lines*, Wiley Ser. Microwave Optical Engrg., Wiley-Interscience, New York, 1999.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 637–667

A HYPERBOLIC SYSTEM OF EQUATIONS OF BLOOD FLOW IN AN ARTERIAL NETWORK*

WEIHUA RUAN†, M. E. CLARK‡, MEIDE ZHAO‡, AND ANTHONY CURCIO‡

Abstract. We study a coupled system of the Navier–Stokes equation and the equation of conservation of mass in a network. The system models the blood circulation in arterial networks. A special feature of the system is that the equations are coupled through boundary conditions at joints of the network. We use a fixed point method to prove the existence and uniqueness of the classic solution to the initial-boundary value problem and discuss the continuity of dependence of the solution and its derivatives on initial, boundary, and forcing functions and their derivatives. We develop a numerical scheme that generates discretized solutions, and we also prove the convergence of the scheme.

Key words. arterial network, blood flow, hyperbolic partial differential equations, initial-boundary value problems, numerical scheme

AMS subject classifications. 76Z05, 35L45, 35Q80, 65M06, 92C35

DOI. 10.1137/S0036139902415294

1. Introduction.

In this paper, we study a system of one-spatial-dimensional first-order quasi-linear hyperbolic partial differential equations defined on networks. By network, we mean a finite collection of smooth curves joining a finite number of vertices, and with a direction assigned to each curve (see Figure 1.1). The mathematical system arises from a long-term study of fluid dynamical models that simulate blood flow in arterial networks (cf. [1, 2, 9, 11, 13, 17, 18, 19, 20, 21, 22, 23, 24, 25] and

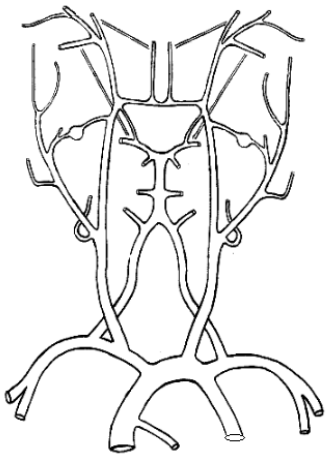


FIG. 1.1. *A schematic diagram of an arterial network.*

*Received by the editors September 27, 2002; accepted for publication (in revised form) June 19, 2003; published electronically December 31, 2003.
http://www.siam.org/journals/siap/64-2/41529.html

†Department of Mathematics, Computer Science, and Statistics, Purdue University Calumet, Hammond, IN 46323-2094 (ruanw@nwi.calumet.purdue.edu). The research of this author was partially supported by VasSol, Inc.

‡VasSol, Inc., 833 W. Jackson Blvd., Chicago, IL 60607 (m-clark4@uiuc.edu, meide@vassolinc.com, apc@vassolinc.com).

references therein). Recently, these models have been used in technologies for medical diagnostics [3, 6, 7, 8]. In particular, a technology called CANVAS (computer-assisted noninvasive vascular analysis and simulation) has been developed to help stroke patients. CANVAS uses data from magnetic resonance imaging to determine volumetric flow within vessels in the patient's brain [26]. The vessel flows are used to determine the boundary conditions of the model [8]. This approach is based on a model formulated by Clark and Kufahl [9, 13]. The technology has displayed its capability in helping doctors predict outcomes of major medical procedures. It is the extensive applications of these models that motivate their mathematical study. Of particular importance are whether the mathematical system is well posed (solution exists, is unique, and is stable), and whether the solutions generated by the computer algorithm approximate the true solutions.

In this paper, we study a generalization of a model given by [20, 23, 24], prove the existence and uniqueness of the solution, prove the continuous dependence of the solution on the initial, boundary, and forcing functions, and develop a numerical scheme that approximates the solution.

To explain our system, let us first describe the original model of [20, 23, 24]. Suppose that an arterial network consists of n vessels. We parameterize each vessel with a spatial variable $x \in (0,1)$ in accordance with the assigned direction of blood flow through the vessel. In the vessel, the flow of blood is governed by the one-spatial-dimensional equation of conservation of mass and an approximation of the Navier–Stokes momentum equation:

$$\begin{aligned} &\frac{\partial Q_i}{\partial x} + \frac{\partial A_i}{\partial t} = 0, \\ &\frac{\partial Q_i}{\partial t} + \alpha_i \frac{\partial}{\partial x}\left(\frac{Q_i^2}{A_i}\right) = -\frac{A_i}{\rho_i}\frac{\partial P_i}{\partial x} - \frac{\kappa_i Q_i}{A_i}, \end{aligned} \qquad x \in (0,1)\,,\ t > 0, \tag{1.1}$$

where Q_i is the flow rate, P_i is the pressure, A_i is the cross-sectional area of the vessel, and α_i, ρ_i, κ_i are positive constants. The above equations are valid if we assume that the radial component of the fluid velocity is far less than the axial component and that the axial velocity profile is proportional to $1-(r/R)^{\gamma}$ (R being the radius of the vessel) for some $\gamma \neq 0$. The latter assumption is satisfied by the Poiseuille flow profile considered in [9] and the plug flow profile considered in [17, 18]. (The reader is referred to [18, 19, 25] for a derivation of the above equations from three-dimensional mass balance and Navier–Stokes equations.)

The initial conditions are given by

$$P_i\,(0,x) = P_i^I\,(x)\,, \quad Q_i\,(0,x) = Q_i^I\,(x)\,, \quad i = 1,\ldots,n.$$

At each end of the vessel, depending on whether it is a source, an internal junction, or a terminal, a boundary condition is imposed. At a source end, either the pressure

$$P_i\,(0,t) = P_i^B\,(t) \tag{1.2}$$

or the flow

$$Q_i\,(0,t) = Q_i^B\,(t) \tag{1.3}$$

is specified. Various source ends may have different types of boundary conditions. At an internal junction, suppose $j_1,\ldots,j_\nu$ are the incoming vessels to and $j_{\nu+1},\ldots,j_\mu$

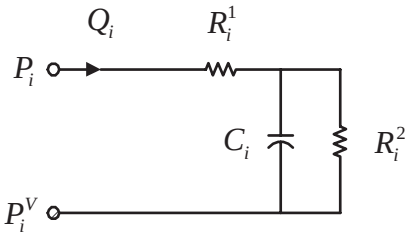


FIG. 1.2. *Windkessel model for a terminal boundary condition.*

are the outgoing vessels from the junction. We have mass and pressure continuities at junctions given by

$$\begin{aligned} &\sum_{l=1}^{\nu} Q_{j_l}(1,t) = \sum_{l'=\nu+1}^{\mu} Q_{j_{l'}}(0,t), \\ &P_{j_l}(1,t) = P_{j_{l'}}(0,t), \quad 1 \le l \le \nu,\ \nu+1 \le l' \le \mu. \end{aligned} \tag{1.4}$$

At a terminal end, we may either specify the pressure,

$$P_i(1,t) = P_i^B(t), \tag{1.5}$$

or the flow,

$$Q_i(1,t) = Q_i^B(t), \tag{1.6}$$

or use the windkessel model of peripheral beds [13, 20, 24] as the boundary condition,

$$\frac{\partial P_i}{\partial t} - \eta_i \frac{\partial Q_i}{\partial t} + \delta_i P_i - \varepsilon_i Q_i = W_i^B(t), \quad x = 1, \tag{1.7}$$

where η_i, δ_i, and ε_i are positive constants and W_i^B is a continuous function. This equation models the peripheral beds by a circuit that consists of a resistance R_i^1 in series with the parallel combination of a resistance R_i^2 and a capacitor C_i; see Figure 1.2. The resulting equation is

$$C_i \frac{\partial}{\partial t}\left(P_i - P_i^V\right) - R_i^1 C_i \frac{\partial Q_i}{\partial t} + \frac{P_i - P_i^V}{R_i^2} - \frac{R_i^1}{R_i^2} Q_i = 0,$$

where P_i^V is the venous pressure. This can be rewritten into (1.7). Again, boundary conditions for different terminals need not be the same.

Finally, the cross-sectional area A_i of the ith vessel is a function of x and P_i. A particular example used in [9, 13] is

$$A_i(x, P_i) = A_i^0(x) + \beta \ln \frac{P_i}{P_i^0},$$

where β is a positive constant and A_i^0 is a positive function which represents the cross-sectional area at certain constant pressure P_i^0.

In this paper, we study a more general system, which consists of the equations

$$\begin{aligned} &\frac{\partial P_i}{\partial t} + a_i \frac{\partial Q_i}{\partial x} = f_i, \\ &\frac{\partial Q_i}{\partial t} + b_i \frac{\partial P_i}{\partial x} + 2c_i \frac{\partial Q_i}{\partial x} = g_i, \end{aligned} \qquad x \in (0,1),\ t > 0, \tag{1.8}$$

and the initial and boundary conditions described above. For convenience, we also use the vector form

$$(U_i)_t + B_i (U_i)_x = F_i, \tag{1.9}$$

where $U_i = (P_i, Q_i)$, $F_i = (f_i, g_i)$, and

$$B_i = \begin{pmatrix} 0 & a_i \\ b_i & 2c_i \end{pmatrix}.$$

Equation (1.1) is a special case of this system, where

$$a_i = \frac{1}{(A_i)_{P_i}}, \quad b_i = \frac{A_i}{\rho_i} - \frac{\alpha_i Q_i^2 (A_i)_{P_i}}{A_i^2}, \quad c_i = \frac{\alpha_i Q_i}{A_i}, \quad f_i = 0, \quad g_i = \frac{Q_i^2 (A_i)_x}{A_i^2} - \frac{\kappa_i Q_i}{A_i}.$$

We do not assume any particular form for these functions, only that they are general differentiable functions of (x, t, P_i, Q_i). A basic assumption is $a_i > 0$. Other assumptions will follow.

This problem is interesting not only in fluid mechanics but also in mathematics. Navier–Stokes equations and conservation laws have been studied for over a century. Initial-boundary value problems of such equations in a one-dimensional network have been studied for decades [1, 9, 18, 20, 24, 25]. However, so far, most analyses on the models have been performed numerically rather than mathematically. To give a mathematical analysis is challenging. Unlike the problem of fluid flow in a rigid tube network, the distensibility of vessels greatly increases the complexity of the problem (cf. [19]). For example, as is well known, a first-order quasi-linear system of hyperbolic equations on a finite one-dimensional spatial interval need not have a solution. Even if it has a solution for an interval of time, the solution may not exist for all time. A particular example is given by Čanić [4]. In a network, it is important to know whether the coupling at junctions poses problems to solvability. The effect of the Windkessel boundary condition (1.7) on the solvability also needs to be examined.

In this paper, we study the existence and uniqueness of the local solution, develop a numerical scheme to approximate the solution, and prove the convergence of the scheme. This is the first step towards an analysis of the system. We consider only classic solutions (i.e., solutions whose derivatives are continuous and satisfy the differential equations) in this paper. The problems of existence and uniqueness of the weak solution are interesting but also more difficult. In particular, it is more difficult to establish the convergence of the scheme to the weak solution; this will be the subject of our future study. (See [4] for an example of a numerical scheme that gives a good approximation of the weak solution to a scaler initial-value problem.)

This paper is divided into two parts. The first part consists of sections 2 and 3, which deal with the problem of solvability using a fixed point approach. Substituting a pair of functions (p_i, q_i) for (P_i, Q_i) in the coefficients a_i, b_i, c_i and forcing functions f_i, g_i, the system becomes linear. That is, all the functions a_i, etc., are independent of unknowns. If the linear system has a unique solution, then one can establish a mapping from (p_i, q_i) to the linear problem solution (P_i, Q_i). If one also shows that this mapping has a unique fixed point, then the fixed point is necessarily the unique solution of the quasilinear system. Hence, we shall first give a condition for the linear system to have a unique solution, then examine under what conditions the mapping has a unique fixed point. We investigate the first aspect of the problem in section 2 and the latter in section 3. We also prove a result on the continuity of dependence

of solutions on the initial, boundary, and forcing functions for linear and quasi-linear systems. Thus, we complete the analysis of the well-posedness of the problem. In the second part of the paper, which consists of section 4 only, we give a numerical scheme that approximates the solution and prove its convergence. Our scheme is a set of finite-difference equations based on the normal form of the differential equations. Although these approaches are standard in the analysis of quasi-linear equations, the network feature of the system and the peculiarities of the boundary conditions make the problem more complicated. In the final section, we give a short discussion.

2. The linear system. In this section, we analyze (1.8) as a linear system with a_i, b_i, c_i, f_i, and g_i independent of P_i and Q_i. We give conditions for the system to have a unique global solution. The conditions are most naturally given in terms of the eigenvalues of the matrix B_i, which have the form

$$\lambda_i^R = c_i + u_i, \quad \lambda_i^L = c_i - u_i,$$

where

$$u_i = \sqrt{c_i^2 + a_i b_i}.$$

These eigenvalues are real if

$$c_i^2 + a_i b_i > 0, \quad x \in (0,1), \quad t > 0, \quad i = 1, \ldots, n. \tag{2.1}$$

In this case,

$$\lambda_i^L(x,t) < \lambda_i^R(x,t) \tag{2.2}$$

and the system is hyperbolic. Under this condition, we show that the linear system has a unique solution if

$$\lambda_i^L(x,t) < 0, \quad \lambda_i^R(x,t) > 0, \quad x = 0, 1, \ i = 1, \ldots, n.$$

This is clearly equivalent to

$$a_i b_i > 0, \quad t \geq 0, \quad i = 1, \ldots, n, \tag{2.3}$$

at $x = 0, 1$ only. It need not hold for $x \in (0,1)$.

THEOREM 2.1. *Assume that the functions a_i, b_i, c_i, f_i, and g_i are independent of (P_i, Q_i). Suppose that these functions and the initial and boundary functions P_i^I, Q_i^I, P_i^B, Q_i^B, and W_i^B all have bounded first-order derivatives. Suppose also that $a_i > 0$ and that the conditions* (2.1) *and* (2.3) *hold. Then, for any $T > 0$ there is a unique solution in a bounded subset of the space $C\left([0,1] \times [0,T], \mathbb{R}^{2n}\right)$ to the linear system* (1.8) *with the initial and boundary conditions given in section* 1.

Proof. We first show that the system has a unique solution for $0 < t < \delta$ for some $\delta > 0$. The proof is based on the method of characteristics and a fixed point principle. For systems defined on only one branch, this is a standard approach. In our case, special care is needed to handle the junction condition (1.4) and the Windkessel boundary condition (1.7).

Consider the ith branch of the network. From any point (τ, ξ) on the left, right, and lower boundary of the rectangle $D =: [0,1] \times [0,T]$, we construct the left-going

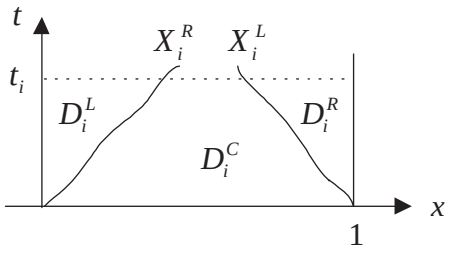


FIG. 2.1. *Three parts of* D_i.

and right-going characteristic curves $x = x_i^L(t;\xi,\tau)$ and $x = x_i^R(t;\xi,\tau)$ by

$$\frac{dx_i^L}{dt} = \lambda_i^L\big(x_i^L, t\big), \quad x_i^L(\tau) = \xi,$$
$$\frac{dx_i^R}{dt} = \lambda_i^R\big(x_i^R, t\big), \quad x_i^R(\tau) = \xi,$$

respectively, where λ_i^L and λ_i^R are the two eigenvalues of the matrix B_i. By the uniqueness of solutions of these differential equations, a left-going characteristic curve cannot intersect with another left-going characteristic curve, and the same is true for right-going characteristic curves. Let X_i^L and X_i^R be the right-most left-going and left-most right-going characteristic curves:

$$x = x_i^L(t;1,0) \quad \text{and} \quad x = x_i^R(t;0,0)$$

starting from the lower boundary of D, respectively. It can be shown from (2.2) that the two curves can have at most one intersection. Let t_i be the value of t at the intersection. If the two curves do not intersect in D, we simply define $t_i = T$. By condition (2.3), X_i^L cannot reach the right vertical line $x = 1$, and X_i^R cannot reach the vertical line $x = 0$ at any $t > 0$. Thus, the rectangle $D_i =: [0,1] \times [0,t_i]$ can be divided into three parts:

$$D_i = D_i^L \cup D_i^C \cup D_i^R,$$

where D_i^L is between the vertical line $x = 0$ and the characteristic curve X_i^R, D_i^C is between the two characteristic curves, and D_i^R is between X_i^L and $x = 1$ (see Figure 2.1). We show that there is a $\delta_i \le t_i$ such that the solution (P_i, Q_i) for the ith branch exists in the restriction of D_i to the strip $\{0 \le t \le \delta_i\}$.

We first observe that the initial conditions alone determine the solution completely in the central region D_i^C. This follows from the theory of first-order linear hyperbolic systems and the fact that from any point $(x,t) \in D_i^C$ the two characteristic curves, followed backwards, must land on the horizontal line $t = 0$. This imperative is a consequence of (2.2). To extend the solution to other parts of D_i, we make a change of unknowns and derive a set of integral equations. Note that $l_i^R =: (-\lambda_i^L, a_i)$ and $l_i^L =: (-\lambda_i^R, a_i)$ are the left eigenvectors of B_i corresponding to λ_i^R and λ_i^L, respectively. Introduce new unknowns

$$r_i = l_i^R U_i \equiv -\lambda_i^L P_i + a_i Q_i, \quad s_i = l_i^L U_i \equiv -\lambda_i^R P_i + a_i Q_i. \tag{2.4}$$

The system (1.8) can be written in terms of r_i and s_i by multiplying the left eigenvectors by (1.9) and substituting in

$$P_i = \frac{1}{2u_i}\left(r_i - s_i\right), \quad Q_i = \frac{1}{2u_i a_i}\left(\lambda_i^R r_i - \lambda_i^L s_i\right). \tag{2.5}$$

This results in the equations

$$\partial_i^R r_i = F_i^R\left(x,t,r_i,s_i\right), \quad \partial_i^L s_i = F_i^L\left(x,t,r_i,s_i\right), \tag{2.6}$$

where

$$\partial_i^R = \frac{\partial}{\partial t} + \lambda_i^R \frac{\partial}{\partial x}, \quad \partial_i^L = \frac{\partial}{\partial t} + \lambda_i^L \frac{\partial}{\partial x}, \tag{2.7}$$

and

$$F_i^R\left(x,t,r_i,s_i\right) = l_i^R F_i + \left(\partial_i^R l_i^R\right) U_i, \quad F_i^L\left(x,t,r_i,s_i\right) = l_i^L F_i + \left(\partial_i^L l_i^L\right) U_i. \tag{2.8}$$

(A differential operator acting on a vector means that it acts on each component of the vector.)

Let $(x,t) \in D_i$. We integrate the first equation of (2.6) along the right-going characteristic curve $x^R\left(t;\xi,\tau\right)$, which passes through (x,t) and reaches the left or lower boundary of D_i at (ξ,τ). It can be shown that for $(x,t) \in D_i^C \cup D_i^R$, $\tau = 0$, and for $(x,t) \in D_i^L$, $\xi = 0$. In the former case, we obtain

$$r_i\left(x,t\right) = r_i^I\left(\xi\right) + \int_0^t F_i^R\left(x_i^R\left(t';\xi,0\right),t',r_i,s_i\right)dt'. \tag{2.9}$$

In the latter case, we have

$$r_i\left(x,t\right) = r_i\left(0,\tau\right) + \int_\tau^t F_i^R\left(x_i^R\left(t';0,\tau\right),t',r_i,s_i\right)dt'. \tag{2.10}$$

Similarly, by integrating the second equation of (2.6) along the left-going characteristic curve $x_i^L\left(t;\xi,\tau\right)$ that passes through both (x,t) and (ξ,τ) (which is on either the right or lower boundary of D_i), the equations are

$$s_i\left(x,t\right) = s_i^I\left(\xi\right) + \int_0^t F_i^L\left(x_i^L\left(t';\xi,0\right),t',r_i,s_i\right)dt' \quad \text{if } (x,t) \in D_i^L \cup D_i^C \tag{2.11}$$

and

$$s_i\left(x,t\right) = s_i\left(1,\tau\right) + \int_\tau^t F_i^L\left(x_i^L\left(t';1,\tau\right),t',r_i,s_i\right)dt' \quad \text{if } (x,t) \in D_i^R. \tag{2.12}$$

These are the integral equations we need.

For any $\delta_i \le t_i$, we use D_{i,δ_i}^L, D_{i,δ_i}^C, and D_{i,δ_i}^R to denote the restrictions of D_i^L, D_i^C, and D_i^R, respectively, to the strip $\{0 \le t \le \delta_i\}$. We first extend the solution to a left region D_{i,δ_i}^L where δ_i is to be determined. For this, we need the boundary condition on the left end of the branch. The left end is either a source or a junction.

For a source with the pressure boundary condition (1.2), we define $\hat{s}_i = s_i/\varepsilon$, where $\varepsilon < 1$ is any constant. Using the first equation of (2.5) in the integral equations

(2.10) and (2.11),

(2.13)
$$\begin{pmatrix} r_i(x,t) \\ \hat{s}_i(x,t) \end{pmatrix} = \begin{pmatrix} 2u_i(0,\tau) P_i^B(\tau) + \varepsilon \hat{s}_i(0,\tau) + \int_\tau^t F_i^R\left(x_i^R(t';0,\tau), t', r_i, \varepsilon \hat{s}_i\right) dt' \\ \frac{1}{\varepsilon} s_i^I(\xi) + \frac{1}{\varepsilon} \int_0^t F_i^L\left(x_i^L(t';\xi,0), t', r_i, \varepsilon \hat{s}_i\right) dt' \end{pmatrix}.$$

This is a fixed point equation for $(r_i, \hat{s}_i)$ if we define the right-hand side as a mapping of an operator K on $(r_i, \hat{s}_i)$ in a bounded subset of $C(D^L_{i,\delta_i} \cup D^C_{i,\delta_i}, \mathbb{R}^2)$. In a standard approach, it can be shown that K is a contraction mapping if δ_i is sufficiently small. Hence, the fixed point exists and is unique. Therefore, the solution (r_i, s_i) can be uniquely extended to $D^L_{i,\delta_i} \cup D^C_{i,\delta_i}$.

For a source with the flow boundary condition (1.3), we define $\hat{s}_i = s_i/\varepsilon$, where $\varepsilon > 0$ and is so small that

$$\varepsilon \left| \frac{\lambda_i^L(0,\tau)}{\lambda_i^R(0,\tau)} \right| < 1, \quad \tau \in (0, t_i).$$

The fixed point equation is then

(2.14)
$$\begin{pmatrix} r_i(x,t) \\ \hat{s}_i(x,t) \end{pmatrix} = \begin{pmatrix} \frac{2a_i u_i(0,\tau)}{\lambda_i^R(0,\tau)} Q_i^B(\tau) + \frac{\lambda_i^L(0,\tau)}{\lambda_i^R(0,\tau)} \varepsilon \hat{s}_i(0,\tau) + \int_\tau^t F_i^R\left(x_i^R(t';0,\tau), t', r_i, \varepsilon \hat{s}_i\right) dt' \\ \frac{1}{\varepsilon} s_i^I(\xi) + \frac{1}{\varepsilon} \int_0^t F_i^L\left(x_i^L(t';\xi,0), t', r_i, \varepsilon \hat{s}_i\right) dt' \end{pmatrix}.$$

By a similar argument, the solution can again be uniquely extended.

If the left end of the branch is a junction, we shall extend the solution simultaneously on all the branches that are connected to the same junction. Thus, also extend the solution to D^R_{i,δ_i} on the branches incoming to the junction. Let $j_1, \ldots, j_\nu$ be the incoming and $j_{\nu+1}, \ldots, j_\mu$ the outgoing branches at the junction. Equations (1.4) and (2.5) give rise to a $2\mu \times \mu$ homogeneous system of linear equations for $r_i(1,\tau)$, $s_i(1,\tau)$, $i = j_1, \ldots, j_\nu$, and $r_i(0,\tau)$, $s_i(0,\tau)$, $i = j_{\nu+1}, \ldots, j_\mu$:

$$\frac{1}{u_1(1,\tau)} (r_1(1,\tau) - s_1(1,\tau)) - \frac{1}{u_i(1,\tau)} (r_i(1,\tau) - s_i(1,\tau)) = 0, \quad i = j_2, \ldots, j_\nu,$$

$$\frac{1}{u_1(1,\tau)} (r_1(1,\tau) - s_1(1,\tau)) - \frac{1}{u_i(0,\tau)} (r_i(0,\tau) - s_i(0,\tau)) = 0, \quad i = j_{\nu+1}, , \ldots, j_\mu,$$

$$\sum_{l=1}^{\nu} \frac{1}{u_{j_l} a_{j_l}} \left(\lambda_{j_l}^R r_{j_l} - \lambda_{j_l}^L s_{j_l}\right)(1,\tau) - \sum_{l'=\nu+1}^{\mu} \frac{1}{u_{j_{l'}} a_{j_{l'}}} \left(\lambda_{j_{l'}}^R r_{j_{l'}} - \lambda_{j_{l'}}^L s_{j_{l'}}\right)(0,\tau) = 0.$$

This system can be solved for $s_{j_1}(1,\tau), \ldots, s_{j_\nu}(1,\tau)$, $r_{j_{\nu+1}}(0,\tau), \ldots, r_{j_\mu}(0,\tau)$ because the coefficient matrix

$$\begin{pmatrix} -\frac{1}{u_{j_1}(1,\tau)} & \frac{1}{u_{j_2}(1,\tau)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{u_{j_1}(1,\tau)} & 0 & \cdots & -\frac{1}{u_{j_\mu}(0,\tau)} \\ -\frac{\lambda_{j_1}^L(1,\tau)}{u_{j_1} a_{j_1}(1,\tau)} & -\frac{\lambda_{j_2}^L(1,\tau)}{u_{j_2} a_{j_2}(1,\tau)} & \cdots & -\frac{\lambda_{j_\mu}^R(0,\tau)}{u_{j_\mu} a_{j_\mu}(0,\tau)} \end{pmatrix}$$

has the determinant

$$\frac{(-1)^{\nu+1}}{\prod_{l=1}^{\nu} u_{j_l}(1,\tau)\prod_{l'=\nu+1}^{\mu} u_{j_{l'}}(0,\tau)}\left(-\sum_{l=1}^{\nu}\frac{\lambda_{j_l}^{L}(1,\tau)}{a_{j_l}(1,\tau)}+\sum_{l'=\nu+1}^{\mu}\frac{\lambda_{j_{l'}}^{R}(0,\tau)}{a_{j_{l'}}(0,\tau)}\right).$$

Since $\lambda_i^L < 0 < \lambda_i^R$ at the junction, the determinant is not zero. Hence, we can express $s_{j_1}(1,\tau),\dots,s_{j_\nu}(1,\tau)$, $r_{j_{\nu+1}}(0,\tau),\dots,r_{j_\mu}(0,\tau)$ in terms of other unknowns as

$$s_i(1,\tau)=\sum_{l=1}^{\nu} m_{j_l}^{i}(\tau)\, r_{j_l}(1,\tau)+\sum_{l'=\nu+1}^{\mu} m_{j_{l'}}^{i}(\tau)\, s_{j_{l'}}(0,\tau),\quad i=j_1,\dots,j_\nu,$$

$$r_i(0,\tau)=\sum_{l=1}^{\nu} n_{j_l}^{i}(\tau)\, r_{j_l}(1,\tau)+\sum_{l'=\nu+1}^{\mu} n_{j_{l'}}^{i}(\tau)\, s_{j_{l'}}(0,\tau),\quad i=j_{\nu+1},\dots,j_\mu,$$

for some functions m_j^i, n_j^i. Choose an $\varepsilon>0$ such that

$$\varepsilon\max\left\{\sum_{l=1}^{\mu}\left|m_{j_l}^{i}(\tau)\right|,\sum_{l=1}^{\mu}\left|n_{j_l}^{i}(\tau)\right|\right\}<1,\quad i=j_1,\dots,j_\mu,\ \tau\in[0,t_i],$$

and introduce

$$\hat r_{j_l}=\frac{r_{j_l}}{\varepsilon},\quad \hat s_{j_{l'}}=\frac{s_{j_{l'}}}{\varepsilon},\quad l=1,\dots,\nu,\ l'=\nu+1,\dots,\mu.$$

Then, in view of (2.9)–(2.12), the integral equations for the 2μ unknowns $\hat r_{j_l}$, s_{j_l}, $r_{j_{l'}}$, $\hat s_{j_{l'}}$, $l=1,\dots,\nu$, $l'=\nu+1,\dots,\mu$, constitute a fixed point equation, $w=Kw$, where

$$w=\left(\hat r_{j_1},\dots,\hat r_{j_\nu},s_{j_1},\dots,s_{j_\nu},r_{j_{\nu+1}},\dots,r_{j_\mu},\hat s_{j_{\nu+1}},\dots,\hat s_{j_\mu}\right) \tag{2.15}$$

and

$$\begin{aligned}Kw=\Bigg(&\frac{1}{\varepsilon}r_{j_1}^{I}(\xi_{j_1})+\frac{1}{\varepsilon}\int_0^t F_{j_1}^{R}\left(x_{j_1}^{R},t',\varepsilon\hat r_{j_1},s_{j_1}\right)dt',\dots,\\ &\varepsilon\left(\sum_{k=1}^{\nu} m_{j_k}^{1}\hat r_{j_k}(1,\tau)+\sum_{k'=\nu+1}^{\mu} m_{j_{k'}}^{1}\hat s_{j_{k'}}(0,\tau)\right)\\ &+\int_\tau^t F_{j_1}^{L}\left(x_{j_1}^{L},t',\varepsilon\hat r_{j_1},s_{j_1}\right)dt',\dots,\\ &\varepsilon\left(\sum_{k=1}^{\nu} n_{j_k}^{1}\hat r_{j_k}(1,\tau)+\sum_{k'=\nu+1}^{\mu} n_{j_{k'}}^{1}\hat s_{j_{k'}}(1,\tau)\right)\\ &+\int_\tau^t F_{j_{\nu+1}}^{R}\left(x_{j_{\nu+1}}^{R},t',r_{j_{\nu+1}},\varepsilon\hat s_{j_{\nu+1}}\right)dt',\dots,\\ &\frac{1}{\varepsilon}s_{j_{\nu+1}}^{I}\left(\xi_{j_{\nu+1}}\right)+\frac{1}{\varepsilon}\int_0^t F_{j_{\nu+1}}^{L}\left(x_{j_{\nu+1}}^{L},t',r_{j_{\nu+1}},\varepsilon\hat s_{j_{\nu+1}}\right)dt',\dots\Bigg).\end{aligned} \tag{2.16}$$

It can be shown by a standard argument that K is a contraction mapping in the space

$$X_j=:\prod_{l=1}^{\nu} C\left(D_{j_l,\delta_j}^{C}\cup D_{j_l,\delta_j}^{R},\mathbb{R}^2\right)\times\prod_{l=\nu+1}^{\mu} C\left(D_{j_l,\delta_j}^{L}\cup D_{j_l,\delta_j}^{L},\mathbb{R}^2\right)$$

if δ_j is sufficiently small. Hence, it has a unique fixed point in X_j. This extends the solution (r_i, s_i) for the neighboring branches of the junction.

We now extend the solution (r_i, s_i) to a right region D^R_{i,δ_i}. This has already been considered if the right end is a junction. Thus, only terminal ends need to be discussed. For the boundary condition of either (1.5) or (1.6) type, the argument is similar to the above discussion about source ends. We only sketch the steps in these two cases. The boundary condition of (1.7) type, however, requires more effort.

If condition (1.5) is assumed, then, by (2.5),

$$s_i(1,t) = r_i(1,t) - 2u_i P_i^B(t).$$

Let $\hat{r}_i = r_i/\varepsilon$ with $0 < \varepsilon < 1$. Then, the fixed point equation for $(\hat{r}_i, s_i)$ has the form

(2.17)
$$\begin{pmatrix} \hat{r}_i(x,t) \\ s_i(x,t) \end{pmatrix} = \begin{pmatrix} \frac{1}{\varepsilon} r_i^I(\xi) + \frac{1}{\varepsilon}\int_0^t F_i^R\left(t', x_i^R(t';\xi,0), \varepsilon\hat{r}_i, s_i\right) dt' \\ \varepsilon\hat{r}_i(1,\tau) - 2u_i(1,\tau) P_i^B(\tau) + \int_\tau^t F_i^L\left(t', x_i^L(t';1,\tau), \varepsilon\hat{r}_i, s_i\right) dt' \end{pmatrix}.$$

As before, the mapping defined by the right-hand side is contractive if δ_i is small enough. Hence, the solution is uniquely extended into D^R_{i,δ_i}. If condition (1.6) is assumed, we find again from (2.5) that

$$\lambda_i^L s_i(1,t) = \lambda_i^R r_i(1,t) - 2u_i(1,t) a_i(1,t) Q_i^B(t).$$

Since $\lambda_i^L(1,t) < 0$, the equation can be uniquely solved for s_i. Choose $\varepsilon > 0$ sufficiently small such that

$$\varepsilon\left|\frac{\lambda_i^R(1,\tau)}{\lambda_i^L(1,\tau)}\right| < 1 \quad \text{for } \tau \in [0,t_i],$$

and let $\hat{r}_i = r_i/\varepsilon$. The fixed point equation for $(\hat{r}_i, s_i)$ has the form

(2.18)
$$\begin{pmatrix} \hat{r}_i(x,t) \\ s_i(x,t) \end{pmatrix} = \begin{pmatrix} \frac{1}{\varepsilon} r_i^I(\xi) + \frac{1}{\varepsilon}\int_0^t F_i^R\left(x_i^R(t';\xi,0), t', \varepsilon\hat{r}_i, s_i\right) dt' \\ \frac{\lambda_i^R(1,\tau)}{\lambda_i^L(1,\tau)}\varepsilon\hat{r}_i(1,\tau) - \frac{2a_i u_i(1,\tau)}{\lambda_i^L(1,\tau)} Q_i^B(\tau) + \int_\tau^t F_i^L\left(x_i^L(t';1,\tau), t', \varepsilon\hat{r}_i, s_i\right) dt' \end{pmatrix}.$$

Again, the mapping is contractive in a bounded subset of $C(D^C_{i,\delta_i} \cup D^R_{i,\delta_i}, \mathbb{R}^2)$ if δ_i is sufficiently small. Thus, the solution is uniquely extended to D^R_{i,δ_i}.

In the case where the boundary condition (1.7) is assumed, we integrate it with respect to t to obtain

$$(P_i - \eta_i Q_i)(1,t) = \left(P_i^I - \eta_i Q_i^I\right)(1) + \int_0^t \left(W_i^B(t') - \delta_i P_i(1,t') + \varepsilon_i Q_i(1,t')\right) dt'.$$

Substituting (2.5) into this equation, we can write

$$\begin{aligned} &m_i(t) r_i(1,t) - n_i(t) s_i(1,t) \\ &\quad = m_i(0) r_i^I(1) - n_i(0) s_i^I(1) + \int_0^t H_i(t', r_i(1,t'), s_i(1,t'))\, dt', \end{aligned}$$

where

$$m_i(t) = \frac{a_i(1,t) - \eta_i \lambda_i^R(1,t)}{2a_i u_i(1,t)}, \quad n_i(t) = \frac{-a_i(1,t) + \eta_i \lambda_i^L(1,t)}{2a_i u_i(1,t)}$$

and

$$H_i(t,r,s) = W_i^B(t) + \frac{\varepsilon_i \lambda_i^R(1,t) - \delta_i a_i(1,t)}{2a_i u_i(1,t)} r - \frac{\varepsilon_i \lambda_i^L(1,t) - \delta_i a_i(1,t)}{2a_i u_i(1,t)} s.$$

Since $a_i > 0$, $u_i > 0$, $\eta_i > 0$, and $\lambda_i^L(1,t) < 0$, it follows that $n_i(t) < 0$. Hence, there exists $\varepsilon > 0$ such that

$$\varepsilon \left| \frac{m_i(\tau)}{n_i(\tau)} \right| < 1 \quad \text{for } \tau \in [0, t_i].$$

Let $\hat{r}_i = r_i/\varepsilon$. The integral equations for $\hat{r}_i$ and s_i then have the form

(2.19)

$$\begin{aligned}
\hat{r}_i(x,t) &= \frac{1}{\varepsilon} r_i^I(\xi) + \frac{1}{\varepsilon} \int_0^t F_i^R\left(x_i^R(t';\xi,0), t', \varepsilon \hat{r}_i, s_i\right) dt', \\
s_i(x,t) &= \varepsilon \frac{m_i(\tau)}{n_i(\tau)} \hat{r}_i(1,\tau) - \frac{1}{n_i(\tau)} \left(M_i + \int_0^t H_i\left(t', \varepsilon \hat{r}_i(1,t'), s_i(1,t')\right) dt' \right) \\
&\qquad + \int_\tau^t F_i^L\left(x_i^L(t';1,\tau), t', \varepsilon \hat{r}_i, s_i\right) dt',
\end{aligned}$$

where $M_i = m_i(0) r_i^I(1) - n_i(0) s_i^I(1)$ is a constant. The extension of the solution to D_{i,δ_i}^R is thus guaranteed.

Finally, if we let δ be the minimum of all δ_i occurring above, we see that $\delta > 0$, and the solution exists and is unique in $(x,t) \in D_\delta =: [0,1] \times [0,\delta]$. Observe that δ depends only on the bounds of the system functions (e.g., a_i), the initial and boundary functions (e.g., P_i^I), and their first-order derivatives in $D = [0,1] \times [0,T]$. Hence, it is independent of t, and we can extend the solution successively in the time intervals $[0,\delta]$, $[\delta, 2\delta]$, etc. In this way, the solution is obtained in D in a finite number of steps. □

It can be seen from the above proof that the linear system need not have a solution if condition (2.3) fails at any end point of a branch. In the quasilinear case, since a_i and b_i depend on the unknowns P_i and Q_i, this condition may fail at some time. Therefore the solution does not generally exist for all time.

We next derive an estimate of the deviation of the solution in term of the deviations of the initial, boundary, and forcing functions. This estimate is needed in the next section. For any vector function $v = (v_1, \ldots, v_k)$ defined in $C(X; \mathbb{R}^k)$ we use $|v|_{C(X)}$ to denote the norm $\max_i\{|v_i|_{C(X)}\}$, where X represents a closed subset of either $\mathbb{R}$ or $\mathbb{R}^2$ and $|v_i|_{C(X)}$ is the maximum norm.

LEMMA 2.2. *Let $U = (P,Q)$ and $\tilde{U} = (\tilde{P}, \tilde{Q})$ be two solutions of the linear problem* (1.9) *with different initial, boundary, and forcing functions. Suppose the conditions of Theorem* 2.1 *hold for both solutions. Then there exists a constant $M > 0$, independent of initial, boundary, and forcing functions, such that*

$$\begin{aligned}
&\left| U - \tilde{U} \right|_{C(D_\delta)} \\
&\le M \left(\left| P^I - \tilde{P}^I \right|_{C[0,1]} + \left| Q^I - \tilde{Q}^I \right|_{C[0,1]} + \left| P^B - \tilde{P}^B \right|_{C[0,\delta]} + \left| Q^B - \tilde{Q}^B \right|_{C[0,\delta]} \right. \\
&\qquad \left. + \delta \left| f - \tilde{f} \right|_{C(D_\delta)} + \delta \left| g - \tilde{g} \right|_{C(D_\delta)} + \delta \left| W - \tilde{W} \right|_{C[0,\delta]} \right)
\end{aligned} \tag{2.20}$$

Proof. We need prove (2.20) only for a $\delta \leq \min_i \{\delta_i\}$, where δ_i represents the constants occurring in the proof of Theorem 2.1. This is because for larger δ we can divide the interval $[0, \delta]$ into subintervals, each having a length less than $\min_i \{\delta_i\}$, and apply (2.20) in each subinterval. We can then take the maximum on each side of the inequalities to derive the inequality in $[0, \delta]$. In what follows, D_δ^C, D_δ^L, and D_δ^R are the restrictions of D_i^C, D_i^L, and D_i^R, respectively, to the strip $\{0 \leq t \leq \delta\}$.

By linearity, $U - \tilde{U}$ is the solution of the system with the initial, boundary, and forcing functions $P_i^I - \tilde{P}_i^I$, $Q_i^I - \tilde{Q}_i^I$, $P_i^B - \tilde{P}_i^B$, $Q_i^B - \tilde{Q}_i^B$, $W_i^B - \tilde{W}_i^B$, $f_i - \tilde{f}_i$, and $g_i - \tilde{g}_i$. Let r_i, $\hat{r}_i$, s_i, $\hat{s}_i$ be defined as in the proof of Theorem 2.1, corresponding to $U - \tilde{U}$. We show that these quantities have upper bounds in the form of the right-hand side of (2.20) in D_δ^C, D_δ^L, and D_δ^R.

In D_δ^C, (2.9) and (2.11) hold. Notice that the functions F_i^R and F_i^L are linear in r_i and s_i. Hence, there exists a constant M (we will use M generically for any constant bounds that are independent of solutions) such that

$$R_i^C(t) + S_i^C(t) \leq \left|r_i^I\right|_{C[0,1]} + \left|s_i^I\right|_{C[0,1]} + M \int_0^t \left(R_i^C(t') + S_i^C(t') + T_i^C(t')\right) dt',$$

where

$$R_i^C(t) = \sup_{\{x:(x,t)\in D_\delta^C\}} |r_i(x,t)|, \quad S_i^C(t) = \sup_{\{x:(x,t)\in D_\delta^C\}} |s_i(x,t)|, \tag{2.21}$$

and

$$T_i^C(t) = \sup_{\{x:(x,t)\in D_\delta^C\}} \left(\left|f_i(x,t) - \tilde{f}_i(x,t)\right| + |g_i(x,t) - \tilde{g}_i(x,t)|\right). \tag{2.22}$$

Hence, by Gronwall's inequality (see, e.g., [16, p. 327]),

$$R_i^C(t) + S_i^C(t) \leq M\left(\left|r_i^I\right|_{C[0,1]} + \left|s_i^I\right|_{C[0,1]} + \delta \sup_{t\in(0,\delta)} T_i^C(t)\right)$$

for $t \in [0, \delta]$. This proves that R_i^C and S_i^C have upper bounds in the form of the right-hand side of (2.20).

In D_δ^L, if the left end is a source, we use either (2.13) or (2.14) according to the type of the boundary condition. The resulting inequality has the form

$$\begin{aligned} &R_i^L(t) + \hat{S}_i^L(t) \\ &\quad \leq \sigma \hat{S}_i^L(t) + M\left(\left|s_i^I\right|_{C[0,1]} + \left|\xi_i^B\right|_{C[0,\delta]} + \int_0^t \left(R_i^L(\tau) + \hat{S}_i^L(\tau) + T_i^L(\tau)\right) d\tau\right), \end{aligned}$$

where ξ_i^B is either P_i^B or Q_i^B, depending on the boundary condition, and R_i^L, $\hat{S}_i^L$, and T_i^L are defined in the same way as in (2.21)–(2.22), with D_δ^C substituted by $D_\delta^L \cup D_\delta^C$, and $\sigma > 0$ is a positive constant such that $\sigma = \varepsilon$ if the boundary condition is (1.2) and

$$\sigma = \varepsilon \sup_{t\in(0,\delta)} \left|\frac{\lambda_i^L(0,t)}{\lambda_i^R(0,t)}\right| < 1$$

if the boundary condition is (1.3). Replacing M by $(1-\sigma)M$, we can write

$$R_i^L(t) + \hat{S}_i^L(t) \leq M\left(\left|s_i^I\right|_{C[0,1]} + \left|\xi_i^B\right|_{C[0,\delta]} + \int_0^t \left(R_i^L(\tau) + \hat{S}_i^L(\tau) + T_i^L(\tau)\right) d\tau\right).$$

Hence, by Gronwall's inequality,

$$R_i^L(t) + \hat{S}_i^L(t) \le M\left(\left|s_i^I\right|_{C[0,1]} + \left|\xi_i^B\right|_{C[0,\delta]} + \delta \max_{t\in(0,\delta)} T_i^L(t)\right).$$

This proves that both $R_i^L(t)$ and $S_i^L(t)$ have upper bounds in the form of the right-hand side of (2.20).

If the left end is a junction, the solutions on the branches $j_1, \ldots, j_\mu$ connecting to the junction constitute fixed points of the operator K, which is defined in (2.16). Let

$$W(t) = \sum_{l=1}^{\nu}\left(\hat{R}_{j_l}^R(t) + S_{j_l}^R(t)\right) + \sum_{l'=\nu+1}^{\mu}\left(R_{j_{l'}}^L(t) + \hat{S}_{j_{l'}}^L(t)\right),$$

where $\hat{R}_i^R$ and S_i^R are defined as in (2.21) with D_δ^C substituted by $D_\delta^C \cup D_\delta^R$. Then, from $w = Kw$, we can deduce

$$\begin{aligned} W(t) \le{}& \sigma\left(\sum_{l=1}^{\nu}\hat{R}_{j_l}^R(t) + \sum_{l'=\nu+1}^{\mu}\hat{S}_{j_{l'}}^L(t)\right) \\ &+ M\left(\sum_{l=1}^{\nu}\left|r_{j_l}^I\right|_{C[0,1]} + \sum_{l'=\nu}^{\mu}\left|s_{j_{l'}}^I\right|_{C[0,1]} + \int_0^t (W(\tau) + T(\tau))\,d\tau\right), \end{aligned}$$

where

$$T(\tau) = \sum_{l=1}^{\nu} T_{j_l}^R(\tau) + \sum_{l'=\nu+1}^{\mu} T_{j_{l'}}^L(\tau)$$

and $T_i^R(t)$ is defined as in (2.22) with D_δ^C substituted by $D_\delta^C \cup D_\delta^R$. Replacing M by $(1-\sigma)M$, we obtain

$$W(t) \le M\left(\sum_{l=1}^{\nu}\left|r_{j_l}^I\right|_{C[0,1]} + \sum_{l'=\nu}^{\mu}\left|s_{j_{l'}}^I\right|_{C[0,1]} + \int_0^t (W(\tau) + T(\tau))\,d\tau\right).$$

Hence, by Gronwall's inequality,

$$W(t) \le M\left(\sum_{l=1}^{\nu}\left|r_{j_l}^I\right|_{C[0,1]} + \sum_{l'=\nu}^{\mu}\left|s_{j_{l'}}^I\right|_{C[0,1]} + \delta \max_{t\in(0,\delta)} T(t)\right).$$

This leads to an upper bound in the form of the right-hand side of (2.20) for $R_i^R(t)$, $S_i^R(t)$, $i = j_1, \ldots, j_\nu$, and $R_i^L(t)$, $S_i^L(t)$, $i = j_{\nu+1}, \ldots, j_\mu$.

The only remaining case is when the right end of the branch is a terminal. The fixed point equation to be used is either (2.17), (2.18), or (2.19), depending on the type of the boundary condition. In the former two cases, the treatment is similar to that for sources. Hence, we consider only the third case. From (2.19), we obtain

$$\begin{aligned} &\hat{R}_i^R(t) + S_i^R(t) \\ &\quad \le \sigma\hat{R}_i^R(t) + M\left(\left|r_i^I\right|_{C[0,1]} + \int_0^t \left(\hat{R}_i^R(t') + S_i^R(t') + \left|W_i^B(t')\right| + T_i^R(t')\right)dt'\right), \end{aligned}$$

where

$$\sigma = \varepsilon \max_{t\in[0,\delta]} \left| \frac{m_i(t)}{n_i(t)} \right| < 1.$$

Hence, by Gronwall's inequality,

$$\hat{R}_i^R(t) + S_i^R(t) \le M \left(\left| r_i^I \right|_{C[0,1]} + \delta \max_{t\in(0,\delta)} T_i^R(t) + \delta \max_{t\in(0,\delta)} \left| W_i^B(t) \right| \right),$$

which gives the desired upper bounds of R_i^R and S_i^R.

We have, thus, obtained an upper bound in the form of the right-hand side of (2.20) for the quantities $|r_i - \tilde{r}_i|_{C(D_\delta)}$ and $|s_i - s_i|_{C(D_\delta)}$. The conclusion of the lemma follows now from (2.5). □

3. The quasilinear system. In this section, we study the quasilinear system where the coefficients a_i, b_i, c_i, f_i, and g_i depend on both (x,t) and (P_i, Q_i). Under certain conditions, we show that the system has a unique local solution. We then present a theorem on the continuity of dependence of the solution on initial, boundary, and forcing functions.

The basic idea in the proof of the existence of a solution is to construct an iterative sequence. Substituting any vector function (p_i, q_i) for (P_i, Q_i) in a_i, etc., the system becomes linear. Thus, we can use Theorem 2.1 to get a solution (P_i, Q_i). This defines a mapping S from $u =: (p_i, q_i)$ to $U =: (P_i, Q_i)$, and the solution for the quasilinear system is a fixed point of S. If there is a subset of a Banach space that is invariant under S, then we can construct a sequence

$$u_{k+1} = Su_k, \quad k = 0, 1, \ldots .$$

In the case where the limit exists and is unique, it gives rise to fixed point of S. This is our approach in this section.

In this approach, conditions (2.1) and (2.3) are repeatedly used. One might want to impose them for all the values of the variables. This would give the existence and uniqueness for the global solution, as in the case of the linear system. However, such a requirement is so restrictive that even the original system (1.1) cannot meet it. Therefore, we will impose them only for $t = 0$, and obtain the local solution for the quasilinear system.

THEOREM 3.1. *Assume that the initial and boundary functions P_i^I, Q_i^I, P_i^B, Q_i^B, W_i^B and the system functions a_i, b_i, c_i, f_i, g_i all have continuous first-order derivatives with respect to each variable. Suppose that $a_i > 0$ for all the values of its arguments, and that conditions* (2.1) *and* (2.3) *hold at $t = 0$. Suppose also that the initial functions P_i^I, Q_i^I satisfy any relevant boundary conditions at $t = 0$. Then, for some $\delta > 0$, there is a unique solution for $0 \le t < \delta$ to the quasilinear system* (1.8) *with the initial and boundary conditions described in section* 1.

Proof. We first consider the simpler case where $U^I =: (P^I, Q^I) = 0$. Let $v = \{v_i\}$, $v_i = (p_i, q_i)$ be a family of vector functions (not necessarily constituting a solution) that satisfy the initial and boundary conditions. Substitute v for U in the functions a_i, b_i, c_i, f_i, and g_i. Then, the system becomes linear, and we can invoke Theorem 2.1 to obtain a solution U to the linear system. This defines a mapping $S : v \mapsto U$. A solution of the quasilinear system is then a fixed point of S. We will choose a subset X_{δ,M_0} of a Banach space such that (1) $SX_{\delta,M_0} \subset X_{\delta,M_0}$ and (2) S is contracting in X_{δ,M_0}. For any scalar or vector function $f \in C^k(D_\delta)$, let $|f|_{k,\delta}$ denote the maximum norm of all

the kth-order derivatives of f in D_δ. (If f is a vector function, $|f|_{k,\delta} = \max_i\{|f_i|_{k,\delta}\}$.) Let $C_B(D_\delta, \mathbb{R}^{2n})$ denote the subset of the vector-valued functions in $C(D_\delta, \mathbb{R}^{2n})$ that satisfy the initial and boundary conditions. We seek X_{δ,M_0} in the form

$$X_{\delta,M_0} = \left\{ v \in C_B\left(D_\delta, \mathbb{R}^{2n}\right) : \ |v|_{0,\delta} \le M_0, |v|_{1,\delta} \le M_1 \right\}, \tag{3.1}$$

where M_0 is an arbitrary positive constant and M_1 is a constant to be determined. Note that, by the vanishing initial condition, for any M_1, $|U|_{1,\delta} \le M_1$ implies $|U|_{0,\delta} \le M_1\delta$. Hence, for any M_0, we can ensure $|U|_{0,\delta} \le M_0$ by reducing δ. It remains, therefore, only to show that for M_1 sufficiently large and δ sufficiently small, $|v|_{1,\delta} \le M_1$ implies $|Sv|_{1,\delta} \le M_1$. Throughout this proof, we use M to represent any positive constant that may depend on M_1 but is otherwise independent of v and δ, and use $\tilde{M}$ for any constant that is independent of M_1, v, and δ. The values of M or $\tilde{M}$ in different occurrences need not be equal.

Let $U = Sv$, and let r_i and s_i be defined by (2.4). On each branch, we show that

$$\max\left\{\left|(r_i)_x\right|, \left|(s_i)_x\right|\right\} \le M_1 \tag{3.2}$$

and

$$\max\left\{\left|(r_i)_t\right|, \left|(s_i)_t\right|\right\} \le M_1 \tag{3.3}$$

in D_δ^C, D_δ^L, and D_δ^R if M_1 is large and δ is small. (Recall that D_δ^C, etc., are the intersections $D_i^C \cap D_\delta$, etc., respectively.) In fact, only (3.2) needs to be shown. To see this, first observe that the vanishing initial condition and the compatibility of the initial and boundary conditions gives

$$\max_i \left\{ \left|P_i^B\right|_{C[0,\delta]}, \left|Q_i^B\right|_{C[0,\delta]} \right\} \le M\delta.$$

Hence, we obtain from Lemma 2.2 with $\tilde{U} = 0$ that

$$|U|_{0,\delta} \le M\delta. \tag{3.4}$$

From (2.6) and (2.8), there are constants $\tilde{M}$ and M such that

$$\begin{aligned} \left|\partial_i^R r_i\right| &\le \left|l_i^R F_i\right| + \left|\partial_i^R l_i^R\right| |U_i| \le \tilde{M} + M\delta, \\ \left|\partial_i^L s_i\right| &\le \left|l_i^L F_i\right| + \left|\partial_i^L l_i^L\right| |U_i| \le \tilde{M} + M\delta \end{aligned} \tag{3.5}$$

for each $i = 1, \ldots, n$. Hence, (3.3) follows from (3.2), (3.5), and the definition of ∂_i^L and ∂_i^R in (2.7). We also note that (2.5) and (3.5) imply

$$\left|\partial_i^R U_i\right|_{0,\delta} \le \tilde{M} + M\delta, \quad \left|\partial_i^R U_i\right|_{0,\delta} \le \tilde{M} + M\delta \tag{3.6}$$

for all i. This will be used later.

We first consider the middle region D_δ^C, where the solution (r_i, s_i) satisfies the integral equations (2.9) and (2.11) with $r_i^I = s_i^I = 0$. Differentiating the equations with respect to x, we have

$$\begin{aligned} (r_i)_x &= \left(l_i^R\right)_x U_i(x,t) + \int_0^t \left[\left(l_i^R F_i\right)_x + \left(\partial_i^R l_i^R\right)(U_i)_x - \left(l_i^R\right)_x \left(\partial_i^R U_i\right)\right] \left(x_i^R\right)_x dt, \\ (s_i)_x &= \left(l_i^L\right)_x U_i(x,t) + \int_0^t \left[\left(l_i^L F_i\right)_x + \left(\partial_i^L l_i^L\right)(U_i)_x - \left(l_i^L\right)_x \left(\partial_i^L U_i\right)\right] \left(x_i^L\right)_x dt. \end{aligned} \tag{3.7}$$

Here, we used an identity from [10, p. 469]:

$$\begin{aligned}\frac{d}{d\xi}\int_a^b f\left(x\left(t\right),t\right)Dg\left(x\left(t\right),t\right)dt&\\ =f\left(x\left(b\right),b\right)g_x\left(x\left(b\right),b\right)x_\xi\left(b\right)-f\left(x\left(a\right),a\right)g_x\left(x\left(a\right),a\right)x_\xi\left(a\right)&\\ +\int_a^b\left[f_x\left(x\left(t\right),t\right)Dg\left(x\left(t\right),t\right)-Df\left(x\left(t\right),t\right)g_x\left(x\left(t\right),t\right)\right]x_\xi\left(t\right)dt,&\end{aligned}\tag{3.8}$$

where $x\left(t\right)$ is a function such that $x\left(b\right)=\xi$ and $D=\frac{\partial}{\partial t}+x'\left(t\right)\frac{\partial}{\partial x}$. (Here and below, $x'\left(t\right)=dx/dt$.) Let

$$R_i^C\left(t\right)=\sup_{\left\{x:(x,t)\in D_\delta^C\right\}}\left\{\left|\left(r_i\right)_x\left(x,t\right)\right|\right\},\quad S_i^C\left(t\right)=\sup_{\left\{x:(x,t)\in D_\delta^C\right\}}\left\{\left|\left(s_i\right)_x\left(x,t\right)\right|\right\}.\tag{3.9}$$

From (3.4), (3.6), and (3.7), we derive

$$R_i^C\left(t\right)+S_i^C\left(t\right)\le M\delta+M\int_0^t\left(1+R_i^C\left(t'\right)+S_i^C\left(t'\right)\right)dt'$$

for $t\in[0,\delta]$. Hence, Gronwall's inequality gives

$$\left|\left(r_i\right)_x\right|\le M\delta e^{M\delta},\quad \left|\left(s_i\right)_x\right|\le M\delta e^{M\delta}$$

in D_δ^C. This proves (3.2) in D_δ^C if M_1 is sufficiently large and δ is sufficiently small.

We next consider the left triangular region D_δ^L in the case where the branch is connected to a source. Let $\hat{s}_i=s_i/\varepsilon$ for any $\varepsilon>0$. Then, the pair $(r_i,\hat{s}_i)$ satisfies the fixed point equations of either (2.13) or (2.14), depending on the type of the boundary condition. Differentiating the equations with respect to x and using a slightly modified version of (3.8), we have

$$\begin{aligned}\left(r_i\right)_x&=\left(\zeta_i-l_i^RF_i-\left(\partial_i^Rl_i^R\right)U_i\right)\left(0,\tau\right)\tau_x+\left(l_i^R\right)_xU_i\left(x,t\right)-\left(l_i^R\right)_xU_i\left(x_i^R\right)_x\left(0,\tau\right)\\ &\quad+\int_\tau^t\left[\left(l_i^RF_i\right)_x+\left(\partial_i^Rl_i^R\right)\left(U_i\right)_x-\left(l_i^R\right)_x\left(\partial_i^RU_i\right)\right]\left(x_i^R\right)_xdt,\\ \left(\hat{s}_i\right)_x&=\frac{1}{\varepsilon}\left(l_i^L\right)_xU_i\left(t,x\right)+\frac{1}{\varepsilon}\int_0^t\left[\left(l_i^LF_i\right)_x+\left(\partial_i^Ll_i^L\right)\left(U_i\right)_x-\left(l_i^L\right)_x\left(\partial_i^LU_i\right)\right]\left(x_i^L\right)_xdt,\end{aligned}\tag{3.10}$$

where

$$\zeta_i=2\left(u_iP_i^B\right)_t+\varepsilon\left(\hat{s}_i\right)_t$$

if the boundary condition is given by (1.2), and

$$\zeta_i=2\left(\frac{a_iu_i}{\lambda_i^R}Q_i^B\right)_t+\varepsilon\left(\frac{\lambda_i^L}{\lambda_i^R}\right)_t\hat{s}_i+\varepsilon\left(\frac{\lambda_i^L}{\lambda_i^R}\right)\left(\hat{s}_i\right)_t$$

if the boundary condition is given by (1.3). (Modification of (3.8) is caused by the lower limit of the integral in the first equation of (3.10), which also depends on x.) This equation is valid for any ε. Thus, we may choose ε so small that

$$\sigma=:\varepsilon\left|\lambda_i^L\tau_x\left(0,t\right)\right|\max\left\{1,\left|\left(\frac{\lambda_i^L\left(0,t\right)}{\lambda_i^R\left(0,t\right)}\right)\right|\right\}<1,\quad t\in\left[0,\delta\right].$$

To proceed further, we need an estimate of $|\tau_x(0,t)|$. Observe that $\tau(x)$ satisfies the equation

$$x_i^R(\tau; x, t) = 0,$$

where $x_i^R(\tau; x, t)$ is the solution of the initial value problem

$$\frac{dx_i^R}{ds} = \lambda_i^R\left(x_i^R, s\right), \quad x_i^R(t; x, t) = x.$$

By differentiation,

$$\lambda_i^R(0, \tau(x))\,\tau_x + \left.\frac{\partial x_i^R}{\partial x}\right|_{(\tau(x);x,t)} = 0. \tag{3.11}$$

Let $w_i = \partial x_i^R/\partial x$. Then w_i is the solution of the linear equation

$$\frac{dw_i}{ds} = \left(\lambda_i^R\right)_x\left(x_i^R(s; x, t), s\right) w_i, \quad w_i(t) = 1.$$

Solving the equation,

$$w_i(s) = \exp\left(\int_t^s \left(\lambda_i^R\right)_x\left(x_i^R(s'; x, t), s'\right) ds'\right).$$

Returning to (3.11), we find

$$\tau_x = \frac{-1}{\lambda_i^R(0, \tau(x))} \exp\left(\int_t^{\tau(x)} \left(\lambda_i^R\right)_x\left(x_i^R(s'; x, t), s'\right) ds'\right).$$

Observe that $0 < \tau(x) < t \le \delta$ and the integrand is bounded. Hence,

$$|\tau_x| \le \tilde{M} e^{M\delta}. \tag{3.12}$$

This is the estimate we need. By this estimate, for any M_1, we can choose δ small enough such that the constants σ and ε are independent of M_1. Let $R_i^L(t)$ and $\hat{S}_i^L(t)$ be defined as in (3.9) except that s_i is substituted by $\hat{s}_i$ and D_δ^C is substituted by $D_\delta^L \cup D_\delta^C$. We derive from (3.10) and the identity

$$(\hat{s}_i)_t = \partial_i^L \hat{s}_i - \lambda_i^L (\hat{s}_i)_x$$

that

$$R_i^L(t) + \hat{S}_i^L(t) \le \sigma \hat{S}_i^L(t) + \tilde{M} + M\delta + M\int_0^t \left(1 + R_i^L(t') + \hat{S}_i^L(t')\right) dt'.$$

Replacing M and $\tilde{M}$ by $M(1-\sigma)$ and $\tilde{M}(1-\sigma)$, respectively, and applying Gronwall's inequality, we obtain

$$R_i^L(t) + \hat{S}_i^L(t) \le \left(\tilde{M} + M\delta\right) e^{M\delta}.$$

Since $|s_i| \le |\hat{s}_i|$, it follows that

$$\max\left\{|(r_i)_x|, |(s_i)_x|\right\} \le \left(\tilde{M} + M\delta\right) e^{M\delta}$$

in $D_\delta^L \cup D_\delta^C$. This proves (3.2) in $D_\delta^L \cup D_\delta^C$ if M_1 is large and δ is small.

We next consider the case where the left end of the branch is a junction. As before, we shall simultaneously consider the branches that are connected to the same junction.This also includes the right triangular regions D_δ^R for the branches that are

connected to the junction from the left. We consider the fixed point equation $w = Kw$, where w and Kw are defined in (2.15) and (2.16), respectively. Differentiating the equations, we obtain (3.10) in $D_\delta^L \cup D_\delta^C$ for $i = j_{\nu+1}, \ldots, j_\mu$ and

(3.13)

$$(\hat{r}_i)_x = \frac{1}{\varepsilon}\left(l_i^R\right)_x U_i(t,x) + \frac{1}{\varepsilon}\int_0^t \left[\left(l_i^R F_i\right)_x + \left(\partial_i^R l_i^R\right)(U_i)_x - \left(l_i^R\right)_x \left(\partial_i^R U_i\right)\right]\left(x_i^R\right)_x dt,$$

$$\begin{aligned}(s_i)_x = {} & \left(\theta_i - l_i^L F_i - \left(\partial_i^L l_i^L\right) U_i\right)(1,\tau)\,\tau_x + \left(l_i^L\right)_x U_i(x,t) - \left(l_i^L\right)_x U_i\left(x_i^L\right)_x(1,\tau) \\ & + \int_\tau^t \left[\left(l_i^L F_i\right)_x + \left(\partial_i^L l_i^L\right)(U_i)_x - \left(l_i^L\right)_x \left(\partial_i^L U_i\right)\right]\left(x_i^L\right)_x dt\end{aligned}$$

in $D_\delta^C \cup D_\delta^R$ for $i = j_1, \ldots, j_\nu$, where

$$\begin{aligned}\zeta_i = {} & \varepsilon \sum_{l=1}^{\nu} \left(\left(n_{j_l}^i\right)_t \hat{r}_{j_l}(1,\tau) + n_{j_l}^i \left(\hat{r}_{j_l}\right)_t\right) \\ & + \varepsilon \sum_{l'=\nu+1}^{\mu} \left(\left(n_{j_{l'}}^i\right)_t \hat{s}_{j_{l'}}(0,\tau) + n_{j_{l'}}^i \left(\hat{s}_{j_{l'}}\right)_t(0,\tau)\right), \\ \theta_i = {} & \varepsilon \sum_{l=1}^{\nu} \left(\left(m_{j_l}^i\right)_t \hat{r}_{j_l}(1,\tau) + m_{j_l}^i \left(\hat{r}_{j_l}\right)_t\right) \\ & + \varepsilon \sum_{l'=\nu+1}^{\mu} \left(\left(m_{j_{l'}}^i\right)_t \hat{s}_{j_{l'}}(0,\tau) + m_{j_{l'}}^i \left(\hat{s}_{j_{l'}}\right)_t(0,\tau)\right),\end{aligned}$$

and m_j^i, n_j^i are defined in the proof of Theorem 2.1. Note that the estimate (3.12) holds for τ_x in both (3.10) and (3.13), although in the latter case, τ is the t-coordinate of the intersection of the left-going characteristic curve x_i^L with the vertical line $x = 1$. The derivation is identical. Hence, there is a constant ε, independent of M_1, such that for $t \in [0, \delta]$,

$$\begin{aligned}\varepsilon\,|\tau_x| \left(\sum_{k=1}^{\nu} \left|m_{j_k}^i(t)\right| + \sum_{k'=\nu+1}^{\mu} \left|m_{j_{k'}}^i(t)\right|\right) < 1, \\ \varepsilon\,|\tau_x| \left(\sum_{k=1}^{\nu} \left|n_{j_k}^i(t)\right| + \sum_{k'=\nu+1}^{\mu} \left|n_{j_{k'}}^i(t)\right|\right) < 1.\end{aligned}$$

Let σ be the maximum of the quantities on the left-hand side of the above inequalities. Define $\hat{R}_i^R$, S_i^R, R_i^L, and $\hat{S}_i^L$ as in (3.9), with obvious modifications. We see that the function

$$W(t) = \sum_{l=1}^{\nu} \left(\hat{R}_{j_l}^R(t) + S_{j_l}^R(t)\right) + \sum_{l'=\nu+1}^{\mu} \left(R_{j_{l'}}^L(t) + \hat{S}_{j_{l'}}^L(t)\right)$$

satisfies the inequality

$$\begin{aligned}(1-\sigma)\,W(t) &\le \sum_{l=1}^{\nu} \left((1-\sigma)\,\hat{R}_{j_l}^R(t) + S_{j_l}^R(t)\right) + \sum_{l'=\nu+1}^{\mu} \left(R_{j_{l'}}^L(t) + (1-\sigma)\,\hat{S}_{j_{l'}}^L(t)\right) \\ &\le \tilde{M} + M\delta + M\int_0^t (1 + W(t'))\,dt'.\end{aligned}$$

Hence, by rescaling and using Gronwall's inequality, we achieve

$$W(t) \leq \left(\tilde{M} + M\delta\right) e^{M\delta}.$$

This proves that

$$\max\left\{\left|\left(r_i\right)_x\right|, \left|\left(s_i\right)_x\right|\right\} \leq M_1$$

in D_δ^R for $i = j_1, \ldots, j_\nu$ and in D_δ^L for $i = j_{\nu+1}, \ldots, j_\mu$ if M_1 is sufficiently large and δ is sufficiently small. Thus, we have proved (3.2) in this case.

It remains to treat the branches that are connected to terminals. If the terminal boundary condition is either (1.5) or (1.6), the argument is parallel to the one given above for sources. Hence, we consider only the case where the boundary condition is (1.7). The fixed point equation in this case is (2.19). Differentiating (2.19) with respect to x gives (3.13) with

$$\zeta_i = \varepsilon \left(\frac{m_i}{n_i}\right)_t \tau_x \hat{r}_i(1,\tau) + \varepsilon \frac{m_i}{n_i} \tau_x \left(\hat{r}_i\right)_t (1,\tau) - \left(\frac{1}{n_i}\right)_t \int_0^t H_i\left(t', r_i(1,t'), s_i(1,t')\right) dt'.$$

Let δ be sufficiently small such that $|\tau_x|$ is bounded by a constant independent of M_1. Choose $\varepsilon > 0$ such that

$$\sigma =: \varepsilon \left|\lambda_i^R(1,t)\right| \left|\frac{m_i}{n_i} \tau_x(1,t)\right| < 1$$

for $t \in [0, \delta]$. Note that $(\frac{m_i}{n_i})_t$ and $(\frac{1}{n_i})_t$ are bounded (by a constant depending on M_1). Hence,

$$\hat{R}_i^R(t) + S_i^R(t) \leq \sigma \hat{R}_i^R(t) + \tilde{M} + M\delta + M \int_0^t \left(1 + \hat{R}_i^R(t') + S_i^R(t')\right) dt'.$$

This leads to

$$\hat{R}_i^R(t) + S_i^R(t) \leq \left(\tilde{M} + M\delta\right) e^{M\delta}$$

in D_δ^R upon rescaling of constants. Hence, (3.2) holds in D_δ^R.

This completes the proof of (3.2) for all cases. By choosing appropriate values of M_1 and δ, we thus obtain a set X_{δ,M_0} in the form of (3.1), which is invariant under the mapping S.

We now show that S is a contraction in X_{δ,M_0}. Let $U = Sv$, $\tilde{U} = S\tilde{v}$ for some $v, \tilde{v} \in X_\delta$, and let $W = U - \tilde{U}$. W satisfies the vanishing initial and external boundary conditions, and its differential equations take the form of (1.8) with the coefficients

$$a_i = a_i(x,t,v), \; b_i = b_i(x,t,v), \; c_i = c_i(x,t,v)$$

and the forcing functions f_i and g_i replaced by

$$\hat{f}_i =: f_i(x,t,v) - f_i(x,t,\tilde{v}) + \left(a_i(x,t,v) - a_i(x,t,\tilde{v})\right) \frac{\partial \tilde{Q}_i}{\partial x} \tag{3.14}$$

and

$$\begin{aligned} \hat{g}_i =: {} & g_i(x,t,v) - g_i(x,t,\tilde{v}) + \left(b_i(x,t,v) - b_i(x,t,\tilde{v})\right) \frac{\partial \tilde{P}_i}{\partial x} \\ & + 2\left(c_i(x,t,v) - c_i(x,t,\tilde{v})\right) \frac{\partial \tilde{Q}_i}{\partial x}, \end{aligned} \tag{3.15}$$

respectively. By the Lipschitz property and the boundedness $|\tilde{U}|_{1,\delta} \le M_1$, there is a constant M such that

$$|\hat{f}|_{0,\delta} \le M\,|v-\tilde{v}|_{0,\delta}\,, \quad |\hat{g}|_{0,\delta} \le M\,|v-\tilde{v}|_{0,\delta}\,.$$

Hence, by Theorem 2.2,

$$|Sv - S\tilde{v}|_{0,\delta} \le M\delta\,|v-\tilde{v}|_{0,\delta}\,.$$

Therefore, S is contracting in X_{δ,M_0} if δ is sufficiently small.

The rest is standard (cf., e.g., [10]). Starting with a $v_0 \in X_{\delta,M_0}$, we generate an iterative sequence $v_{k+1} = Sv_k$. Clearly, each v_k lies in X_{δ,M_0}, and the sequence converges uniformly. The limit then satisfies the integral equations in the proof of Theorem 2.1 and, hence, is differentiable. Therefore, it is the solution of the quasi-linear differential equations. This proves the existence and uniqueness of the solution when $U^I = 0$.

If $U^I \neq 0$, we regard U^I as a vector function of x and t and introduce $\tilde{U} = U - U^I$. It follows that $\tilde{U}$ is a solution of the quasi-linear equations (1.8), with the forcing functions $\tilde{f}_i$ and $\tilde{g}_i$ given by

$$\tilde{f}_i = f_i - \left(Q_i^I\right)_x a_i, \quad \tilde{g}_i = g_i - \left(P_i^I\right)_x b_i - \left(Q_i^I\right)_x 2c_i$$

and the boundary functions given by

$$\tilde{P}_i^B = P_i^B - P_i^I, \quad \tilde{Q}_i^B = Q_i^B - Q_i^I, \quad \tilde{W}_i^B = W_i^B - \delta_i P_i^I + \varepsilon_i Q_i^I.$$

Since $\tilde{U}$ has vanishing initial values, it can be uniquely solved for an interval of $t \in [0,\delta]$. This gives rise to a solution U. ∎

Remark. Examples can be constructed to show that if the condition (2.3) fails at $t = 0$, then the local solution need not exist or may be not unique. In particular, if (2.3) fails at a source end, then the system is underdetermined, and if it fails at a terminal end, the system is overdetermined. See section 5 for further discussion.

We next give a result for the continuity of dependence of the solution and its derivatives on the initial, boundary, and forcing functions and their derivatives. This follows from an argument similar to the proofs of Lemma 2.2 and Theorem 3.1.

COROLLARY 3.2. *Let $U = (P,Q)$ and $\tilde{U} = (\tilde{P},\tilde{Q})$ be two solutions of the quasi-linear problem of Theorem* 3.1. *Suppose the conditions of that theorem hold for the initial and boundary functions of both solutions. Then there exists a constant $M > 0$, independent of initial, boundary, and forcing functions, such that*

$$\begin{aligned}
&\left|U-\tilde{U}\right|_{k,\delta} \\
&\le M\left(\left|P^I-\tilde{P}^I\right|_{C^k[0,1]} + \left|Q^I-\tilde{Q}^I\right|_{C^k[0,1]} + \left|P^B-\tilde{P}^B\right|_{C^k[0,\delta]} + \left|Q^B-\tilde{Q}^B\right|_{C^k[0,\delta]}\right. \\
&\qquad\qquad \left. + \delta\left|f-\tilde{f}\right|_{C^k\left(\overline{D_\delta}\right)} + \delta\,|g-\tilde{g}|_{C^k\left(\overline{D_\delta}\right)} + \delta\left|W-\tilde{W}\right|_{C^k[0,\delta]}\right) \qquad (3.16)
\end{aligned}$$

for $k = 0, 1$.

Proof. For $k = 0$, the result follows from substituting one of the solutions into the coefficients, modifying the forcing functions by (3.14)–(3.15), and using Lemma 2.2. For $k = 1$, we differentiate the equations and apply the lemma to the resulting equations for the derivatives of the solution. The process is standard and is omitted. ∎

4. A finite-difference scheme. In this section, we present a finite-difference scheme that computes discretized solutions and also prove the convergence of the scheme.

The scheme is based on the equations in (2.6). Substituting (2.4) and (2.8) into (2.6), we obtain the normal form of the equations

$$-\lambda_i^L P_{i,t} + a_i Q_{i,t} + \lambda_i^R \left(-\lambda_i^L P_{i,x} + a_i Q_{i,x}\right) = d_i^R,$$

$$-\lambda_i^R P_{i,t} + a_i Q_{i,t} + \lambda_i^L \left(-\lambda_i^R P_{i,x} + a_i Q_{i,x}\right) = d_i^L,$$

where

$$d_i^R(x,t,P_i,Q_i) = -\lambda_i^L f_i + a_i g_i, \qquad d_i^L(x,t,P_i,Q_i) = -\lambda_i^R f_i + a_i g_i.$$

Let h and k be the spatial and temporal step sizes, respectively. Hence, $hN = 1$ for some integer N. We impose the finite-difference equations as

$$\begin{aligned} &\frac{1}{k}\left[-\lambda_{i,n}^{L,m}\left(p_{i,n}^{m+1} - p_{i,n}^{m}\right) + a_{i,n}^{m}\left(q_{i,n}^{m+1} - q_{i,n}^{m}\right)\right] \\ &\quad + \frac{\lambda_{i,n}^{R,m}}{h}\left[-\lambda_{i,n}^{L,m}\left(p_{i,n}^{m} - p_{i,n-1}^{m}\right) + a_{i,n}^{m}\left(q_{i,n}^{m} - q_{i,n-1}^{m}\right)\right] = d_{i,n}^{R,m} \end{aligned} \tag{4.1}$$

for $n = 1, \ldots, N$ and

$$\begin{aligned} &\frac{1}{k}\left[-\lambda_{i,n}^{R,m}\left(p_{i,n}^{m+1} - p_{i,n}^{m}\right) + a_{i,n}^{m}\left(q_{i,n}^{m+1} - q_{i,n}^{m}\right)\right] \\ &\quad + \frac{\lambda_{i,n}^{L,m}}{h}\left[-\lambda_{i,n}^{L,m}\left(p_{i,n+1}^{m} - p_{i,n}^{m}\right) + a_{i,n}^{m}\left(q_{i,n+1}^{m} - q_{i,n}^{m}\right)\right] = d_{i,n}^{L,m} \end{aligned} \tag{4.2}$$

for $n = 0, \ldots, N-1$, where $a_{i,n}^m$, etc., are the values of the respective functions a_i, etc., at the point $\left(nh, mk, p_{i,n}^m, q_{i,n}^m\right)$. (In this section, n is always the running index for the spatial variable, not the number of branches.) The initial condition is simply

$$p_{i,n}^0 = P_i^I(nh), \quad q_{i,n}^0 = Q_i^I(nh). \tag{4.3}$$

If for a fixed m the quantities $p_{i,n}^m$ and $q_{i,n}^m$ are constructed for $n = 0, \ldots, N$, then (4.1) and (4.2) determine $p_{i,n}^{m+1}$ and $q_{i,n}^{m+1}$ for $n = 1, \ldots, N-1$. The quantities for $n = 0$ and N are determined by boundary conditions. At a source end, if the boundary condition is given by (1.2), we impose

$$p_{i,0}^{m+1} = P_i^B((m+1)k) \tag{4.4}$$

and solve $q_{i,0}^{m+1}$ from (4.2) with $n = 0$. If the boundary condition is (1.3), we impose

$$q_{i,0}^{m+1} = Q_i^B((m+1)k) \tag{4.5}$$

and solve $p_{i,0}^{m+1}$ from (4.2). At a junction with $j_1, \ldots, j_\nu$ incoming branches and $j_{\nu+1}, \ldots, j_\mu$ outgoing branches, we prescribe

$$p_{j_1,N}^{m+1} = p_{j_{l'},0}^{m+1} =: p^{m+1} \tag{4.6}$$

for $l = 1, \ldots, \nu$, $l' = \nu+1, \ldots, \mu$, and

$$\sum_{l=1}^{\nu} q_{j_l,N}^{m+1} = \sum_{l'=\nu+1}^{\mu} q_{j_{l'},0}^{m+1}. \tag{4.7}$$

These equations are solved jointly with (4.1) at $n = N$ for $i = j_1, \ldots, j_\nu$, and with (4.2) at $n = 0$ for $i = j_{\nu+1}, \ldots, j_\mu$. The reason that the quantities p^{m+1}, $q^{m+1}_{j_l,N}$, and $q^{m+1}_{j_{l'},0}$ can be uniquely solved is that the coefficient matrix

$$\begin{pmatrix} 0 & R_1 & R_2 \\ -\frac{1}{k}S_1 & \frac{1}{k}A_1 & 0 \\ -\frac{1}{k}S_2 & 0 & \frac{1}{k}A_2 \end{pmatrix}$$

with

$$\begin{aligned} &R_1 = (1, \ldots, 1), \quad R_2 = (-1, \ldots, -1), \\ &S_1 = \left(\lambda^{L,m}_{j_1,N}, \ldots, \lambda^{L,m}_{j_\nu,N}\right)^T, \quad S_2 = \left(\lambda^{R,m}_{j_1,0}, \ldots, \lambda^{R,m}_{j_\nu,0}\right)^T, \\ &A_1 = \operatorname{diag}\left(a^m_{j_1,N}, \ldots, a^m_{j_\nu,N}\right), \quad A_2 = \operatorname{diag}\left(a^m_{j_{\nu+1},0}, \ldots, a^m_{j_\mu,0}\right) \end{aligned}$$

has the determinant

$$\frac{1}{k^\mu}\left(-\sum_{l=1}^{\nu}\frac{\lambda^{L,m}_{j_l,N}}{a^m_{j_l,N}} + \sum_{l'=\nu+1}^{\mu}\frac{\lambda^{R,m}_{j_{l'},0}}{a^m_{j_{l'},0}}\right)\prod_{l=1}^{\nu}a^m_{j_l,N}\prod_{l'=\nu+1}^{\mu}a^m_{j_{l'},0} > 0.$$

(Here we used the fact $\lambda^L_i < 0$, $\lambda^R_i > 0$, and $a_i > 0$.) At a terminal end with the boundary condition (1.5) (resp., (1.6)), we impose

$$p^{m+1}_{i,N} = P^B_i\left((m+1)\,k\right) \quad \left(\text{resp., } q^{m+1}_{i,N} = Q^B_i\left((m+1)\,k\right)\right) \tag{4.8}$$

and solve the other quantity from (4.1) with $n = N$. If the boundary condition is (1.7), we impose

$$\begin{aligned} &\frac{1}{k}\left(p^{m+1}_{i,N} - p^m_{i,N}\right) - \frac{\eta_i}{k}\left(q^{m+1}_{i,N} - q^m_{i,N}\right) + \frac{\delta_i}{2}\left(p^{m+1}_{i,N} + p^m_{i,N}\right) \\ &\qquad - \frac{\varepsilon_i}{2}\left(q^{m+1}_{i,N} + q^m_{i,N}\right) = W^B_i\left(\left(m + \frac{1}{2}\right)k\right). \end{aligned} \tag{4.9}$$

Together with (4.1) for $n = N$, the values of $p^{m+1}_{i,N}$ and $q^{m+1}_{i,N}$ are uniquely determined. This is because the coefficient matrix has the determinant

$$\det\begin{pmatrix} -\frac{\lambda^{L,m}_{i,N}}{k} & \frac{a^m_{i,N}}{k} \\ \frac{1}{k} + \frac{\delta_i}{2} & -\frac{\eta_i}{k} - \frac{\varepsilon_i}{2} \end{pmatrix} < 0.$$

(One might suspect that the simpler condition

$$\frac{1}{k}\left(p^{m+1}_{i,N} - p^m_{i,N}\right) - \frac{\eta_i}{k}\left(q^{m+1}_{i,N} - q^m_{i,N}\right) + \delta_i p^m_{i,N} - \varepsilon_i q^m_{i,N} = W^B_i\,(mk) \tag{4.10}$$

would also suffice. It indeed can determine unique values of $p^{m+1}_{i,N}$ and $q^{m+1}_{i,N}$. However, we are unable to prove the convergence of the scheme with this condition. The difficulty will be clear from the proof of the next theorem.)

It is clear that for any step sizes h and k this scheme generates a discretized solution as long as λ^L_i remains negative at $x = 0$ and $x = 1$. We show that if the ratio k/h is fixed and sufficiently small, then in a time interval the solutions for the

finite-difference equations converge to the solution of the original system of differential equations (1.8) as $h \to 0$.

THEOREM 4.1. *Suppose that the conditions of Theorem* 3.1 *hold and that*

$$a_i(x,t,p,q) > 0, \quad \lambda_i^L(x,t,p,q) < 0$$

for all $(x,t) \in [0,1] \times [0,\delta]$ *and* $(p,q) \in \mathbb{R}^2$, *where* $\delta > 0$ *appears in Theorem* 3.1. *Suppose also that the initial and boundary functions* P_i^I, Q_i^I, P_i^B, Q_i^B, *and* W_i^B *have continuous second derivatives. Let* $\sigma > 0$ *be a positive constant such that*

$$\sigma \max\left\{ \left|\lambda_i^L\right|_{0,\delta}, \left|\lambda_i^R\right|_{0,\delta} \right\} < 1, \tag{4.11}$$

and let the ratio $k/h = \sigma$ *be fixed. Then there is a constant* $\delta_0 > 0$ *such that, as* $h \to 0$, *the solutions of the finite-difference scheme described above converge to the solution of the differential equation* (1.8) *in the strip* $0 \le t \le \delta_0$.

Remark. The condition of $a_i > 0$, $\lambda_i^L < 0$ for all (p,q) is stronger than needed. One may require only that the inequalities hold in a certain range of (p,q) containing the solution (P_i, Q_i) in its interior. The theorem is stated as above to simplify the argument.

Proof. By Theorem 3.1, the system of differential equations has a solution (P_i, Q_i) in D_δ for some $\delta > 0$. Since the initial and boundary functions have continuous second derivatives, it can be shown using standard arguments that the solution (P_i, Q_i) has continuous second-order derivatives in D_δ. (Reduce δ if necessary.) By Taylor's theorem and $k = \sigma h$, we can write

$$\begin{aligned} &\frac{1}{k}\left[-\tilde{\lambda}_{i,n}^{L,m}\left(P_{i,n}^{m+1} - P_{i,n}^m\right) + \tilde{a}_{i,n}^m\left(Q_{i,n}^{m+1} - Q_{i,n}^m\right)\right] \\ &\quad + \frac{\tilde{\lambda}_{i,n}^{R,m}}{h}\left[-\tilde{\lambda}_{i,n}^{L,m}\left(P_{i,n}^m - P_{i,n-1}^m\right) + \tilde{a}_{i,n}^m\left(Q_{i,n}^m - Q_{i,n-1}^m\right)\right] = \tilde{d}_{i,n}^{R,m} + O(h) \end{aligned} \tag{4.12}$$

for $n = 1, \ldots, N$, and

$$\begin{aligned} &\frac{1}{k}\left[-\tilde{\lambda}_{i,n}^{R,m}\left(P_{i,n}^{m+1} - P_{i,n}^m\right) + \tilde{a}_{i,n}^m\left(Q_{i,n}^{m+1} - Q_{i,n}^m\right)\right] \\ &\quad + \frac{\tilde{\lambda}_{i,n}^{L,m}}{h}\left[-\tilde{\lambda}_{i,n}^{R,m}\left(P_{i,n}^m - P_{i,n-1}^m\right) + \tilde{a}_{i,n}^m\left(Q_{i,n}^m - Q_{i,n-1}^m\right)\right] = \tilde{d}_{i,n}^{L,m} + O(h) \end{aligned} \tag{4.13}$$

for $n = 0, \ldots, N-1$, where $P_{i,n}^m$ and $Q_{i,n}^m$ are the values of the corresponding functions at the point (nh, mk), and $\tilde{\lambda}_{i,n}^{L,m}$, etc., represent the values of the corresponding functions at the point $\left(nh, mk, P_{i,n}^m, Q_{i,n}^m\right)$. Let

$$u_{i,n}^m = P_{i,n}^m - p_{i,n}^m, \quad v_{i,n}^m = Q_{i,n}^m - q_{i,n}^m.$$

Our task is to show

$$u_{i,n}^m \to 0, \quad v_{i,n}^m \to 0$$

as $h \to 0$ and $k = \sigma h$. We prove it by showing that there are positive constants δ_0, h_0, and M, independent of m, such that

$$\left|u_{i,n}^m\right| \le Mh, \quad \left|v_{i,n}^m\right| \le Mh \tag{4.14}$$

if $h \le h_0$, $k = \sigma h$, and $0 \le mk \le \delta_0$.

We first derive some recursive relations. Subtract (4.1) and (4.2) from (4.12) and (4.13), respectively, and use the Lipschitz property and the boundedness of the derivatives of P_i and Q_i. In this way we obtain

$$\begin{aligned}
&\frac{1}{k}\left[-\lambda_{i,n}^{L,m}\left(u_{i,n}^{m+1}-u_{i,n}^{m}\right)+a_{i,n}^{m}\left(v_{i,n}^{m+1}-v_{i,n}^{m}\right)\right]\\
&\qquad+\frac{\lambda_{i,n}^{R,m}}{h}\left[-\lambda_{i,n}^{L,m}\left(u_{i,n}^{m}-u_{i,n-1}^{m}\right)+a_{i,n}^{m}\left(v_{i,n}^{m}-v_{i,n-1}^{m}\right)\right]\\
&\quad=O\left(h\right)+\tilde{d}_{i,n}^{R,m}-d_{i,n}^{R,m}+\left(\tilde{\lambda}_{i,n}^{L,m}-\lambda_{i,n}^{L,m}\right)\frac{P_{i,n}^{m+1}-P_{i,n}^{m}}{k}\\
&\qquad-\left(\tilde{a}_{i,n}^{m}-a_{i,n}^{m}\right)\frac{Q_{i,n}^{m+1}-Q_{i,n}^{m}}{k}+\left(\tilde{\lambda}_{i,n}^{R,m}\tilde{\lambda}_{i,n}^{L,m}-\lambda_{i,n}^{R,m}\lambda_{i,n}^{L,m}\right)\frac{P_{i,n}^{m}-P_{i,n-1}^{m}}{h}\\
&\qquad-\left(\tilde{\lambda}_{i,n}^{R,m}\tilde{a}_{i,n}^{m}-\lambda_{i,n}^{R,m}a_{i,n}^{m}\right)\frac{Q_{i,n}^{m}-Q_{i,n-1}^{m}}{h}\\
&\quad=O\left(h\right)+O\left(u_{i,n}^{m}\right)+O\left(v_{i,n}^{m}\right)
\end{aligned}\tag{4.15}$$

and, similarly,

$$\begin{aligned}
&\frac{1}{k}\left[-\lambda_{i,n}^{R,m}\left(u_{i,n}^{m+1}-u_{i,n}^{m}\right)+a_{i,n}^{m}\left(v_{i,n}^{m+1}-v_{i,n}^{m}\right)\right]\\
&\qquad+\frac{\lambda_{i,n}^{L,m}}{h}\left[-\lambda_{i,n}^{R,m}\left(u_{i,n+1}^{m}-u_{i,n}^{m}\right)+a_{i,n}^{m}\left(v_{i,n+1}^{m}-v_{i,n}^{m}\right)\right]\\
&\qquad\qquad=O\left(h\right)+O\left(u_{i,n}^{m}\right)+O\left(v_{i,n}^{m}\right).
\end{aligned}\tag{4.16}$$

Introduce

$$r_{i,n}^{m}=-\lambda_{i,n}^{L,m-1}u_{i,n}^{m}+a_{i,n}^{m-1}v_{i,n}^{m},\qquad s_{i,n}^{m}=-\lambda_{i,n}^{R,m-1}u_{i,n}^{m}+a_{i,n}^{m-1}v_{i,n}^{m}.$$

One can show that (4.14) is equivalent to

$$\left|r_{i,n}^{m}\right|\le Mh,\quad \left|s_{i,n}^{m}\right|\le Mh.\tag{4.17}$$

(Throughout the proof of this theorem, we use M to denote any positive constant that is independent of m.) Using the identity

$$-\lambda_{i,n}^{L,m}u_{i,l}^{m}+a_{i,n}^{m}v_{i,l}^{m}=r_{i,l}^{m}+\left(\lambda_{i,l}^{L,m-1}-\lambda_{i,n}^{L,m}\right)u_{i,l}^{m}+\left(a_{i,l}^{m-1}-a_{i,n}^{m}\right)v_{i,l}^{m},$$

together with

$$\begin{aligned}
\lambda_{i,l}^{L,m-1}-\lambda_{i,n}^{L,m}&=O\left(k\right)+O\left(p_{i,l}^{m-1}-p_{i,n}^{m}\right)+O\left(q_{i,l}^{m-1}-q_{i,n}^{m}\right),\\
a_{i,l}^{m-1}-a_{i,n}^{m}&=O\left(k\right)+O\left(p_{i,l}^{m-1}-p_{i,n}^{m}\right)+O\left(q_{i,l}^{m-1}-q_{i,n}^{m}\right),
\end{aligned}$$

and

$$\begin{aligned}
p_{i,l}^{m-1}-p_{i,n}^{m}&=-u_{i,l}^{m-1}+u_{i,n}^{m}+\left(P_{i,l}^{m-1}-P_{i,n}^{m}\right),\\
q_{i,l}^{m-1}-q_{i,n}^{m}&=-v_{i,l}^{m-1}+v_{i,n}^{m}+\left(Q_{i,l}^{m-1}-Q_{i,n}^{m}\right)
\end{aligned}$$

for $l = n-1,\ n,\ n+1$, we can write

$$\begin{aligned}
-\lambda_{i,n}^{L,m} u_{i,l}^m + a_{i,n}^m v_{i,l}^m &= r_{i,l}^m + u_{i,l}^m O(k) + v_{i,l}^m O(k) + u_{i,l}^m O\left(u_{i,l}^{m-1}, u_{i,n}^m, v_{i,l}^{m-1}, v_{i,n}^m\right) \\
&\quad + v_{i,l}^m O\left(u_{i,l}^{m-1}, u_{i,n}^m, v_{i,l}^{m-1}, v_{i,n}^m\right), \\
-\lambda_{i,n}^{R,m} u_{i,l}^m + a_{i,n}^m v_{i,l}^m &= s_{i,l}^m + u_{i,l}^m O(k) + v_{i,l}^m O(k) + u_{i,l}^m O\left(u_{i,l}^{m-1}, u_{i,n}^m, v_{i,l}^{m-1}, v_{i,n}^m\right) \\
&\quad + v_{i,l}^m O\left(u_{i,l}^{m-1}, u_{i,n}^m, v_{i,l}^{m-1}, v_{i,n}^m\right),
\end{aligned}$$

where

$$O(u_{i,l}^{m-1}, u_{i,n}^m, v_{i,l}^{m-1}, v_{i,n}^m) = O(u_{i,l}^{m-1}) + O(u_{i,n}^m) + O(v_{i,l}^{m-1}) + O(v_{i,l}^m).$$

Substituting these relations into (4.15) and (4.16), we obtain

$$\begin{aligned}
r_{i,n}^{m+1} &= r_{i,n}^m - \sigma\lambda_{i,n}^{R,m}\left(r_{i,n}^m - r_{i,n-1}^m\right) + O_{i,n,n-1}^m, \qquad n = 1, \dots, N, \\
s_{i,n}^{m+1} &= s_{i,n}^m - \sigma\lambda_{i,n}^{L,m}\left(s_{i,n+1}^m - s_{i,n}^m\right) + O_{i,n,n+1}^m, \qquad n = 0, \dots, N-1,
\end{aligned} \tag{4.18}$$

for $m \geq 1$, where

$$\begin{aligned}
O_{i,n,n-1}^m &= O\left(h^2\right) + h\left(O\left(u_{i,n}^m\right) + O\left(v_{i,n}^m\right)\right) + u_{i,n-1}^m O(h) + v_{i,n-1}^m O(h) \\
&\quad + u_{i,n-1}^m O\left(u_{i,n-1}^{m-1}, u_{i,n}^m, v_{i,n-1}^{m-1}, v_{i,n}^m\right) + v_{i,n-1}^m O\left(u_{i,n-1}^{m-1}, u_{i,n}^m, v_{i,n-1}^{m-1}, v_{i,n}^m\right)
\end{aligned}$$

and $O_{i,n,n+1}^m$ is defined similarly, with $n-1$ substituted by $n+1$. These are the recursive relations we need.

We now prove (4.17). Assume $\delta_0 < \sigma/2$; then, $mk \leq \delta_0$ implies $m < N - m$. The proof will be divided into three cases: (1) $m \leq n \leq N-m$, (2) $0 \leq n < m$, and (3) $N - m < n \leq N$. It may be helpful to compare the argument below with the proof of Theorem 2.1, in which the region D_i is divided into D_i^C, D_i^L, and D_i^R.

Case 1: $m \leq n \leq N-m$. Let

$$e_m = \max_{m \leq n \leq N-m} \left\{\left|r_{i,n}^m\right|, \left|s_{i,n}^m\right|\right\}.$$

In view of (4.11), the coefficients of $r_{i,n}^m$, $r_{i,n-1}^m$, $s_{i,n}^m$, and $s_{i,n+1}^m$ in (4.18) are all nonnegative. Hence, from (4.18),

$$e_{m+1} \leq e_m + C\left(h^2 + he_m + e_m e_{m-1} + e_m^2\right), \quad m \geq 1, \tag{4.19}$$

where $C > 0$ is a constant. By initial condition (4.3),

$$u_{i,n}^0 = v_{i,n}^0 = 0.$$

Thus, $e_0 = 0$. Also, by (4.18) with $m = 0$,

$$\begin{aligned}
r_{i,n}^1 &= O(h^2) \qquad \text{for } n = 1, \dots, N, \\
s_{i,n}^1 &= O(h^2) \qquad \text{for } n = 0, \dots, N-1.
\end{aligned} \tag{4.20}$$

This implies $e_1 = O(h^2)$. Consider the linear difference equation with initial condition

$$E_{m+1} = (1 + 3Ch)\, E_m + Ch^2, \quad m \geq 1, \quad E_1 = C_0 h^2,$$

where C_0 is so large that $e_1 \le C_0 h^2$. It has the solution

$$\begin{aligned} E_{m+1} &= C_0 h^2 \left(1+3Ch\right)^m + \frac{h}{3}\left(\left(1+3Ch\right)^m - 1\right) \\ &\le h\left(C_0 h e^{3Chm} + \frac{1}{3}e^{3Chm} - 1\right). \end{aligned}$$

Let δ_0 be so small that $e^{3C\delta_0/\sigma} < 4$. Then there is an $h_0 > 0$ such that $E_m \le h$ for all $h \le h_0$ and $mk \le \delta_0$. This implies that

$$E_{m+1} \ge E_m + C\left(h^2 + hE_m + E_m E_{m-1} + E_m^2\right), \quad E_1 \ge e_1.$$

Hence,

$$e_m \le E_m \le h,$$

which leads to (4.17) with $M = 1$ in Case 1.

Case 2: $0 \le n < m$. The proof in this case depends on the type of boundary condition at the left end of the branch. Suppose that the end is a source with the boundary condition (4.4). Let

$$e_m = \max_{0\le n\le N-m}\left\{\left|r_{i,n}^m\right|, \left|s_{i,n}^m\right|\right\}.$$

(As was the case in the proof of Theorem 2.1, it is more convenient to include the central trapezoidal part $m \le n \le N-m$.) Hence, from (4.18),

$$\begin{aligned} \left|r_{i,n}^{m+1}\right| &\le |e_m| + C\left(h^2 + he_m + e_m e_{m-1} + e_m^2\right) \quad \text{for } n = 1, \dots, N-m, \\ \left|s_{i,n}^{m+1}\right| &\le |e_m| + C\left(h^2 + he_m + e_m e_{m-1} + e_m^2\right) \quad \text{for } n = 0, \dots, N-m. \end{aligned} \tag{4.21}$$

Since, by (4.4), $u_{i,0}^m = 0$, it follows that $r_{i,0}^m = s_{i,0}^m$ for all m. Therefore, e_m satisfies the same difference inequality (4.19). We also have $e_1 = O(h^2)$ by (4.20). Thus, the above analysis gives $e_m \le h$.

Suppose that the boundary condition is given by (4.5); then $v_{i,0}^m = 0$ and

$$r_{i,0}^m = \frac{\lambda_{i,0}^{L,m-1}}{\lambda_{i,0}^{R,m-1}} s_{i,0}^m$$

for all $m \ge 1$. Let $\hat{r}_{i,n}^m = r_{i,n}^m / M$, where M is sufficiently large such that

$$M > \max_m \left\{\left|\frac{\lambda_{i,0}^{L,m}}{\lambda_{i,0}^{R,m}}\right|\right\}.$$

Then (4.18) still holds, with r substituted by $\hat{r}$. Let

$$e_m = \max_{0\le n\le N-m}\left\{\left|\hat{r}_{i,n}^m\right|, \left|s_{i,n}^m\right|\right\}.$$

We again have (4.21) and

$$\left|\hat{r}_{i,0}^{m+1}\right| \le \left|s_{i,0}^{m+1}\right| \le |e_m| + C\left(h^2 + he_m + e_m e_{m-1} + e_m^2\right).$$

Hence, e_m satisfies (4.19) again. Therefore,

$$\left|r_{i,n}^m\right| \le Mh, \quad \left|s_{i,n}^m\right| \le h.$$

Suppose that the left end is a junction. We shall simultaneously treat all the branches connected to the same junction. Let $j_1, \dots, j_\nu$ be the incoming branches and $j_{\nu+1}, \dots, j_\mu$ the outgoing branches. It is easy to see that the boundary conditions (4.6)–(4.7) are satisfied if p and q are substituted by u and v, respectively. Using the identities

$$u_{i,n}^{m+1} = \frac{r_{i,n}^{m+1} - s_{i,n}^{m+1}}{\lambda_{i,n}^m}, \quad v_{i,n}^{m+1} = \frac{\lambda_{i,n}^{R,m} r_{i,n}^{m+1} - \lambda_{i,n}^{L,m} s_{i,n}^{m+1}}{a_{i,n}^m \lambda_{i,n}^m}, \tag{4.22}$$

where

$$\lambda_{i,n}^m = \lambda_{i,n}^{R,m} - \lambda_{i,n}^{L,m} > 0,$$

the equations for r and s have the form

$$\begin{aligned}
&\frac{1}{\lambda_{j_1,N}^m}\left(r_{j_1,N}^{m+1} - s_{j_1,N}^{m+1}\right) - \frac{1}{\lambda_{i,N}^m}\left(r_{i,N}^{m+1} - s_{i,N}^{m+1}\right) = 0, \quad i = j_2, \dots, j_\nu,\\
&\frac{1}{\lambda_{j_1,N}^m}\left(r_{j_1,N}^{m+1} - s_{j_1,N}^{m+1}\right) - \frac{1}{\lambda_{i,0}^m}\left(r_{i,0}^{m+1} - s_{i,0}^{m+1}\right) = 0, \qquad i = j_{\nu+1}, \dots, j_\mu,\\
&\sum_{l=1}^{\nu} \frac{1}{a_{j_l,N}^m \lambda_{j_l,N}^m}\left(\lambda_{j_l,N}^{R,m} r_{j_l,N}^{m+1} - \lambda_{j_l,N}^{L,m} s_{j_l,N}^{m+1}\right)\\
&\qquad - \sum_{l'=\nu+1}^{\mu} \frac{1}{a_{j_{l'},0}^m \lambda_{j_{l'},0}^m}\left(\lambda_{j_{l'},0}^{R,m} r_{j_{l'},0}^{m+1} - \lambda_{j_{l'},0}^{L,m} s_{j_{l'},0}^{m+1}\right) = 0.
\end{aligned}$$

The system can be solved for $s_{j_1,N}^{m+1}, \dots, s_{j_\nu,N}^{m+1}, r_{j_{\nu+1},0}^{m+1}, \dots, r_{j_\mu,0}^{m+1}$ because the coefficient matrix

$$\begin{pmatrix}
-\frac{1}{\lambda_{j_1,N}^m} & \frac{1}{\lambda_{j_2,N}^m} & \cdots & 0\\
\vdots & \vdots & \ddots & \vdots\\
-\frac{1}{\lambda_{j_1,N}^m} & 0 & \cdots & -\frac{1}{\lambda_{j_\mu,0}^m}\\
-\frac{\lambda_{j_1,N}^{L,m}}{\lambda_{j_1,N}^m a_{j_1,N}^m} & -\frac{\lambda_{j_2,N}^{L,m}}{\lambda_{j_2,N}^m a_{j_2,N}^m} & \cdots & -\frac{\lambda_{j_\mu,0}^{R,m}}{\lambda_{j_\mu,0}^m a_{j_\mu,0}^m}
\end{pmatrix}$$

has the determinant

$$\frac{(-1)^{\nu+1}}{\prod_{l=1}^{\nu} \lambda_{j_l,N}^m \prod_{l'=\nu+1}^{\mu} \lambda_{j_{l'},0}^m}\left(-\sum_{l=1}^{\nu} \frac{\lambda_{j_l,N}^{L,m}}{a_{j_l,N}^m} + \sum_{l'=\nu+1}^{\mu} \frac{\lambda_{j_{l'},0}^{R,m}}{a_{j_{l'},N}^m}\right) \neq 0.$$

(Here we used $\lambda_{i,n}^m > 0$, $a_{i,n}^m > 0$, $\lambda_{i,n}^{R,m} > 0$, and $\lambda_{i,n}^{L,m} < 0$.) Let the solution be written as

$$\begin{aligned}
s_{i,N}^{m+1} &= \sum_{l=1}^{\nu} m_{j_l}^i r_{j_i,N}^{m+1} + \sum_{l'=\nu+1}^{\mu} m_{j_{l'}}^i s_{j_{l'},0}^{m+1}, \qquad i = j_1, \dots, j_\nu,\\
r_{i,0}^{m+1} &= \sum_{l=1}^{\nu} n_{j_l}^i r_{j_i,N}^{m+1} + \sum_{l'=\nu+1}^{\mu} n_{j_{l'}}^i s_{j_{l'},0}^{m+1}, \qquad i = j_{\nu+1}, \dots, j_\mu.
\end{aligned} \tag{4.23}$$

Choose a constant M such that

$$M > \max_{i=j_1,\ldots,j_\mu} \left\{ \sum_{l=1}^{\mu} \left| m_{j_l}^i \right|, \sum_{l=1}^{\mu} \left| n_{j_l}^i \right| \right\},$$

and introduce

$$\hat{s}_{j_l,n}^m = s_{j_l,n}^m / M, \quad \hat{r}_{j_{l'},n}^m = r_{j_{l'},n}^m / M$$

for $l = 1, \ldots, \nu$, $l' = \nu+1, \ldots, \mu$. Equations in (4.18) still hold if $\hat{s}_{j_l,n}^m$ and $\hat{r}_{j_{l'},n}^m$ are substituted for $s_{j_l,n}^m$ and $r_{j_{l'},n}^m$, respectively. Let e_m denote the maximum of the quantities

$$\max_{\substack{m \le n \le N \\ 1 \le l \le \nu}} \left\{ \left| r_{j_l,n}^m \right|, \left| \hat{s}_{j_l,n}^m \right| \right\}, \quad \max_{\substack{0 \le n \le N-m \\ \nu+1 \le l' \le \mu}} \left\{ \left| \hat{r}_{j_{l'},n}^m \right|, \left| s_{j_{l'},n}^m \right| \right\}.$$

(Notice again the inclusion of the middle part $m \le n \le N-m$.) Since the coefficients of r and s are all positive, it is easy to see that

$$\left| r_{j_l,n}^{m+1} \right| \le e_m + C\left(h^2 + h e_m + e_m e_{m-1} + e_m^2\right)$$

for $l = 1, \ldots, \nu$, $n = m, \ldots, N$ and

$$\left| s_{j_{l'},n}^{m+1} \right| \le e_m + C\left(h^2 + h e_m + e_m e_{m-1} + e_m^2\right)$$

for $l = \nu+1, \ldots, \mu$, $n = 0, \ldots, m$. Similar inequalities can be derived for $|\hat{s}_{j_l,n}^{m+1}|$, $l = 1, \ldots, \nu$, $n = m, \ldots, N-1$, and for $|\hat{r}_{j_{l'},n}^{m+1}|$, $l' = \nu+1, \ldots, \mu$, $n = 1, \ldots, m$. Furthermore, by (4.23),

$$\left| \hat{s}_{j_l,N}^{m+1} \right| = \frac{1}{M} \left| \sum_{l=1}^{\nu} m_{j_l}^i r_{j_i,N}^{m+1} + \sum_{l'=\nu+1}^{\mu} m_{j_{l'}}^i s_{j_{l'},0}^{m+1} \right| \le \max_{\substack{1 \le l \le \nu \\ \nu+1 \le l' \le \mu}} \left\{ \left| r_{j_l,N}^{m+1} \right|, \left| s_{j_{l'},0}^{m+1} \right| \right\},$$

$$\left| \hat{r}_{j_l,N}^{m+1} \right| = \frac{1}{M} \left| \sum_{l=1}^{\nu} n_{j_l}^i r_{j_i,N}^{m+1} + \sum_{l'=\nu+1}^{\mu} n_{j_{l'}}^i s_{j_{l'},0}^{m+1} \right| \le \max_{\substack{1 \le l \le \nu \\ \nu+1 \le l' \le \mu}} \left\{ \left| r_{j_l,N}^{m+1} \right|, \left| s_{j_{l'},0}^{m+1} \right| \right\}.$$

Therefore, we achieve again the difference inequality (4.19) for e_m. Hence, $e_m \le h$, and consequently,

$$\left| r_{i,n}^m \right| \le h, \quad \left| s_{i,n}^m \right| \le Mh.$$

This not only proves (4.17) for Case 2, but also for the part of Case 3 where the right endpoint is a junction.

Case 3: $N - m \le n \le N$. It remains only to discuss the case where the right end is a terminal. If the boundary condition is given by (4.8), the results follow from arguments similar to those in Case 2, when the source end boundary condition is either (4.4) or (4.5). Thus, we shall discuss only the case when the boundary condition is given by (4.9), which corresponds to the Windkessel-type boundary condition (1.7) for the differential equations.

From (1.7), we derive

$$\frac{1}{k}\left(P_{i,N}^{m+1}-P_{i,N}^{m}\right)-\frac{\eta_i}{k}\left(Q_{i,N}^{m+1}-Q_{i,N}^{m}\right)+\frac{\delta_i}{2}\left(P_{i,N}^{m+1}+P_{i,N}^{m}\right)$$
$$-\frac{\varepsilon_i}{2}\left(Q_{i,N}^{m+1}+Q_{i,N}^{m}\right)=W_i^B\left(\left(m+\frac{1}{2}\right)k\right)+O\left(k^2\right).$$

Subtracting (4.9) from above yields

$$\frac{1}{k}(u_{i,N}^{m+1}-u_{i,N}^{m})-\frac{\eta_i}{k}(v_{i,N}^{m+1}-v_{i,N}^{m})+\frac{\delta_i}{2}(u_{i,N}^{m+1}+u_{i,N}^{m})-\frac{\varepsilon_i}{2}(v_{i,N}^{m+1}+v_{i,N}^{m})=O(k^2).$$

Let

$$f^m=\left(1+\frac{\delta_i k}{2}\right)u_{i,N}^{m}-\left(\eta_i+\frac{\varepsilon_i k}{2}\right)v_{i,N}^{m},\quad m=0,1,\ldots.$$

The equation for f^m has the form

$$f^{m+1}=f^m+k\left(\varepsilon_i v_{i,N}^{m}-\delta_i u_{i,N}^{m}\right)+O\left(k^3\right).$$

Since $f^0=0$, the difference equation has the solution

$$f^{m+1}=k\sum_{j=0}^{m}\left(\varepsilon_i v_{i,N}^{j}-\delta_i u_{i,N}^{j}\right)+O\left(k^2\right).$$

From (4.22), we obtain

$$s_{i,N}^{m+1}=\frac{M_i^m}{N_i^m}r_{i,N}^{m+1}-\frac{k}{N_i^m}\sum_{j=0}^{m}\left(\varepsilon_i v_{i,N}^{j}-\delta_i u_{i,N}^{j}\right)+O\left(k^2\right),\tag{4.24}$$

where

$$M_i^m=\frac{1}{\lambda_{i,n}^m}\left(1+\frac{\delta_i k}{2}-\left(\eta_i+\frac{\varepsilon_i k}{2}\right)\frac{\lambda_{i,n}^{R,m}}{a_{i,n}^m}\right),$$
$$N_i^m=\frac{1}{\lambda_{i,n}^m}\left(1+\frac{\delta_i k}{2}-\left(\eta_i+\frac{\varepsilon_i k}{2}\right)\frac{\lambda_{i,n}^{L,m}}{a_{i,n}^m}\right).$$

(Notice that $N_i^m>0$, and hence (4.24) is valid.) Let $\hat{s}_{i,n}^m=s_{i,n}^m/M$, where M is a constant to be determined later. Also let

$$e_m=\max_{\substack{m\le n\le N\\ 0\le j\le m}}\left\{\left|r_{i,n}^{j}\right|,\left|\hat{s}_{i,n}^{j}\right|\right\}.$$

Unlike previous cases where e_m depends on the mth level quantities, here it is more convenient to let e_m be the maximum of all the lower level quantities. Then, by (4.18) modified with $\hat{s}$ substituted for s,

$$\left|r_{i,n}^{m+1}\right|\le e_m+C\left(h^2+he_m+e_me_{m-1}+e_m^2\right)\tag{4.25}$$

for $n=m,\ldots,N$ and

$$\left|\hat{s}_{i,n}^{m+1}\right|\le e_m+C\left(h^2+he_m+e_me_{m-1}+e_m^2\right)$$

for $n = m, \ldots, N-1$, where C is a positive constant. Also, by (4.24) and the relation $mk \leq \delta_0$,

$$\left|\hat{s}_{i,N}^{m+1}\right| \leq \frac{1}{M}\left|\frac{M_i^m}{N_i^m}\right|\left|r_{i,N}^{m+1}\right| + \delta_0 C' e_m + O\left(h^2\right),$$

where $C' > 0$ is constant. Hence, from (4.25) we see that if M is sufficiently large and δ_0 is sufficiently small, we can ensure

$$\left|\hat{s}_{i,N}^{m+1}\right| \leq e_m + C\left(h^2 + he_m + e_m e_{m-1} + e_m^2\right).$$

(This is where the simpler boundary condition (4.10) fails. Instead of $O\left(h^2\right)$, it can provide only $O(h)$, which is inconsistent with (4.19).) Thus, e_m satisfies the relation (4.19), which leads to $e_m \leq h$. We have, thus, shown that

$$\left|r_{i,n}^m\right| \leq h, \quad \left|s_{i,n}^m\right| \leq Mh.$$

This completes the proof of Case 3, and also of the entire theorem. ∎

5. Discussion. We have given a rather thorough treatment to the initial-boundary value problem of the first-order quasilinear system (1.8) with various source and terminal boundary conditions. From our results, it can be seen that the junction condition (1.4), which stems from the conservation of mass and Navier–Stokes momentum, is consistent with the differential equations. Also, the Windkessel-type terminal boundary condition does not cause problems for the solvability. However, due to the nature of the first-order hyperbolic equations, the existence of a global solution generally is not guaranteed (cf. [14, 15]). This problem may disappear if more accurate models are used. For example, in (1.8) and its special case (1.1), only the effect of viscosity on the wall of the vessels is taken into consideration. If we include viscosity more comprehensively, a term of $\mu\nabla^2 Q_i$ appears in the right-hand sides of the second equations of (1.8) and (1.1). The system then becomes parabolic instead of hyperbolic. It is well known that parabolic systems have better regularity properties than hyperbolic ones. Therefore, it may be possible to prove the existence of a global solution. Another possible approach is to use the results of Čanić and Kim [5] that shock waves can develop only in an unrealistically long vessel. With appropriate a priori estimates, one might find a range of lengths of vessels within which global solutions exist. We are currently investigating these issues.

We have developed a numerical scheme for the computation of solutions and proved its convergence. Although our scheme uses a nonstaggered method similar to the one developed by Raines, Jaffrin, and Shapiro [23, 24], they are substantially different. (By nonstaggered, we mean that the values of P_i and Q_i are approximated at the same mesh points, unlike the staggered method developed in [9, 13].) This is because ours is based on the normal form of the equations and takes into account the characteristic directions. This might explain why our scheme converges even if the network has loops, while the other can break down (cf. [13]).

REFERENCES

[1] M. Anliker, R. Rockwell, and E. Ogden, *Non-linear analysis of flow pulses and shock waves in arteries*, Z. Angew. Math. Phys., 22 (1971), pp. 217–246.

[2] M. Anliker, J. Stettler, P. Niederer, and R. Holenstein, *Prediction of shape changes of propagating flow and pressure pulses in human arteries*, in Dynamics, Control Theory and Regulation, Symposium, Erlangen, Germany, 1977, R. Bauer and R. Busse, eds., Springer-Verlag, New York, 1978, pp. 15–34.

[3] C. Almeder, F. Breitenecker, S. Wassertheurer, K. Kaser, J. Krocza, and M. Suda, *Modelling of the human arterial network for an expert system for preoperative predictions*, in Proceedings of the 2nd Mathmod, IMACS Symposium on Mathematical Modelling, Vienna, Austria, 1997, p. 1110.

[4] S. Čanić, *Blood flow through compliant vessels after endovascular repair: Wall deformations induced by the discontinuous wall properties*, Comput. Vis. Sci., 4 (2002), pp. 147–155.

[5] S. Čanić and E.H. Kim, *Mathematical analysis of the quasilinear effects in a hyperbolic model of blood flow through compliant axi-symmetric vessels*, Math. Meth. Appl. Sci., 26 (2003), pp. 1161–1186.

[6] F.T. Charbel, M.E. Clark, M. Misra, K. Hannigan, W.E. Hoffman, and J.I. Ausman, *The application of a computerized model of the cerebral circulation in skull base surgery*, in Proceedings of the 2nd International Skull Base Congress, San Diego, 1996, p. 210.

[7] F.T. Charbel, M. Misra, M.E. Clark, and J.I. Ausman, *Computer simulation of cerebral blood flow in Moyamoya and the results of surgical therapies*, Clinical Neurol. Neurosurg., 99 (1997), Supp. 2, pp. 563–573.

[8] M.E. Clark, M. Zhao, F. Loth, N. Alperin, L. Sadler, K. Guppy, and F.T. Charbel, *A patient-specific computer model for prediction of clinical outcomes in the cerebral circulation using MR flow measurements*, in Medical Image Computing and Computer-Assisted Intervention, Cambridge, England, 1999.

[9] M.E. Clark and R.H. Kufahl, *Simulation of the cerebral macrocirculation*, in Cardiovascular Systems Dynamics, MIT Press, Cambridge, MA, 1978, pp. 380–390.

[10] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, John Wiley & Sons, New York, 1962.

[11] L. Formaggia, F. Nobile, and A. Quarteroni, *A one dimensional model for blood flow: Application to vascular prosthesis*, in Mathematical Modeling and Numerical Simulation in Continuum Mechanics (Yamaguchi, 2000), Lecture Notes in Comput. Sci. Eng. 19, Springer, New York, 2002, pp. 137–153.

[12] G.E. Forsythe and W.R. Wasow, *Finite-Difference Methods for Partial Differential Equations*, John Wiley & Sons, New York, 1960.

[13] R.H. Kufahl and M.E. Clark, *A circle of Willis simulation using distensible vessels and pulsatile flow*, J. Biomech. Eng., 107 (1985), pp. 112–122.

[14] T.-T. Li and W. Yu, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Duke University Mathematics Series V, Duke University Press, Durham, NC, 1985.

[15] T.-T. Li, *Global Classical Solutions for Quasilinear Hyperbolic Systems*, Research in Applied Mathematics 32, P.G. Giarlet and L.-L. Lions, series eds., John Wiley & Sons, New York, 1994.

[16] R. McOwen, *Partial Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1995.

[17] M.S. Olufsen, *A one-dimensional fluid dynamic model of the systemic arteries*, in Computational Modeling in Biological Fluid Dynamics (Minneapolis, MN, 1999), IMA Vol. Math. Appl. 124, Springer, New York, 2001, pp. 167–187.

[18] M.S. Olufsen, C.S. Peskin, W.Y. Kim, E.M. Pedersen, A. Nadim, and J. Larsen, *Numerical simulation and experimental validation of blood flow in arteries with structured-tree outflow conditions*, Ann. Biomed. Eng., 28 (2000), pp. 1281–1299.

[19] J.T. Ottesen, *Valveless pumping in a fluid-filled closed elastic tube-system: One-dimensional theory with experimental validation*, J. Math. Biol., 46 (1995), pp. 309–332.

[20] G. Porenta, D.F. Young, and T.R. Rogge, *A finite-element model of blood flow in arteries including taper, branches, and obstructions*, J. Biomech. Eng., 108 (1986), pp. 161–167.

[21] A. Quarteroni, *Mathematical modelling of the cardiovascular system*, in Proceedings of the International Congress of Mathematicians, Beijing, 2002, Vol. III, Higher Education Press, Beijing, 2002, pp. 839–849.

[22] A. Quarteroni, A. Veneziani, and P. Zunino, *Mathematical and numerical modeling of solute dynamics in blood flow and arterial walls*, SIAM J. Numer. Anal., 39 (2001), pp. 1488–1511.

[23] J.K. Raines, M.Y. Jaffrin, and A.H. Shapiro, *A computer simulation of the human arterial system*, in Proceedings of the 1971 Summer Computer Conference, Vol. 2, pp. 171–178.

[24] J.K. Raines, M.Y. Jaffrin, and A.H. Shapiro, *A computer simulation of arterial dynamics in the human leg*, J. Biomechanics, 7 (1974), pp. 77–91.

[25] N.P. Smith, A.J. Pullan, and P.J. Hunter, *An anatomically based model of transient coronary blood flow in the heart*, SIAM J. Appl. Math., 62 (2002), pp. 990–1018.

[26] M. Zhao, F.T. Charbel, N. Alperin, F. Loth, and M.E. Clark, *Improved phase-contrast flow quantification by three-dimensional vessel localization*, Magn. Reson. Imaging, 18 (2000), pp. 697–706.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 668–687

A VARIATIONAL APPROACH TO NONRIGID MORPHOLOGICAL IMAGE REGISTRATION*

M. DROSKE† AND M. RUMPF†

Abstract. A variational method for nonrigid registration of multimodal image data is presented. A suitable deformation will be determined via the minimization of a morphological, i.e., contrast invariant, matching functional along with an appropriate regularization energy. The aim is to correlate the morphologies of a template and a reference image under the deformation. Mathematically, the morphology of images can be described by the entity of level sets of the image and hence by its Gauss map. A class of morphological matching functionals is presented which measure the defect of the template Gauss map in the deformed state with respect to the deformed Gauss map of the reference image. The problem is regularized by considering a nonlinear elastic regularization energy. Existence of homeomorphic, minimizing deformation is proved under assumptions on the class of admissible deformations. With respect to actual medical applications, suitable generalizations of the matching energies and the boundary conditions are presented. Concerning the robust implementation of the approach, the problem is embedded in a multiscale context. A discretization based on multilinear finite elements is discussed, and the first numerical results are presented.

Key words. image processing, image registration, mathematical morphology, nonlinear elasticity

AMS subject classifications. 49J45, 65N55, 65M60, 74B20, 68U10

DOI. 10.1137/S0036139902419528

1. Introduction.

Nowadays classical image acquisition machinery, such as computer tomography and magnetic resonance tomography, and a variety of novel sources for images, such as functional magnetic resonance imaging (MRI), three dimensional ultrasound, or densiometric computer tomography (DXA), deliver various three dimensional images of the same human body. Due to different body positioning, temporal differences of the image generation, and differences in the measurement process, the images frequently cannot simply be overlaid. Indeed, corresponding structures are situated at usually nonlinearly transformed positions. In case of intraindividual registration, the variability of the anatomy cannot be described by a rigid transformation, since many structures like, e.g., the brain cortex, may evolve very differently in the growing process. Frequently, if the image modality differs, there is also no correlation of image intensities at corresponding positions. What still remains, at least partially, is the local image structure or "morphology" of corresponding objects. Thus, the matching of two dimensional and especially three dimensional images—also known as image registration—with respect to their morphology is one of the fundamental tasks in image processing.

One aims to correlate two images—a reference image R and a template image T—via an energy relaxation over a set of, in general, nonrigid spatial deformations. Let us denote the reference image by $R : \Omega \to \mathbb{R}$ and the template image by $T : \Omega \to \mathbb{R}$. Here, both images are supposed to be defined on a bounded domain $\Omega \in \mathbb{R}^d$ for $d = 1, 2$, or 3 with Lipschitz boundary and satisfy the *cone condition* (cf., e.g., [4]). We ask

*Received by the editors December 12, 2002; accepted for publication (in revised form) May 16, 2003; published electronically January 30, 2004. This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG) SPP 1114 "Mathematical methods for time series analysis and digital image processing."

http://www.siam.org/journals/siap/64-2/41952.html

†Numerical Analysis and Scientific Computing, Lotharstr. 65, University of Duisburg, 47048 Duisburg, Germany (droske@math.uni-duisburg.de, rumpf@math.uni-duisburg.de).

for a deformation $\phi : \Omega \to \Omega$ such that $T \circ \phi$ is optimally correlated to R. There is a large and diverse body of literature on registration. In particular, Grenander, Miller, and coworkers contributed different physically motivated and mathematically profound approaches [12, 11, 24, 21, 32]. For an overview, in particular, on the mathematical modeling, see the references therein. For unimodal images one defines similarity measures, for instance, by the simple choice $\|T \circ \phi - R\|^2_{L^2}$ [15, 28, 33, 41]. In case T and R are images of different modality, we are left to define what is meant by the correlation of local structures in the image.

Viola and Wells [45], Wells et al. [47], and Collignon et al. [16] presented an information theoretic approach for registration of multimodal images. It is based on the idea of maximizing the so-called mutual information of the deformed template image and the reference image. The mutual information consists of the entropies of both images and the negative joint entropy. It can be interpreted as a measure of variability and uncertainty. Thus, the joint entropy of the images is low, where one image can stochastically be well described by the other and vice versa. Since the entropies of random variables are integrals containing the corresponding density functions, here the intensities, the corresponding local structure analysis is rather implicitly encoded in the global functionals. Viola and Wells performed the maximization process by using a stochastic descent method in which the gradients are computed via a Parzen windowing function, while Collignon et al. used Powell's method for the optimization. The method is currently restricted to an expression in global parametric form such as rigid transformations or a lower dimensional space of smooth deformations. A different approach of image registration via the matching of objects in images is due to Monasse [35]. He classifies objects by moments, and a registration is achieved by aligning these moments under scaling and rigid body motion.

Here, we introduce a different approach based on the definition of a matching energy, which effectively measures the local morphological "defect" of the deformed template and the reference image. The congruence of the shapes instead of the equality of the intensities is the main object of the registration approach presented here. First, let us define the morphology $M[I]$ of an image I as the set of level sets of I:

$$M[I] := \{\mathcal{M}^I_c \,|\, c \in \mathbb{R}\}, \tag{1.1}$$

where $\mathcal{M}^I_c := \{x \in \Omega \,|\, I(x) = c\}$ is a single level set for the gray value c. (For a general overview on image morphology, we refer to [40].) That is, $M[\gamma \circ I] = M[I]$ for any reparametrization $\gamma : \mathbb{R} \to \mathbb{R}$ of the gray values. Obviously, $M[I]$ is uniquely identified by the set of tangent spaces $\mathcal{T}_x\mathcal{M}^I_{I(x)}$ of all level sets $\mathcal{M}^I_c$ or up to the orientation by the normal field N_I on $\mathcal{M}^I_c$. Hence, again up to the orientation, the morphology $M[I]$ can be identified with the normal map (Gauss map)

$$N_I : \Omega \to \mathbb{R}^d;\ x \mapsto \frac{\nabla I}{\|\nabla I\|}\,. \tag{1.2}$$

Two images I_1 and I_2 are called morphologically equivalent if $M[I_1] = M[I_2]$. Let us emphasize that we deal here with classical level sets, which might not be defined everywhere. The problem related to vanishing image gradients and thus undefined normals will be addressed in section 4, where we allow for such singularities as long as the measure of the corresponding set in appropriate terms is not too large. A weaker definition of level sets has been introduced by Caselles, Coll, and Morel [10]. They consider the so-called upper topographic map $\{\{x \,|\, \phi(x) \geq \lambda\} \,|\, \lambda \in \mathbb{R}\}$ to characterize

the morphology of an image ϕ. This map uniquely describes the morphology, and they prove stability with respect to discretization and quantization.

Morphological methods in image processing are characterized by an invariance with respect to the morphology. Explicitly speaking, a method is called morphological if, when it is applied to morphologically identical images, the resulting images are still morphologically identical [1, 39, 43]. Hence, such methods only affect the morphology of the image, which coincides with the geometry of the level sets. Now, aiming for a morphological registration method, we will ask for a deformation $\phi : \Omega \to \Omega$ such that

$$M[T \circ \phi] = M[R] \,.$$

Thus, we try to align the normal fields (cf. Desolneux, Moisan, and Morel [19], where tangent spaces are identified in rigorous statistical terms). We set up a matching functional which locally measures the twist of the tangent spaces of the template image at the deformed position and the deformed reference image or the defect of the corresponding normal fields. See Figure 1 for an example of a registration on a pair of images for which a smooth deformation ϕ exists such that $T \circ \phi$ and R are morphologically equivalent. Figure 2 shows the known exact deformation and the deformation computed by the method we propose here.

As known from other approaches, the corresponding minimization, if settled over an infinite dimensional space of deformations and not ab initio restricted to a small finite dimensional function space, turns out to be ill posed [8, 44]. Hence, we have to ask for a suitable regularization. Various regularization approaches have been considered in the literature [11, 12, 18, 26]. On one hand, a regularization of the energy is taken into account, typically adding a convex energy functional based on gradients to the actual matching energy. The regularization energy is regarded as a penalty for "elastic stresses" resulting from the deformation of the images. This competitive approach is related to the well-known classical Tikhonov regularization of the originally ill-posed problem. On the other hand, viscous flow techniques are taken into account. They compute smooth paths from some initial deformation towards the set of minimizers of the matching energy [15, 27].

The paper is organized as follows. In section 2 the morphological matching energies are discussed, and in section 3 the regularization via nonlinear elasticity functionals will be introduced. Then, in section 4 we prove existence of homeomorphic, minimizing deformations. With respect to the actual application to medical data, the model is further generalized in sections 5 and 6, where an additional feature-based matching functional is introduced and generalized boundary conditions are discussed. Finally, in section 7 we describe the finite element discretization and the minimization algorithm.

In the present paper, we will prove the existence of a minimizing deformation for a variational approach, which is formulated for three dimensional images. It is left to the reader to transfer the assumptions and the existence results to the simpler two dimensional case. Here and in what follows we make use of the summation convention. That is, we implicitly sum over every index which appears twice in an expression.

Let us emphasize that the focus of the paper is on the presentation of a new concept in morphological image registration. Details on the implementation will be discussed in a forthcoming publication. Hence, the computational results are currently restricted to two dimensions.

2. A morphological registration energy. In this section we will construct a suitable matching energy, which measures the defect of the morphology of the refer-

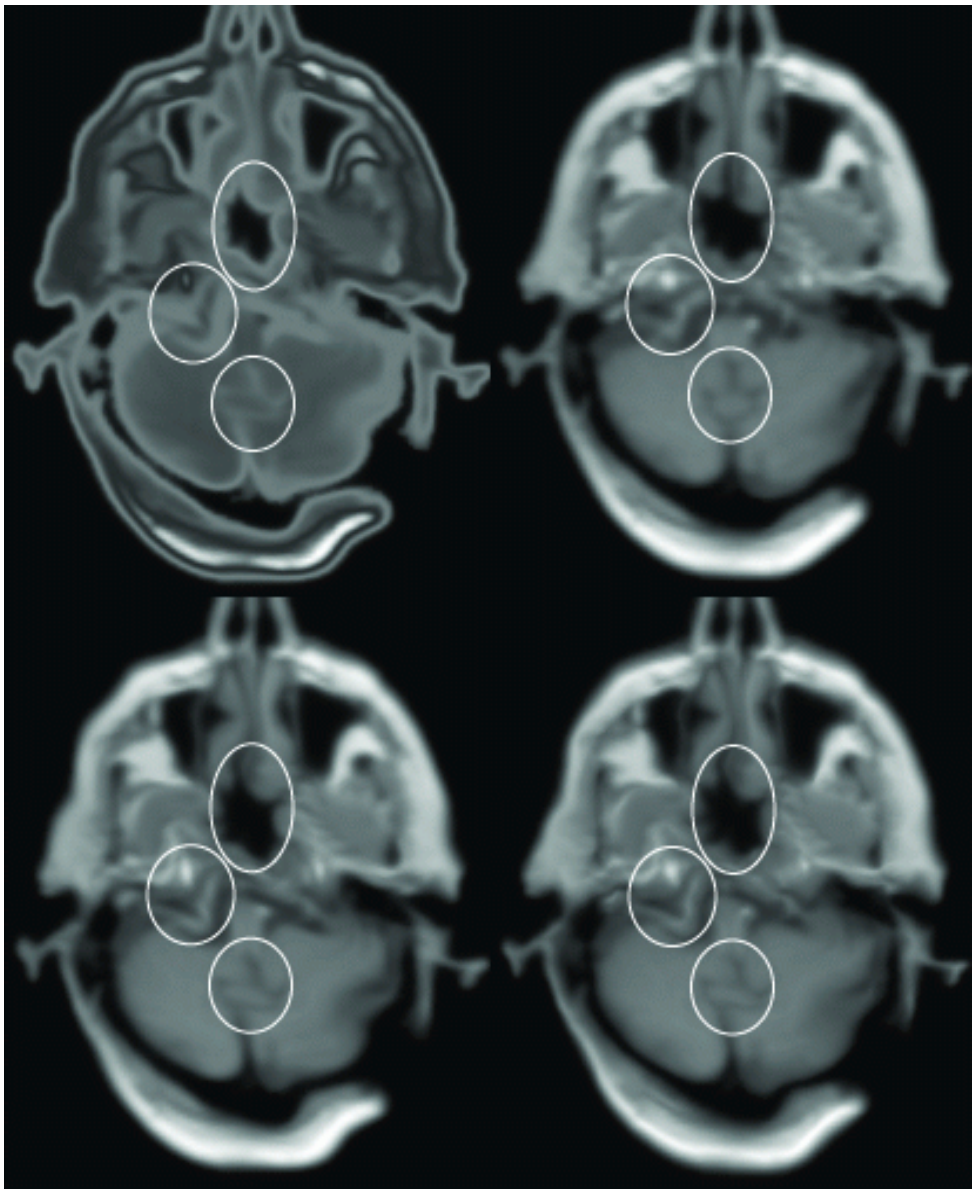


FIG. 1. *Test example. Top left: reference image* $R = \beta \circ T \circ \psi$*, generated from the template image by applying an artificial volume preserving distortion* ψ *and a nonmonotone contrast transformation* β*. Top right: template image* T*. Bottom left: reference image* $T \circ \psi$ *before contrast transformation. Bottom right: registration result* $T \circ \phi$*, template image applied to the computed deformation* ϕ*. All images have a resolution of* 257^2*. Areas of special interest are marked by white circles. See Figure* 2 *for the corresponding deformation.*

ence image R and the deformed template image T. Thus, with respect to the above identification of morphologies and normal fields, we ask for a deformation ϕ such that

$$N_T \circ \phi \,||\, N_R^{\phi}, \tag{2.1}$$

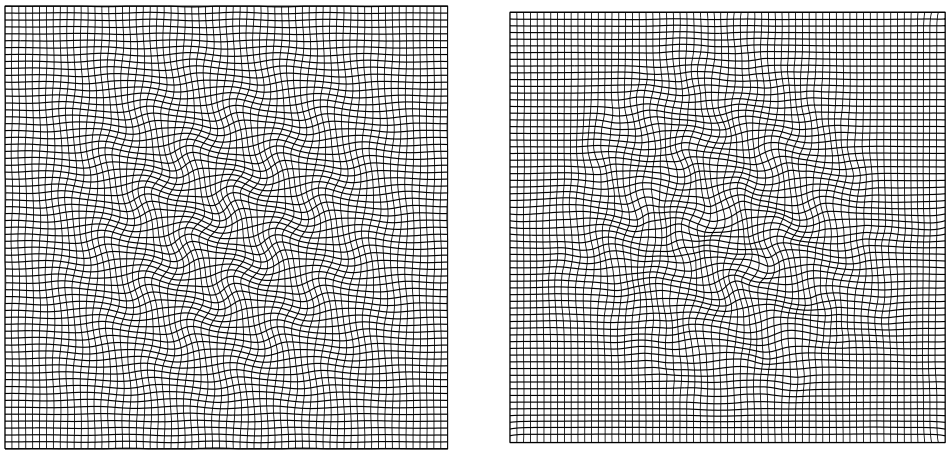


FIG. 2. *Exact deformation ψ (left) and computed deformation ϕ for the example in Figure* 1.

where N_R^ϕ is the transformed normal of the reference image R on $\mathcal{T}_{\phi(x)}\phi(\mathcal{M}^R_{R(x)})$ at position $\phi(x)$. From the transformation rule for the exterior vector product $D\phi\, u \wedge D\phi\, v = \operatorname{Cof} D\phi(u \wedge v)$ for all $v, w \in \mathcal{T}_x\mathcal{M}^R_{R(x)}$, one derives

$$N_R^\phi = \frac{\operatorname{Cof} D\phi\, N_R}{\|\operatorname{Cof} D\phi\, N_R\|},$$

where $\operatorname{Cof} A = \det A \cdot A^{-T}$ for invertible $A \in \mathbb{R}^{d,d}$. In a variational setting, optimality can be expressed in terms of energy minimization. Hence, we consider a matching energy

$$E_m[\phi] := \int_\Omega g(N_T \circ \phi, N_R, \operatorname{Cof} D\phi)\, \mathrm{d}\mu$$

for some function $g : S^{d-1} \times S^{d-1} \times \mathbb{R}^{d,d} \to \mathbb{R}^+;\ (u, v, A) \mapsto g(u, v, A)$. Here S^{d-1} denotes the unit sphere in $\mathbb{R}^d$ and μ the Lebesgue measure. This matching energy depends on the deformation of normal fields, and we are going to relax the energy via a minimizing deformation for fixed image morphologies and hence fixed normal fields. Recently, in image restoration or inpainting, energies have been introduced which depend on the normal fields of images represented by BV functions [5, 6, 7]. There, the energy is minimized over an appropriate set of BV functions on a destroyed image region.

As a boundary condition we require $\phi = \mathbb{1}$ on $\partial\Omega$, where $\mathbb{1}$ indicates the identity mapping on Ω and simultaneously the identity matrix. So far, we have assumed that the normal fields N_T and N_R are well defined on the whole domain Ω. To be not too restrictive with respect to the space of images, we have to take into account the problem of degenerate Gauss maps. Hence, let us define the set $\mathcal{D}_I := \{x \in \Omega \,|\, \nabla I = 0\}$ for $I = T$ or R, where no image normal can be defined. At first, we resolve this problem of undefined normals at least formally by introducing a 0-homogeneous extension $g_0 : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d,d} \to \mathbb{R}^+$ of g in the first and second argument:

$$g_0(v, w, A) = \begin{cases} 0, & v = 0 \text{ or } w = 0, \\ g(\frac{v}{\|v\|}, \frac{w}{\|w\|}, A), & \text{else.} \end{cases} \tag{2.2}$$

Based on g_0 we can redefine the matching energy E_m and obtain

$$E_m[\phi] := \int_\Omega g_0(\nabla T \circ \phi, \nabla R, \operatorname{Cof} D\phi)\, \mathrm{d}\mu. \tag{2.3}$$

In the later analysis we have to take special care of the singularity of g_0 for eliminating the first or second argument. Indeed, we will assume that the measure of D_T and D_R is in a suitable sense sufficiently small. Furthermore, in the existence theory we will explicitly control the impact of these sets on the energy. As a first choice for the energy density g, let us consider

$$g(v, w, A) := \left(v - \frac{Aw}{\|Aw\|} \right)^2 \tag{2.4}$$

for $v, w \in S^{d-1}$, which corresponds to the energy

$$\int_\Omega \|N_T \circ \phi - N_R^\phi\|^2 .$$

We observe that the energy E_m vanishes if $T \circ \phi = \gamma \circ R$ for a monotone gray value transformation $\gamma : \mathbb{R} \to \mathbb{R}$. If we want E_m to vanish also for nonmonotone transformations γ, we are lead to the symmetry assumption:

$$g(v, w, A) = g(-v, w, A) = g(v, -w, A) . \tag{2.5}$$

Example 2.1. A useful class of matching functionals E_m is obtained choosing functions g which depend on the scalar product $v \cdot u$ or alternatively on $(\mathbb{1} - v \otimes v)u$ (where $\mathbb{1} - v \otimes v = (\delta_{ij} - v_i\, v_j)_{ij}$ denotes the projection of u onto the plane normal to v) for $u = \frac{Aw}{\|Aw\|}$ and $v, w \in S^{d-1}$, i.e.,

$$g(v, w, A) = \hat{g}\left((\mathbb{1} - v \otimes v) \frac{Aw}{\|Aw\|} \right). \tag{2.6}$$

Let us remark that $\hat{g}((\mathbb{1} - v \otimes v)u)$ is convex in u if $\hat{g}$ is convex. With respect to arbitrary gray value transformations mapping morphologically identical images onto each other, we might consider $\hat{g}(s) = \|s\|^\gamma$ for some $\gamma \geq 1$.

3. Hyperelastic, polyconvex regularization. Suppose a minimizing deformation ϕ of E_m is given. Then, obviously for any deformation ψ which exchanges the level sets $\mathcal{M}_c^R$ of the image R, the concatenation $\psi \circ \phi$ still is a minimizer. But ψ can be arbitrarily irregular. Hence, minimizing solely the matching energy is an ill-posed problem. Thus, we consider a regularized energy

$$E[\phi] = E_m[\phi] + E_{reg}[\phi] . \tag{3.1}$$

Due to the fact that the matching energy already includes first order derivatives of the deformation ϕ, one might consider a regularization energy which involves higher order derivatives of ϕ [34]. In particular, the existence of minimizers would basically rely on usual compactness arguments. But on the background of elasticity theory, we aim to model the image domain as an elastic body responding to forces induced by the matching energy. Hence, we have to confine ourselves to energies as they appear in the usual mechanical approach to elastic bodies. It will turn out in section 4 that we have nice consistency of the type of nonlinearity in the matching energy with

respect to the Jacobian of the deformation and the well-known structure of nonlinear elastic functionals. We have to emphasize that we do not attempt to model the actual material of the objects represented by the image.

At first, let us briefly recall some background from elasticity. For details we refer to the comprehensive introductions in the books by Ciarlet [13] and Marsden and Hughes [31]. We interpret Ω as an isotropic elastic body and suppose that the regularization energy plays the role of an elastic energy while the matching energy can be regarded as an external potential contributing to the energy. Furthermore we suppose $\phi = \mathbb{1}$ to represent the stress free deformation. Let us consider the deformation of length, area, and volume under a deformation ϕ. It is well known that the norm of the Jacobian of the deformation $\|D\phi\|_2$ controls the isotropically averaged change of length under the deformation, where $\|A\|_2 := \operatorname{tr}(A^T A) = \sum_{i,j} A_{ij} A_{ij}$ for $A \in \mathbb{R}^{d,d}$. Second, the local volume transformation under a deformation ϕ is represented by $\det D\phi$. If $\det D\phi$ changes sign, self-penetration may be observed. Furthermore for $d = 3$, the norm of the matrix of the cofactors of the Jacobian of the deformation $||\operatorname{Cof} D\phi||_2 = \operatorname{tr}(\operatorname{Cof} D\phi^T \operatorname{Cof} D\phi)$ is the proper measure for the averaged change of area.

Example 3.1. Based on these considerations, we can define a simple physically reasonable isotropic elastic energy for $d = 3$, which separately cares about length, area, and volume deformation and especially penalizes volume shrinkage:

$$E_{reg}[\phi] := \int_\Omega a||D\phi||_2^p + b||\operatorname{Cof} D\phi||_2^q + \Gamma(\det D\phi)\,\mathrm{d}\mu \tag{3.2}$$

with $\Gamma(D) \to \infty$ for $D \to 0, \infty$, e.g., $\Gamma(D) = \gamma D^2 - \delta \ln D$. In nonlinear elasticity such material laws have been proposed by Ogden [38], and for $p = q = 2$ we obtain the Mooney–Rivlin model [13].

More general than in the above example, we will consider a so-called polyconvex energy functional (see [17])

$$E_{reg}[\phi] := \int_\Omega W(D\phi, \operatorname{Cof} D\phi, \det D\phi)\,\mathrm{d}\mu, \tag{3.3}$$

where $W : \mathbb{R}^{d,d} \times \mathbb{R}^{d,d} \times \mathbb{R} \to \mathbb{R}$ is supposed to be convex. Besides suitable growth conditions to be stated later, we furthermore assume that W and thereby $E_{reg}[\phi]$ again penalizes volume shrinkage, i.e., $W(A, C, D) \overset{D\to 0}{\longrightarrow} \infty$. This will enable us to successfully control singularity sets. Such energies have already been introduced to the related optical flow problem by Hinterberger et al. [29]. However, their focus was neither on morphological registration nor on the control of singularity sets.

4. An existence result. In this section we will discuss under which conditions there exists a minimizing deformation of the total energy $E[\cdot]$. Let us emphasize that the problem stated here significantly differs from most other regularized image registration problems, e.g., all intensity based approaches, where the matching energy is defined solely in terms of the deformation ϕ and the regularization energy is of higher order and considers the Jacobian $D\phi$ of the deformation. In our case already the matching energy incorporates the cofactor of the Jacobian. Thus, with respect to the direct method in the calculus of variations, we cannot use standard compactness arguments due to Rellich's embedding theorem to deal with the matching energy on a minimizing sequence [17]. Instead, we will need suitable convexity assumptions on the function g.

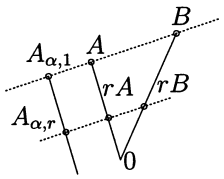


FIG. 3. *Construction as of the proof in Remark* 4.1.

Remark 4.1 (lack of lower semicontinuity for certain functionals E_m). Recalling Example 2.1, we might wish to choose a matching energy with an integrand $g_0(v,w,A) := \hat{g}((\mathbb{1} - \frac{v}{\|v\|} \otimes \frac{v}{\|v\|})\frac{Aw}{\|Aw\|})$ for $\hat{g} \in C^0(\mathbb{R}^d, \mathbb{R}^+_0)$. It is well known that the essential condition to ensure weak sequential lower semicontinuity of functionals depending on the Jacobian of a deformation is quasi-convexity [36, 37]. With our special choice of the class of energies (2.3), this requires the convexity of g in the argument A (cf. [17, sect. 5.1]).

We easily verify that a function

$$f : \mathbb{R}^{d,d} \to \mathbb{R}; A \mapsto f(A)$$

that is 0-homogeneous on $\mathbb{R}^{d,d}$ and convex has to be constant, and thus an existence result for our image matching problem via the direct calculus of variations can only be expected for trivial matching energies, i.e., for $\hat{g} \equiv \text{const}$. Indeed, suppose $A, B \in \mathbb{R}^{d,d}$ with $f(A) - f(B) = \delta > 0$ and define $A_{\alpha,r} := r\,A + \alpha\,(A - B)$ for $r > 0$ and $\alpha > 0$; then for $s = \frac{\alpha}{r}$ we obtain (cf. Figure 3)

$$\begin{aligned} f(A_{\alpha,r}) &= f(r\,A + s\,r\,(A - B)) \\ &\geq f(r\,A) + s\,(f(r\,A) - f(r\,B)) \\ &= f(A) + \frac{\alpha}{r}(f(A) - f(B)) = f(A) + \frac{\alpha\,\delta}{r} \to \infty \end{aligned}$$

for $r \to 0$. Finally, we deduce $f(A-B) = \infty$, which contradicts our assumptions on f. Thus, the definition of the matching energy via the above integrand $\hat{g}((\mathbb{1} - v \otimes w)\frac{Aw}{\|Aw\|})$ and especially our first choice of a matching energy in (2.4) are not appropriate with respect to a positive answer to the question of existence of minimizers via direct methods.

So far, we have not discussed the singularities of the normal fields. Hence, let us introduce assumptions which allow the normals to be undefined on a small set. Indeed, we take into account the space of bounded functions I, which are differentiable and whose gradients are unequal to zero on $\Omega \setminus \mathcal{D}_I$. In particular, the set of degenerate points is defined as $\mathcal{D}_I := \{x \in \Omega \,|\, \nabla I = 0\}$. We suppose that for the Lebesgue measure μ,

$$\mu(B_\epsilon(\mathcal{D}_I)) \overset{\epsilon \to 0}{\longrightarrow} 0\,,$$

where $B_\epsilon(\mathcal{D}_I) := \bigcup\{B_\epsilon(x) \,|\, x \in \mathcal{D}_I\}$. Let us introduce a corresponding set of functions

$$\mathcal{I}(\Omega) := \Big\{ I : \Omega \to \mathbb{R} \;\Big|\; I \in C^1(\bar{\Omega}), \exists\, \mathcal{D}_I \subset \Omega \text{ such that } \nabla I \neq 0 \text{ on } \Omega \setminus \mathcal{D}_I, \\ \mu(B_\epsilon(\mathcal{D}_I)) \overset{\epsilon \to 0}{\longrightarrow} 0 \Big\}\,.$$

The existence proof for minimizers of nonlinear elastic energies via the calculus of variations and direct methods dates back to the work of Ball [3]. In particular, the incorporated control of the volume transformation in this theory turns out to be the key to proving existence of minimizing, continuous, and injective deformations for the image matching problem discussed here. We consider the following energy (cf. (2.3) and (3.3)):

$$E[\phi] := E_m[\phi] + E_{reg}[\phi]\,, \tag{4.1}$$

with

$$E_{reg}[\phi] := \int_\Omega W(D\phi, \operatorname{Cof} D\phi, \det D\phi)\ \mathrm{d}\mu,$$
$$E_m[\phi] := \int_\Omega g_0(\nabla T \circ \phi, \nabla R, \operatorname{Cof} D\phi)\,\mathrm{d}\mu$$

for g_0 defined in (2.2).

Let us denote by L^p for $p \in [1,\infty]$ the usual Lebesgue spaces of functions on Ω into $\mathbb{R}$, $\mathbb{R}^d$, and $\mathbb{R}^{d,d}$, respectively; by $\|\cdot\|_p$, the corresponding norm; and by $H^{1,p}$, the Banach space of functions in L^p with weak first derivatives also in L^p. For ease of presentation, we do not exploit the full generality of the corresponding existence theory. Here the reader is, for instance, referred to [3, 4, 14, 22, 23, 46]. We confine ourselves to a basic model and state the following theorem.

THEOREM 4.2 (existence of minimizing deformations). *Suppose $d = 3$ and $T, R \in \mathcal{I}(\Omega)$ and consider the total energy defined in* (4.1) *for deformations ϕ in the set of admissible deformations*

$$\mathcal{A} := \{\phi : \Omega \to \Omega \mid \phi \in H^{1,p}(\Omega), \operatorname{Cof} D\phi \in L^q(\Omega),$$
$$\det D\phi \in L^r(\Omega), \det D\phi > 0 \text{ a.e. in } \Omega, \phi = \mathbb{1} \text{ on } \partial\Omega\},$$

where p, $q > 3$ and $r > 1$. Suppose $W : \mathbb{R}^{3,3} \times \mathbb{R}^{3,3} \times \mathbb{R}^+ \to \mathbb{R}$ is convex and there exist constants $\beta, s \in \mathbb{R}$, $\beta > 0$, and $s > \frac{2q}{q-3}$ such that

$$W(A,C,D) \geq \beta\left(\|A\|_2^p + \|C\|_2^q + D^r + D^{-s}\right) \quad \forall A, C \in \mathbb{R}^{3,3},\ D \in \mathbb{R}^+. \tag{4.2}$$

Furthermore, assume that $g_0(v,w,A) = g(\frac{v}{\|v\|}, \frac{w}{\|w\|}, A)$ for some function $g : S^2 \times S^2 \times \mathbb{R}^{3,3} \to \mathbb{R}_0^+$, which is continuous in $\frac{v}{\|v\|}$, $\frac{w}{\|w\|}$ and convex in A, and for a constant $m < q$ the estimate

$$g(v,w,A) - g(u,w,A) \leq C_g\ \|v-u\|\ (1 + \|A\|_2^m)$$

holds for all $u, v, w \in S^2$ and $A \in \mathbb{R}^{3,3}$. Then $E[\cdot]$ attains its minimum over all deformations $\phi \in \mathcal{A}$ and the minimizing deformation ϕ is a homeomorphism and, in particular, $\det D\phi > 0$ *a.e. in* Ω.

Proof. The proof of this result is based on the well-known weak continuity results for the principle invariants of the Jacobian of the deformation. We observe that the total energy is polyconvex. Furthermore, the volume of the neighborhood sets $B_\epsilon(\mathcal{D}_T)$ and $B_\epsilon(\mathcal{D}_R)$ of the singularity sets $\mathcal{D}_T$ and $\mathcal{D}_R$, respectively, can be controlled. Hence, we can basically confine ourselves to a set where the integrand fulfills Carathéodory's conditions. At first, let us recall some well-known, fundamental weak convergence results: Given a sequence of deformations $(\phi^k)_k$ in $H^{1,p}$, with $\operatorname{Cof} D\phi^k \in L^q$ and $\det D\phi^k \in L^r$, such that the sequence converges weakly in the sense $\phi^k \rightharpoonup \phi$ in $H^{1,p}$,$\operatorname{Cof} D\phi^k \rightharpoonup C$ in L^q, and $\det D\phi^k \rightharpoonup D$ in L^r, then $C = \operatorname{Cof} D\phi$ and $D = \det D\phi$

(weak continuity). For the proof we refer to Ball [3] or the book of Ciarlet [13, sect. 7.5, 7.6].

The proof of the theorem proceeds in four steps.

Step 1. Due to the assumption on the image set $\mathcal{I}(\Omega)$, $E_m[\phi]$ is well defined for $\phi \in \mathcal{A}$. In particular, $g_0(\nabla T \circ \phi, \nabla R, \operatorname{Cof} D\phi)$ is measurable. Obviously $\mathbb{1} \in \mathcal{A}$ and $E[\mathbb{1}] < \infty$; thus $\underline{E} := \inf_{\phi\in\mathcal{A}} E[\phi] < \infty$, and due to the growth conditions and the assumption of g we furthermore get $\underline{E} \geq 0$. Let us consider a minimizing sequence $(\phi^k)_{k=0,1,\ldots} \subset \mathcal{A}$ with $E[\phi^k] \to \inf_{\phi\in\mathcal{A}} E[\phi]$. We denote by $\overline{E}$ an upper bound of the energy E on this sequence. Due to the growth condition on W we get that $\{(D\phi^k, \operatorname{Cof} D\phi^k, \det D\phi^k)\}_{k=0,1,\ldots}$ is uniformly bounded in $L^p(\Omega) \times L^q(\Omega) \times L^r(\Omega)$. By Poincaré's inequality applied to $(\phi^k - \mathbb{1})$ we obtain that $\{\phi^k\}_{k=0,1,\ldots}$ is uniformly bounded in $H^{1,p}(\Omega)$. Because of the reflexivity of $L^p \times L^q \times L^r$ for p, q, $r > 1$, we can extract a weakly convergent subsequence, again denoted by an index k, such that

$$(D\phi^k, \operatorname{Cof} D\phi^k, \det D\phi^k) \rightharpoonup (D\phi, C, D)$$

in $L^p \times L^q \times L^r$ with $C : \Omega \to \mathbb{R}^{3\times 3}$, $D : \Omega \to \mathbb{R}$. Applying the above results on weak convergence, we achieve $C = \operatorname{Cof} D\phi$ and $D = \det D\phi$. In addition, by Rellich's embedding theorem we know that $\phi^k \to \phi$ strongly in $L^p(\Omega)$, and by Sobolev's embedding theorem we obtain $\phi \in C^0(\bar{\Omega})$.

Step 2. Next, we control the set where the volume shrinks by a factor of more than ϵ for the limit deformation. Let us define

$$S_\epsilon = \{x \in \Omega \,|\, \det D\phi \leq \epsilon\}$$

for $\epsilon \geq 0$. Let as assume without loss of generality that the sequence of energy values $E[\phi^k]$ is monotone decreasing and that for given $\epsilon > 0$ we denote by $k(\epsilon)$ the smallest index such that

$$E[\phi^k] \leq E[\phi^{k(\epsilon)}] \leq \underline{E} + \epsilon \qquad \forall k \geq k(\epsilon)\,.$$

From Step 1 we know that $\Psi^k := (D\phi^k, \operatorname{Cof} D\phi^k, \det D\phi^k)$ converges weakly to $\Psi := (D\phi, \operatorname{Cof} D\phi, \det D\phi)$ in $L^p \times L^q \times L^r$. Hence, applying Mazur's lemma, we obtain a sequence of convex combinations of Ψ^k and ϕ^k which converges strongly to Ψ and ϕ in $L^p \times L^q \times L^r \times L^p$. Thus, there exists a family of weights $((\lambda_i^k)_{k(\epsilon)\leq i\leq k})_{k\geq k(\epsilon)}$ with $\lambda_i^k \geq 0$, $\sum_{k(\epsilon)}^k \lambda_i^k = 1$, such that

$$\lambda_i^k \Psi^i \to \Psi \;\text{ and }\; \lambda_i^k \phi^i \to \phi\,.$$

Now, taking into account the growth conditions, the convexity of W, and Fatou's lemma, we estimate

$$\begin{aligned}
\beta\epsilon^{-s}\mu(S_\epsilon) &\leq \beta \int_{S_\epsilon} (\det D\phi)^{-s}\,\mathrm{d}\mu \leq \int_{S_\epsilon} W(\Psi)\,\mathrm{d}\mu \\
&= \int_{S_\epsilon} \liminf_{k\to\infty} W(\lambda_i^k \Psi^i)\,\mathrm{d}\mu \leq \int_{S_\epsilon} \liminf_{k\to\infty} \lambda_i^k W(\Psi^i)\,\mathrm{d}\mu \\
&\leq \liminf_{k\to\infty} \lambda_i^k \int_{S_\epsilon} W(\Psi^i)\,\mathrm{d}\mu \\
&\leq \liminf_{k\to\infty} \lambda_i^k \int_{\Omega} W(\Psi^i) + g_0(\nabla T \circ \phi^i, \nabla R, \operatorname{Cof} D\phi^i)\,\mathrm{d}\mu \\
&\leq \overline{E}
\end{aligned}$$

and claim $\mu(S_\epsilon) \leq \frac{\bar{E}\epsilon^s}{\beta}$. As one consequence, S_0 is a null set and we know that $\det D\phi > 0$ a.e. on Ω. Thus, taking this together with the results from Step 1, we de-

duce that the limit deformation ϕ is in the set of admissible deformation $\mathcal{A}$. Following Ball [4], we furthermore obtain that ϕ is injective and ϕ is a homeomorphism.

Step 3. Now, we deal with the singularity set of the images T. By our assumption on the image set $\mathcal{I}(\Omega)$, we know that for given $\delta > 0$ there exist $\epsilon_T > 0$ such that $\mu(B_{\epsilon_T}(\mathcal{D}_T)) \leq \delta$. From this and the injectivity (cf. Theorem 1(ii) in [4]), we especially deduce the estimate

$$\begin{aligned}\mu\left(\phi^{-1}(B_{\epsilon_T}(\mathcal{D}_T)) \setminus S_\epsilon\right) &\leq \frac{1}{\epsilon}\int_{\phi^{-1}(B_{\epsilon_T}(\mathcal{D}_T))} \det D\phi \,\mathrm{d}\mu \\ &= \frac{1}{\epsilon}\int_{B_{\epsilon_T}(\mathcal{D}_T)} \mathrm{d}\mu \leq \frac{\delta}{\epsilon}\,.\end{aligned}$$

Hence, we can control the preimage of $B_\epsilon(\mathcal{D}_T)$ with respect to ϕ but restricted to $\Omega \setminus S_\epsilon$. Due to the continuous differentiability of both images T and R, we can assume that

$$\|\nabla T(x)\| \geq \gamma(\epsilon) \text{ on } \Omega \setminus B_\epsilon(\mathcal{D}_T), \tag{4.3}$$

where $\gamma : \mathbb{R}_0^+ \to \mathbb{R}$ is a strictly monotone function with $\gamma(0) = 0$.

Step 4. Due to Egorov's theorem and the strong convergence of ϕ^k in $L^p(\Omega)$, there is a set K_ϵ with $\mu(K_\epsilon) < \epsilon$ such that a subsequence, again denoted by ϕ^k, converges uniformly on $\Omega \setminus K_\epsilon$. Let us now define the set

$$R_{\epsilon,\delta} := \phi^{-1}(B_{\epsilon_T}(\mathcal{D}_T)) \cup S_\epsilon \cup K_\epsilon\,,$$

whose measure can be estimated in terms of ϵ and δ, i.e.,

$$\mu(R_{\epsilon,\delta}) \leq \frac{\delta}{\epsilon} + \frac{\bar{E}\epsilon^s}{\beta} + \epsilon\,.$$

On $\Omega \setminus R_{\epsilon,\delta}$ the sequence $(\nabla T \circ \phi^k)_{k=0,1,\dots}$ converges uniformly to $\nabla T \circ \phi$. Next, from the assumption on g and the 0-homogeneous extension property of g_0, we deduce that

$$|g_0(u,w,A) - g_0(v,w,A)| \leq C_\gamma \,\|u-v\|\,(1 + \|A\|_2^m) \tag{4.4}$$

for $u, v, w \in \mathbb{R}^3$, $A \in \mathbb{R}^{3,3}$, and $\|u\|\,, \|v\|\,, \|w\| \geq \gamma$. To use this estimate for $u = \phi^k$ and $v = \phi$ below, we assume that $k(\epsilon)$ is large enough such that $\phi^k(x) \in \Omega \setminus B_{\frac{\epsilon_T}{2}}(\mathcal{D}_T)$ for $x \in \Omega \setminus R_{\epsilon,\delta}$ and

$$C_{\gamma(\frac{\epsilon_T}{2})} \left\|\nabla T \circ \phi^k - \nabla T \circ \phi\right\|_{\infty,\Omega\setminus K_\epsilon} \leq \epsilon$$

for all $k \geq k(\epsilon)$. Now we are able to estimate $E[\phi]$ using especially the convexity of W and $g(v,w,\cdot)$, the estimate (4.4), and Fatou's lemma:

$$\begin{aligned}E[\phi] &= \int_\Omega W(\Psi) + g_0(\nabla T \circ \phi, \nabla R, \operatorname{Cof} D\phi)\,\mathrm{d}\mu \\ &\leq \int_\Omega \liminf_{k\to\infty} \lambda_i^k W(\Psi^i)\,\mathrm{d}\mu + 2\,C_g \int_{R_{\epsilon,\delta}} 1 + \|\operatorname{Cof} D\phi\|^m \,\mathrm{d}\mu \\ &\quad + \int_{\Omega\setminus R_{\epsilon,\delta}} \liminf_{k\to\infty} \lambda_i^k g_0(\nabla T \circ \phi, \nabla R, \operatorname{Cof} D\phi^i)\,\mathrm{d}\mu \\ &\leq \liminf_{k\to\infty} \lambda_i^k \int_\Omega W(\Psi^i)\,\mathrm{d}\mu + b(\mu(R_{\epsilon,\delta})) \\ &\quad + \liminf_{k\to\infty} \lambda_i^k \int_{\Omega\setminus R_{\epsilon,\delta}} g_0(\nabla T \circ \phi, \nabla R, \operatorname{Cof} D\phi^i) - g_0(\nabla T \circ \phi^i, \nabla R, \operatorname{Cof} D\phi^i) \\ &\qquad\qquad + g_0(\nabla T \circ \phi^i, \nabla R, \operatorname{Cof} D\phi^i)\,\mathrm{d}\mu,\end{aligned}$$

where $b(s) := 2\,C_g(s + (\frac{\bar{E}}{\beta})^{\frac{m}{q}} s^{1-\frac{m}{q}})$. Here we have, in particular, used the a priori estimate $\|\mathrm{Cof}\, D\phi\|_{q,\Omega} \le (\frac{\bar{E}}{\beta})^{\frac{1}{q}}$. We estimate further and obtain

$$\begin{aligned} E[\phi] &\le \liminf_{k\to\infty} \lambda_i^k \int_\Omega W(\Psi^i) + g_0(\nabla T \circ \phi^i, \nabla R, \mathrm{Cof}\, D\phi^i)\,\mathrm{d}\mu + 2\,b(\mu(R_{\epsilon,\delta})) \\ &\quad + C_{\gamma(\frac{\epsilon_T}{2})} \limsup_{k\to\infty} \int_{\Omega\setminus R_{\epsilon,\delta}} \left\|\nabla T\circ\phi - \nabla T\circ\phi^k\right\| \left(1 + \left\|\mathrm{Cof}\, D\phi^k\right\|_2^m\right) \mathrm{d}\mu \\ &\le \liminf_{k\to\infty} \lambda_i^k E[\phi^i] + 2\,b(\mu(R_{\epsilon,\delta})) + \epsilon\, b(\mu(\Omega)) \\ &\le \underline{E} + \epsilon + 2\,b(\mu(R_{\epsilon,\delta})) + \epsilon\, b(\mu(\Omega))\,. \end{aligned}$$

Finally, for given $\bar{\epsilon}$ we choose ϵ, then δ, the dependent ϵ_T small enough, and $k(\bar{\epsilon})$ large enough to ensure that

$$\epsilon + 2\,b(\mu(R_{\epsilon,\delta})) + \epsilon\, b(\mu(\Omega)) \le \bar{\epsilon}$$

and get $E[\phi] \le \underline{E} + \bar{\epsilon}$. This holds true for an arbitrary choice of $\bar{\epsilon}$. Thus we conclude

$$E[\phi] \le \underline{E} = \inf_{\phi\in\mathcal{A}} E[\phi]\,,$$

which is the desired result. □

Remark 4.3. From the proof we have seen that the assumptions on the reference image could be weakened considerably compared to the template image. With respect to the applications, we do not detail this difference here.

Example 4.4. Let us consider

$$g(v, w, A) = \|(\mathbb{1} - v\otimes v)\cdot Aw\|^\gamma \tag{4.5}$$

for $1 \le \gamma < q$. Hence, we obtain an admissible matching energy

$$E_m[\phi] = \int_\Omega \|(\mathbb{1} - (N_T\circ\phi)\otimes(N_T\circ\phi))\cdot \mathrm{Cof}\, D\phi\, N_R\|^\gamma$$

(cf. Example 2.1). Applying Theorem 4.2, we can establish the existence of a minimizing deformation. Recalling Remark 4.1, we recognize that scaling the original energy density by an additional factor $\|\mathrm{Cof}\, D\phi N_R\|^\gamma$ turns the minimization task into a feasible problem. This corresponds to a modification of the area element on the level sets $\mathcal{M}_c^R$. Indeed, $\|\mathrm{Cof}\, D\phi N_R\|$ is the change of the area element at a position x on $\mathcal{M}^R_{R(x)}$ under the deformation.

5. An additional feature-based registration energy. As the energy $E_m[\phi]$ depends on the directions of the image normals only, its minimization will lead to an alignment of the level sets of the two images. However, the alignment of significant level sets, which correspond to significant features, is not taken into account by the energy. In medical applications, such features may be boundaries of organs, bones, or tissue structures. Hence, we will incorporate an additional energy which measures the quality of the match of certain clearly detectable features. Suppose $\mathcal{F}_T$ and $\mathcal{F}_R$ are corresponding selected feature sets in the images T and R. These feature sets may be computed in a previous segmentation step applying, for instance, an active contour algorithm [9, 20] (cf. Figure 4 for an example of a pair of corresponding feature sets).

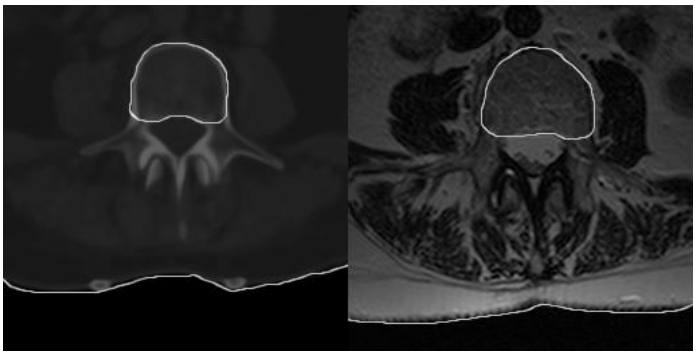


FIG. 4. *Feature sets $\mathcal{F}_R$ and $\mathcal{F}_T$ superimposed on darkened corresponding images for better visibility (cf. Figure* 6 *for registration results).*

We are aiming to penalize a nonproper match of these two sets by a suitable energy. They would be ideally matched if

$$\mathcal{F}_T = \phi(\mathcal{F}_R).$$

The following energy measures the matching quality for a general deformation ϕ:

$$E_f[\phi] = \int_\Omega |d(\phi(\cdot), \mathcal{F}_T) - d(\cdot, \mathcal{F}_R)|^2 \, \mathrm{d}\mu \,, \tag{5.1}$$

where $d(x, A) := \hat{d} \circ \mathrm{dist}(x, A)$ is a function $\hat{d}$ of the distance of a point x from a set $A \subset \Omega$. We suppose $\hat{d} : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ to be monotone and $\hat{d}(0) = 0$. In particular, E_f vanishes in case of a perfect match. A suitable choice is $\hat{d}(s) = \alpha s^\delta$ with $0 < \delta \leq 1$ and $\alpha > 0$. We use this energy as a third term in the regularized problem (3.1).

COROLLARY 5.1 (existence of minimizers in the presence of a feature matching energy). *Suppose the assumptions of Theorem* 4.2 *hold. Furthermore, let $\hat{d} : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ be continuous, and consider*

$$E[\phi] = E_m[\phi] + E_{reg}[\phi] + E_f[\phi] \,. \tag{5.2}$$

Then $E[\cdot]$ attains its minimum over all deformations $\phi \in \mathcal{A}$, the minimizing deformation ϕ is a homeomorphism, and, furthermore, $\det D\phi > 0$ *a.e. in* Ω.

Proof. Due to the Lipschitz continuity of $\mathrm{dist}(\cdot, A)$ for arbitrary sets $A \subset \Omega$ with $A \neq \emptyset$ and the continuity of $\hat{d}$, the integrand of E_f is uniformly continuous in ϕ. Hence, the proof of Theorem 4.2 can easily be generalized. ∎

The overall energy will therefore not only align the directions correctly but also penalize displaced features. In this setting it is thus possible to incorporate some a priori knowledge to improve the matching results. Let us emphasize that the morphological registration provides good results if the morphologies encoded by the normal fields of the two images actually coincide up to a deformation. But in cases where the images of different modalities reveal similar but different geometrical structures, which are not strictly equivalent in terms of mathematical morphology, a weak form of "landmarks" is recommendable to support the matching.

6. Generalized boundary conditions. So far we have imposed boundary conditions of Dirichlet type on $\partial\Omega$ for the deformation. This might be a reasonable assumption in the case of objects located in the center of the image at a considerable

distance from the boundary (cf. Figure 1). If the objects cover the whole image domain, we cannot assume that the requested deformation obeys these artificial boundary conditions. In fact, structures visible close to the boundary in the reference image R will not be present in the template image T and vice versa. Hence, we ask for more general boundary conditions. With these applications in mind, we have to tolerate deformations $\phi(\Omega) \not\subset \Omega$ in the admissible set of deformations. But the integrand of the matching energy is defined only on $\phi^{-1}(\mathrm{Im}\,(\phi) \cap \Omega)$. Hence, we replace $\int_\Omega g_0(\cdot)$ by $\int_{\Omega^\phi} g_0(\cdot)$, where $\Omega^\phi := \{x \in \Omega \,|\, \phi(x) \in \Omega\}$, and obtain the new matching energy

$$\tilde{E}_m[\phi] := \int_{\Omega^\phi} g_0(\nabla T \circ \phi, \nabla R, \mathrm{Cof}\, D\phi)\, \mathrm{d}\mu. \tag{6.1}$$

Taking into account this reformulated matching energy, we are basically facing two problems:

(i) Considering a total energy $E[\phi] = \tilde{E}_m[\phi] + E_{reg}[\phi]$, we are lead to irrelevant, trivial solutions. Indeed, taking into account a simple translation ϕ_{trans} with $\phi_{trans}(\Omega) \cap \Omega = \emptyset$, one obtains $\tilde{E}_m[\phi_{trans}] = 0$. Hence, we no longer measure the matching of relevant image features. We propose to avoid this problem by incorporating the above feature energy $E_f[\phi]$, which can be regarded as a weak boundary condition. Indeed, if $\alpha \to \infty$, we enforce an interior boundary condition on the feature sets, i.e., $\phi(\mathcal{F}_R) = \mathcal{F}_T$.

(ii) Injectivity can no longer be expected for a minimizing deformation. It might happen that parts of the domain Ω fold over each other under a deformation ϕ, although ϕ is locally injective, i.e., $\det D\phi > 0$ (cf. the exposition of this problem in [13, sect. 7.9]). Following Ciarlet and Necas [14], we introduce an additional condition on the set of admissible deformations:

$$\int_\Omega \det D\phi \le \mu(\phi(\Omega))\,.$$

Then, we expect the minimizer of the energy $E = \tilde{E}_m + E_{reg} + E_f$ to be injective on Ω, whereas on $\partial\Omega$ we might observe self-contact. In the actual applications considered so far we have not detected any lack of global injectivity due to overlapping parts of the deformed domain. Hence, there was no need to incorporate this nonlinear contact condition in the algorithm.

7. Multiscale minimization and discretization. The total energy is highly nonlinear. Especially the matching energy E_m with the nonlinearity $\nabla T \circ \phi$ depending on the complexity of image data will usually lead to multiple at-least-local minima. Hence, in order to ensure a robust and efficient minimization, we have to consider a global minimization strategy, which is capable of computing large deformations which minimize the total registration energy. Here we propose a continuous annealing method based on a scale of registration problems

$$\tilde{E}^\sigma[\phi] := \tilde{E}^\sigma_m[\phi] + E_{reg}[\phi] + E_f[\phi]\,,$$

where $\sigma > 0$ is the scale parameter. This enables us to compute global instead of only local deformations and usually avoid a tedious preregistration step. The definition of the energy scales for the matching energy is based on a scale space approach for the underlying images (cf. [2]). We choose

$$E^\sigma_m[\phi] := \int_{\Omega^\phi} g_0(\nabla T^\sigma \circ \phi, \nabla R^\sigma, \mathrm{Cof}\, D\phi)\, \mathrm{d}\mu\,,$$

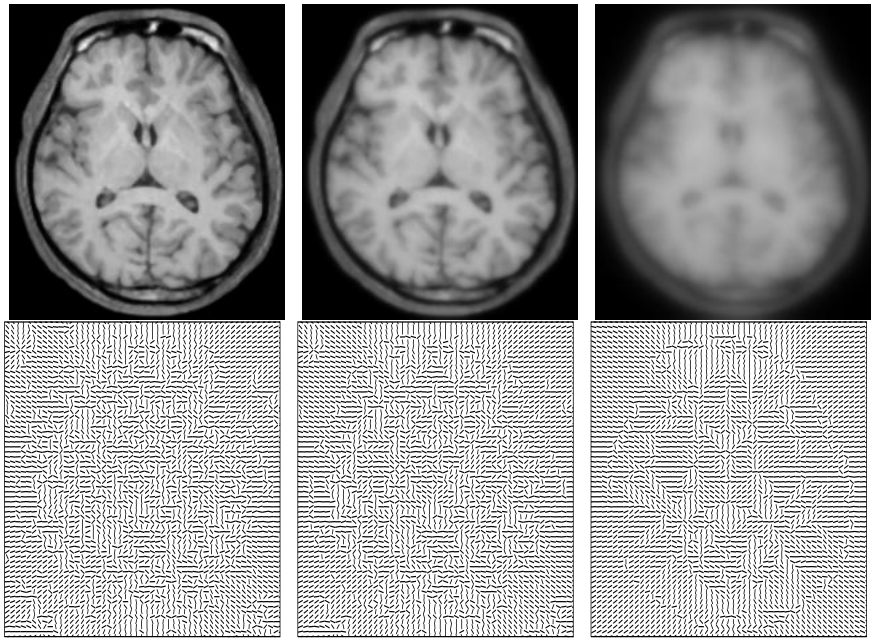


FIG. 5. *Image and the Gauss map and the corresponding grid for the brain slice at scales* $\sigma = h, 2h, 8h$.

where $I^\sigma := G^\sigma[I]$ for $I = T, R$ and G^σ denotes the convolution with a "Gaussian" filter of width σ (cf. Figure 5 for a multiscale of images and the corresponding effect on the Gauss maps N_T^σ and N_R^σ). In fact, we consider the heat equation semigroup and set $I^\sigma := u(\sigma^2/2)$, where u is the solution of the initial boundary value problem

$$\begin{aligned} \partial_t u - \Delta u &= 0 \qquad \text{in } \mathbb{R}^+ \times \Omega\,, \\ \partial_\nu u(t,\cdot) &= 0 \qquad \text{on } \mathbb{R}^+ \times \partial\Omega\,, \\ u(0,\cdot) &= I \qquad \text{in } \Omega\,, \end{aligned} \tag{7.1}$$

and ν denotes the outer normal on $\partial\Omega$. Concerning the spatial discretization, we deal with images as piecewise bilinear, continuous functions on a regular quadrilateral grid. We use the same discrete function space to define discrete nonrigid deformations. Energy functionals and their gradients are numerically evaluated using a midpoint quadrature rule on the grid cells. We assume $\Omega = [0,1]^2$ and start with an initial coarse mesh $\mathcal{M}_0 = \{\Omega\}$, which is iteratively refined by uniform subdivision, where each element is divided into four squares. This refinement process generates sequences of nested meshes $\mathcal{M}_l$, with $0 \le l \le l_{\max}$, consisting of quadrilateral elements E_l^i $(0 \le i < 4^l)$ of edge length $h_l = 2^{-l}$. The set of vertices of $\mathcal{M}_l$ is denoted by $\mathcal{N}_l$. Let V_l be the corresponding space of piecewise bilinear, continuous finite element functions. Suppose $\{\Psi_l^i\}_{i \le (2^l+1)^2}$ to be the nodal basis of V_l. The discrete gradient $\operatorname{grad}_{V_l} \tilde{E}^\sigma \in V_l^2$ of E on grid level l for a deformation $\Phi \in V_l^2$ is then defined by

$$(\operatorname{grad}_{V_l} \tilde{E}^\sigma[\Phi], \Psi_l^j e_k)_h = \langle (\tilde{E}^\sigma)'[\Phi], \Psi_l^j e_k \rangle$$

TABLE 7.1

To obtain a stable descent in the gradient descent algorithm of the global energy E^σ, the derivative in the direction of the descent direction d, i.e., $(d, gradE^\sigma[\phi])_h$, ought to be ≤ 0 in the scalar product. We have shown the impact of the smoothing parameter α for different scales on $\gamma(-d, gradE^\sigma[\phi])$, where $\gamma(u,v) := \frac{u}{\|u\|} \cdot \frac{v}{\|v\|}$. These values have been determined considering the first 50 steps of the gradient descent of the test example. We list the smallest value $\gamma_{\min}$ and the average value γ_{average}.

Scale $\alpha[h_{l_{\max}}]$	0.25	0.5	1.0	2.0	3.0	4.0	5.0
$\gamma_{\min}$	0.9954	0.9585	0.8265	0.6588	0.58481	0.5438	0.5171
γ_{average}	0.9951	0.9556	0.8177	0.6447	0.5586	0.5116	0.4825

for all $j \leq (2^l+1)^2$ and $k = 1, 2$. Here $(\cdot,\cdot)_h$ denotes the usual lumped mass product on V_l and e_k the canonical basis in $\mathbb{R}^2$. Then, on level l the necessary condition for $\Phi^l \in V_l^2$ to be a minimizer of $\tilde{E}^\sigma$ over V_l is given by

$$\operatorname{grad}_{V_l} E^\sigma[\Phi] = 0 \quad \text{for } \Phi \in V_l.$$

Now, we introduce multiple discrete scales. Therefore, we replace the filter G^σ by its discrete counterpart, replacing problem (7.1) by a single implicit Euler time step with time step size $\frac{\sigma^2}{2}$ for a usual finite element discretization with lumped masses (cf. [42]). We denote the corresponding solution operator on the finite element space V_l by $G_l^\sigma : V_l \to V_l$. On each scale, we apply a gradient descent algorithm to minimize the energy. Here we might consider a sequence of scales

$$\sigma^k = 2^{-k}\sigma_0$$

for $k = 0, \ldots, n$. Obviously, solving a coarse scale minimization process on a fine grid introduces a serious amount of redundancy. It is much more efficient to perform such computations also on coarse grid levels. Thus, we introduce a function l_k which selects for each scale an appropriate grid level. In particular, we choose

$$l_k := \min\{l = 0, \ldots, l_{\max} \,|\, h(l) \leq \gamma\sigma_k\}$$

for a scalar $\gamma > 0$, e. g., $\gamma = 1$, which controls the ratio of the cell size $h(l)$ with respect to a filter width σ_k. On each scale we compute the minimum Φ^k of E^{σ_k} over $V_{l_k}^2$ by a gradient descent method and consider the standard prolongation of $\Phi^{k-1} \in V_{l_{k-1}}$ onto V_{l_k} as the initial value if $l_k \neq l_{k-1}$. It turns out to be suitable to regularize the contribution of the matching energy to the descent direction. Hence, a descent direction $d \in V_{l_k}$ at a position $\Phi \in V_{l_k}^2$ is computed by

$$d := -G_{l_k}^\alpha[\operatorname{grad}_{V_{l_k}} E_m^{\sigma_k}[\Phi]] - \operatorname{grad}_{V_{l_k}} \tilde{E}_{reg}[\Phi] - \operatorname{grad}_{V_{l_k}} E_f[\Phi],$$

where $\alpha > 0$ controls the amount of smoothing of the gradient for the registration energy. We have to ensure

$$(d, \operatorname{grad}\tilde{E}^{\sigma_k}[\Phi])_h \leq 0$$

in order to observe stable descent (cf. Table 7.1).

As step size control we consider Armijo's rule [30]. Let us remark that the smoothing by Gaussian convolution is solved efficiently and independently of the filter width σ

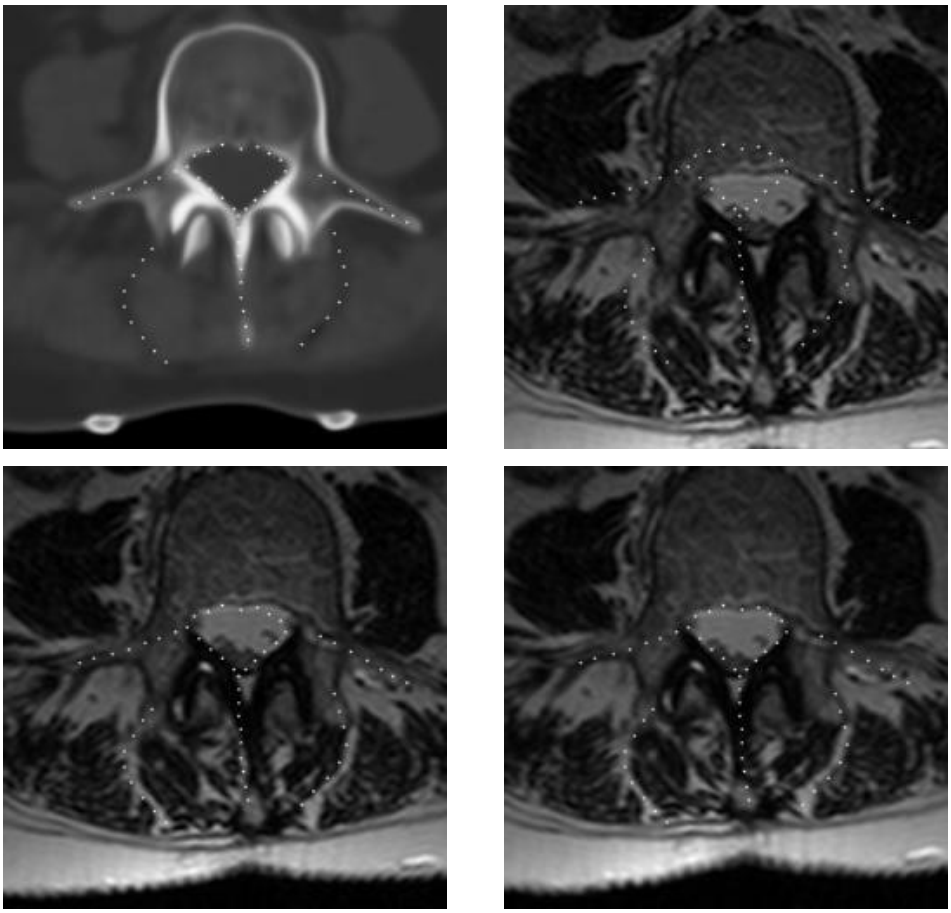


FIG. 6. *Sectional morphological registration on a pair of MR and CT images of a human spine. Dotted lines mark certain features visible in the reference image. They are repeatedly drawn at the same position in the other images. Top left: reference, CT; top right: template, MR, with clearly visible misfit of structures marked by the dotted lines. Bottom left: deformed template after feature-based registration $T \circ \phi_f$, where ϕ_f is the result of a feature-based preregistration (cf. Figure 4 for the feature sets used in this example). Bottom right: deformed template $T \circ \phi$ after final registration, where the dotted feature lines nicely coincide with the same features in the deformed template MR image. All images have a resolution of* 257^2.

by a multigrid solver for the heat equation [25]. In the computation for the registration of real magnetic resonance (MR) and computed tomography (CT) images of a human spine (cf. Figures 6, 7, 8), we chose the parameter α to be $5h_{l_{\max}}$. Furthermore, concerning the elastic regularization that we so far held on to, the Mooney–Rivlin energy, i.e., $p = 2$, q does not have to be specified since the second term of W is redundant in two dimensions. The choices for the further parameters are $a = 0.45$, $\gamma = \frac{1}{2}$, $\delta = 1$. To improve the method's performance we first relax the feature-based energy $E_f[\cdot] + E_{reg}[\cdot]$ to identify an appropriate initial deformation. Then, we continue with the minimization of the global energy $E[\cdot]$.

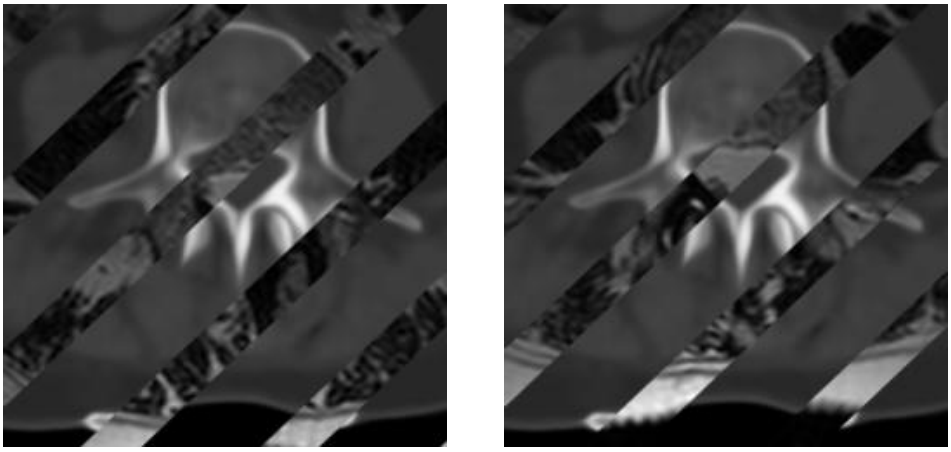


FIG. 7. *Comparison of superimposed template and reference before (left) and after (right) registration.*

Finally, the minimization algorithm can be written in pseudo code as follows.

ALGORITHM 7.1 (multiscale minimization algorithm).

$\Phi_0 := \mathbb{1}$
foreach $k = 0, \dots, n$ **do**
  set level to l_k and grid $\mathcal{M}_{l_k}$
  if $k > 0$ **and** $l_k > l_{k-1}$, **then**
1 prolongate Φ^{k-1} on grid $\mathcal{M}_{l_{k-1}}$ to $\Phi^{k,0}$ on grid $\mathcal{M}_{l_k}$
  end
2 $T_{\sigma_k} := G_{l_k}^{\sigma_k}[T]$, $R_{\sigma_k} := G_{l_k}^{\sigma_k}[R]$,
3 $N_{T_{\sigma_k}} := \frac{\nabla T_{\sigma_k}}{\|\nabla T_{\sigma_k}\|}$, $N_{R_{\sigma_k}} := \frac{\nabla R_{\sigma_k}}{\|\nabla R_{\sigma_k}\|}$,
  $i = 0$
  repeat
4 $d^{k,i} := -G_{l_k}^{\alpha}[\mathrm{grad}_{V_{l_k}} \tilde{E}_m^{\sigma_k}[\Phi^{k,i}]] - \mathrm{grad}_{V_{l_k}} E_{reg}[\Phi^{k,i}] - \mathrm{grad}_{V_{l_k}} E_f[\Phi^{k,i}]$
5 *line-search*: choose step size δ by Armijo's rule
6 $\Phi^{k,i+1} := \Phi^{k,i} + \delta\, d^{k,i}$
  $i := i + 1$
  until $(\|d^{k,i}\| \le TOL$ **or** $i > MAXITER)$;
  $\Phi^k = \Phi^{k,i}$
end

Acknowledgments. The authors thank C. Schaller from the University Hospital at Bonn for many valuable discussions on medical morphology and for providing real data sets. We also thank R. Lachner from BrainLab for providing the test data set, and Ulrich Clarenz and Günther Grün for helpful comments on an early version of the manuscript.

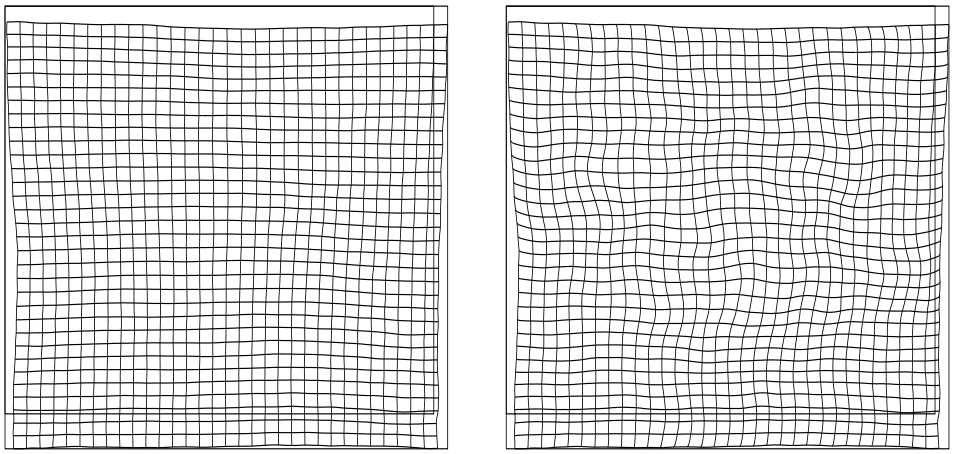


FIG. 8. *Left: deformation after the preregistration solely based on the feature energy. Right: final deformation after the registration including feature and morphological matching energy.*

REFERENCES

[1] L. ALVAREZ, F. GUICHARD, P. L. LIONS, AND J. M. MOREL, *Axioms and fundamental equations of image processing*, Arch. Ration. Mech. Anal., 123 (1993), pp. 199–257.

[2] L. ALVAREZ, J. WEICKERT, AND J. SÁNCHEZ, *A scale-space approach to nonlocal optical flow calculations*, in Scale-Space Theories in Computer Vision. Second International Conference, Scale-Space '99, Corfu, Greece, September 1999, M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert, eds., Lecture Notes in Comput. Sci. 1682, Springer, New York, 1999, pp. 235–246.

[3] J. BALL, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Ration. Mech. Anal., 63 (1977), pp. 337–403.

[4] J. BALL, *Global invertibility of Sobolev functions and the interpenetration of matter*, Proc. Roy. Soc. Edinburgh Sect. A, 88 (1981), pp. 315–328.

[5] C. BALLESTER, V. CASELLES, AND J. VERDERA, *Disocclusion by joint interpolation of vector fields and gray levels*, Multiscale Model. Simul., 2 (2003), pp. 80–123.

[6] C. BALLESTER, V. CASELLES, J. VERDERA, M. BERTALMIO, AND G. SAPIRO, *Filling-in by joint interpolation of vector fields and gray levels*, IEEE Trans. Image Process. Signal Anal., 10 (2001), pp. 1200–1211.

[7] M. BERTALMIO, G. SAPIRO, V. CASELLES, AND C. BALLESTER, *Image inpainting*, in Proceedings of ACM SIGGRAPH 2000, pp. 417–424.

[8] L. G. BROWN, *A survey of image registration techniques*, ACM Comput. Surveys, 24 (1992), pp. 325–376.

[9] V. CASELLES, F. CATTÉ, T. COLL, AND F. DIBOS, *A geometric model for active contours in image processing*, Numer. Math., 66 (1993).

[10] V. CASELLES, B. COLL, AND J. M. MOREL, *Topographics maps and local contrast invariance in natural images*, Int. J. Comput. Vision, 33 (1999), pp. 5–27.

[11] G. E. CHRISTENSEN, S. C. JOSHI, AND M. I. MILLER, *Volumetric transformations of brain anatomy*, IEEE Trans. Medical Imaging, 16 (1997), pp. 864–877.

[12] G. E. CHRISTENSEN, R. D. RABBITT, AND M. I. MILLER, *Deformable templates using large deformation kinematics*, IEEE Trans. Medical Imaging, 5 (1996), pp. 1435–1447.

[13] P. G. CIARLET, *Three-Dimensional Elasticity*, Elsevier, New York, 1988.

[14] P. G. CIARLET AND J. NECAS, *Injectivity and self-contact in nonlinear elasticity*, Arch. Ration. Mech. Anal., 97 (1987), pp. 171–188.

[15] U. CLARENZ, M. DROSKE, AND M. RUMPF, *Towards fast non-rigid registration*, in Inverse Problems, Image Analysis and Medical Imaging, AMS Special Session Interaction of Inverse Problems and Image Analysis, Vol. 313, AMS, Providence, RI, 2002, pp. 67–84.

[16] A. E. A. COLLIGNON, F. MAES, D. DELAERE, D. VANDERMEULEN, P. SUETENS, AND G. MARCHAL, *Automated multi-modality image registration based on information theory*, in Proceedings of the Information Processing in Medical Imaging Conference, Y. Bizais, ed., Kluwer Academic Publishers, Norwell, MA, 1995, pp. 263–274.

[17] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Appl. Math. Sci. 78, Springer, Berlin, 1989.
[18] C. A. Davatzikos, R. N. Bryan, and J. L. Prince, *Image registration based on boundary mapping*, IEEE Trans. Medical Imaging, 15 (1996), pp. 112–115.
[19] A. Desolneux, L. Moisan, and J. M. Morel, *Meaningful alignments*, Int. J. Comput. Vision, 40 (2000), pp. 7–23.
[20] M. Droske, B. Meyer, M. Rumpf, and C. Schaller, *An adaptive level set method for medical image segmentation*, in Proceedings of the Annual Symposium on Information Processing in Medical Imaging, Lecture Notes in Comput. Sci. 2082, R. Leahy and M. Insana, eds., Springer, New York, 2001, pp. 416–422
[21] U. Dupuis, P. Grenander, and M. I. Miller, *Variational problems on flows of diffeomorphisms for image matching*, Quart. Appl. Math., 56 (1998), pp. 587–600.
[22] I. Fonseca, G. Leoni, and J. Malý, *Weak continuity and lower semicontinuity results for determinants*, preprint 03-CNA-015 of the Center for Nonlinear Analysis (www.math.cmu.edu/cna).
[23] I. Fonseca and S. Müller, *Quasi-convex integrands and lower semicontinuity in L^1*, SIAM J. Math. Anal., 23 (1992), pp. 1081–1098.
[24] U. Grenander and M. I. Miller, *Computational anatomy: An emerging discipline*, Quart. Appl. Math., 56 (1998), pp. 617–694.
[25] W. Hackbusch, *Multigrid Methods and Applications*, Springer, Berlin, Heidelberg, 1985.
[26] M. Hanke and C. Groetsch, *Nonstationary iterated Tikhonov regularization*, J. Optim. Theory Appl., 98 (1998), pp. 37–53.
[27] S. Henn and K. Witsch, *A multigrid approach for minimizing a nonlinear functional for digital image matching*, Computing, 64 (2000), pp. 339–348.
[28] S. Henn and K. Witsch, *Iterative multigrid regularization techniques for image matching*, SIAM J. Sci. Comput., 23 (2001), pp. 1077–1093.
[29] W. Hinterberger, O. Scherzer, C. Schnörr, and J. Weickert, *Analysis of optical flow models in the framework of calculus of variations*, Numer. Funct. Anal. Optim., 23 (2002), pp. 69–89.
[30] P. Kosmol, *Optimierung und Approximation*, de Gruyter Lehrbuch, Walter de Gruyter, Berlin, 1991.
[31] J. E. Marsden and T. J. R. Hughes, *Mathematical Foundations of Elasticity*, Prentice–Hall, Englewood Cliffs, NJ, 1983.
[32] M. I. Miller, S. C. Joshi, and G. E. Christensen, *Brain Warping*, Academic Press, San Diego, CA, 1999.
[33] B. Fischer and J. Modersitzki, *Fast diffusion registration*, in Inverse Problems, Image Analysis, and Medical Imaging, Contemp. Math. 313, AMS, Providence, RI, 2002, pp. 117–127.
[34] J. Modersitzki and B. Fischer, *Curvature based image registration*, J. Math. Imaging Vision, 18 (2003), pp. 81–85.
[35] P. Monasse, *Contrast invariant registration of images*, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Phoenix, Arizona, Vol. 6, IEEE, 1999, pp. 3221–3224.
[36] C. Morrey, *Quasi-convexity and lower semicontinuity of multiple integrals*, Pacific J. Math., 2 (1952), pp. 25–53.
[37] C. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer, New York, 1966.
[38] R. W. Ogden, *Non-Linear Elastic Deformations*, John Wiley, New York, 1984.
[39] G. Sapiro, *Geometric Partial Differential Equations and Image Analysis*, Cambridge University Press, Cambridge, UK, 2001.
[40] J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press, New York, 1982.
[41] J. P. Thirion, *Image matching as a diffusion process: An analogy with Maxwell's demon*, Medical Imag. Anal., 2 (1998), pp. 243–260.
[42] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer, Berlin, 1984.
[43] C. Vachier, *Morphological scale-space analysis and feature extraction*, in Proceedings of the IEEE International Conference on Image Processing, Thessaloniki, Greece, 2001.
[44] P. A. van den Elsen, E.-J. J. Pol, and M. A. Viergever, *Medical image matching: A review with classification*, IEEE Engineering in Medicine and Biology, 12 (1993), pp. 26–39.
[45] P. A. Viola and W. M. Wells III, *Alignment by maximization of mutual information*, in Proceedings of the International Conference on Computer Vision, E. Grimson, S. Shafer, A. Blake, and K. Sugihara, eds., IEEE Computer Society Press, Los Alamitos, CA, 1995, pp. 16–23.
[46] V. Šverák, *Regularity properties of deformations with finite energy*, Arch. Ration. Mech. Anal., 100 (1988), pp. 105–127.
[47] W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, *Multi-modal volume registration by maximization of mutual information*, Med. Image Anal., 1(1996), pp. 35–51.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 688–708

ANALYSIS OF TRANSIENT ELECTROMAGNETIC SCATTERING FROM OVERFILLED CAVITIES*

TRI VAN† AND AIHUA WOOD‡

Abstract. In this paper, we consider the time-domain scattering problem of a two-dimensional overfilled cavity embedded in the infinite ground plane. The problem is first discretized in time by the β, γ Newmark time-marching scheme. At each time step, the variational formulation of the semidiscrete problem is derived via a nonlocal boundary condition to truncate the infinite problem domain. Existence and uniqueness of the variational solutions are established. Error analysis of the fully discrete problem is performed. Stability criteria of the time-stepping scheme are also obtained.

Key words. transient scattering, overfilled cavities in the infinite ground plane, time-domain finite element method, Newmark time-marching, error estimates, stability analysis

AMS subject classifications. 65M60, 65M12, 74J20

DOI. 10.1137/S0036139902419255

1. Introduction. The development of mathematical and numerical methods to accurately predict the radar signature of a target is an important area of research in electromagnetics. Of particular interest is the study of electromagnetic scattering from cavities and the calculation of their radar cross sections (RCS). This is because cavity RCS often dominates a target's overall RCS and is computationally challenging. One of the main difficulties in numerically approximating solutions involving cavities is the appearance of spurious modes caused by interior resonances. A variety of techniques have been developed to simulate the scattering by cavities. They include high and low frequency methods [9, 16, 6, 23, 20], the method of moments [29, 19, 30], the time-domain finite difference method [7], the finite element method [15, 22], and several hybrid methods [21, 14, 5]. These methods are limited to a certain range of frequencies and/or small/simple cavities. Recently, the hybrid finite element–boundary integral (FE-BI) methods have gained increasing popularity for their ability to model large and complex cavities [12, 13, 11, 18, 17]. Most of the results reported in the engineering literature appear experimental in nature. Mathematical treatment of scattering problems involving cavities can be found in [2, 3, 4, 25, 28, 27]. It is a common assumption that the cavity opening coincides with the aperture on an infinite ground plane, hence simplifying the modelling of the exterior (to the cavity) domain. This severely limits the application of these methods since many cavity openings are not planar. This paper aims to develop a solid mathematical technique that is capable of characterizing the scattering by overfilled cavities.

In particular, we seek to determine the fields scattered by the protruding cavity upon a given incident wave. Our method decomposes the entire infinite solution domain to two subdomains: the infinite upper half plane over the perfect electrically

*Received by the editors December 12, 2002; accepted for publication (in revised form) June 5, 2003; published electronically January 30, 2004. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the US Government.

http://www.siam.org/journals/siap/64-2/41925.html

†Mission Research Corporation, 3975 Research Blvd., Dayton, OH 45430 (tvan@mrcday.com).

‡Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB, OH, 45433-7765 (aihua.wood@afit.af.mil, aihuawood@aol.com). The research of this author was supported in part by Air Force Office of Scientific Research grant AFOSR-PO-2002005 and in part by a grant from the Dayton Area Graduate Studies Institute.

conducting (PEC) ground plane exterior to the hemisphere enclosing the cavity aperture, and the cavity plus the hemisphere region. The problem is solved exactly in the infinite subdomain, while the other is solved using finite elements. The two regions are coupled over the hemisphere via the introduction of a boundary operator exploiting the field continuity over material interfaces. Other FE-BI schemes, such as those reported in [12] and [11], employ image theory to produce conformal boundaries and hence limit the size of the finite element computational domain. The trade-off, however, is the complexity of the integral equation, which can be very expensive in order to achieve good accuracy. By using the semicircular boundary, exact series solutions are possible which can be accurately computed and are error-controllable. For protruding cavity problems, boundary conditions based on series representations appear more desirable.

In section 1 we discretize the problem in time using the Newmark time-stepping scheme. A similar scheme has been successfully implemented for the scattering by cavities embedded in the infinite ground plane [26]. The key point in the proof of the well-posedness of the resulting semidiscrete problem is the derivation and careful examination of the properties of the boundary operator. The problem is fully discretized in section 4, where stability criteria and error estimates are obtained. We believe this is the first thorough mathematical treatment of scattering problems involving overfilled cavities. It provides a solid mathematical foundation for researchers in both the applied mathematics and the engineering communities to develop accurate and efficient numerical solvers for RCS predictions for targets of a cavity nature. The paper is concluded in section 5.

2. Problem setting. Let $\Omega \subset \mathbb{R}^2$ be the cross section of a z-invariant cavity (or trough) in the infinite ground plane such that its fillings of relative permittivity $\varepsilon_r \geq 1$ protrude above the ground plane. Let $(\boldsymbol{E}^i, \boldsymbol{H}^i)$ be an electromagnetic wave incident on the cavity to generate the scattered field $(\boldsymbol{E}^s, \boldsymbol{H}^s)$. The scattering problem is to find $(\boldsymbol{E}^s, \boldsymbol{H}^s)$.

In the rest of the paper, we denote S as the cavity wall, Γ as the cavity aperture so that $\partial\Omega = S \cup \Gamma$. The infinite ground plane excluding the cavity opening is denoted as Γ_{ext}. Finally, the infinite homogeneous region above the cavity is denoted as $\mathcal{U} = \mathbb{R}^2_+ \setminus \Omega$.

Due to the uniformity in the z-axis, the scattering problem can be decomposed into two fundamental polarizations: transverse magnetic (TM) and transverse electric (TE). Its solution then can be expressed as a linear combination of the solutions to TM and TE problems. In the TM polarization, the magnetic field $\boldsymbol{H}$ is transverse to the z-axis so that $\boldsymbol{E}$ and $\boldsymbol{H}$ are of the form

$$\boldsymbol{E} = (0, 0, E_z), \quad \boldsymbol{H} = (H_x, H_y, 0).$$

In this case, the nonzero component of the total field, also denoted as $\boldsymbol{E}$, satisfies the following equation:

$$\text{(TM)} \quad \begin{cases} -\Delta E_z + \varepsilon_r \dfrac{\partial^2 E_z}{\partial t^2} &= 0 \quad \text{in } \Omega \cup \mathcal{U} \times (0, \infty), \\ E_z &= 0 \quad \text{on } S \cup \Gamma_{ext} \times (0, \infty), \\ E_z|_{t=0} = E_0, \qquad & \left.\dfrac{\partial E_z}{\partial t}\right|_{t=0} = E_{t,0}, \end{cases}$$

where ε_r is the relative electric permittivity, E_0 and $E_{t,0}$ are given initial conditions. The homogeneous region $\mathcal{U}$ above the protruding cavity is assumed to be air, and hence

its permittivity is $\varepsilon_r = 1$. In $\mathcal{U}$, the total field can be decomposed as $E_z = E_z^i + E_z^r + E_z^s$, where E_z^i is the incident field, E_z^r the reflected field, and E_z^s the scattered field. The reflected field exists due to the presence of the infinite ground plane. The incident and reflected electric fields satisfy

$$E_z^i + E_z^r = 0 \quad \text{on } \Gamma_{ext} \subset \{(x, y) : y = 0\}.$$

The scattered field E_z^s solves

$$(\mathrm{TM}^s) \quad \begin{cases} -\Delta E_z^s + \dfrac{\partial^2 E_z^s}{\partial t^2} &= 0 \quad \text{in } \mathcal{U} \times (0, \infty), \\ E_z^s &= E_z - E_z^i - E_z^r \quad \text{on } \Gamma \times (0, \infty), \\ E_z^s &= 0 \quad \text{on } \Gamma_{ext} \times (0, \infty) \end{cases}$$

and satisfies the radiation condition at infinity, that is,

$$\lim_{r \to \infty} \sqrt{r} \left(\frac{\partial E_z^s}{\partial r} + \frac{1}{c} \frac{\partial E_z^s}{\partial t} \right) = 0, \quad t > 0. \tag{2.1}$$

The components of $\boldsymbol{H}$ can be obtained in terms of E_z and its partial derivatives by using Maxwell's equations.

Similarly, in the TE polarization, the electric field $\boldsymbol{E}$ is transverse to the z-axis and hence

$$\boldsymbol{E} = (E_x, E_y, 0), \quad \boldsymbol{H} = (0, 0, H_z).$$

The nonzero component of the total magnetic field, also denoted by $\boldsymbol{H}$, satisfies the following equation:

$$(\mathrm{TE}) \quad \begin{cases} -\nabla \cdot \left(\frac{1}{\varepsilon_r} \nabla H_z \right) + \dfrac{\partial^2 H_z}{\partial t^2} &= 0 \quad \text{in } \Omega \cup \mathcal{U} \times (0, \infty), \\ \dfrac{\partial H_z}{\partial n} &= 0 \quad \text{on } S \cup \Gamma_{ext} \times (0, \infty), \\ H_z|_{t=0} = H_0, & \dfrac{\partial H_z}{\partial t}|_{t=0} = H_{t,0}, \end{cases}$$

where H_0 and $H_{t,0}$ are given initial conditions. In U_R, the total magnetic field can be decomposed into $H_z = H_z^i + H_z^r + H_z^s$, where

$$\frac{\partial H_z^i}{\partial y} + \frac{\partial H_z^s}{\partial y} = 0 \quad \text{on } \{(x, y) : y = 0\}.$$

The scattered field solves

$$(\mathrm{TE}^s) \quad \begin{cases} -\Delta H_z^s + \dfrac{\partial^2 H_z^s}{\partial t^2} &= 0 \quad \text{in } \mathcal{U} \times (0, \infty), \\ \dfrac{\partial H_z^s}{\partial n} &= 0 \quad \text{on } \Gamma_{ext}, \\ \dfrac{\partial H_z^s}{\partial n} &= \dfrac{1}{\varepsilon_r} \dfrac{\partial H_z}{\partial n} - \dfrac{\partial H_z^i}{\partial n} - \dfrac{\partial H_z^r}{\partial n} \quad \text{on } \Gamma \times (0, \infty), \end{cases}$$

where $\frac{\partial}{\partial n}$ is the normal derivative on Γ. The scattered magnetic field also satisfies the same radiation condition defined in (2.1). The components of $\boldsymbol{E}$ can be obtained in terms of H_z and its partial derivatives by using Maxwell's equations.

In the next section, we apply the Newmark time-marching scheme to temporally discretize (TM), (TM^s), (TE), and (TE^s). The resulting equations are called *semi-discrete* equations since only the time variable t is discretized.

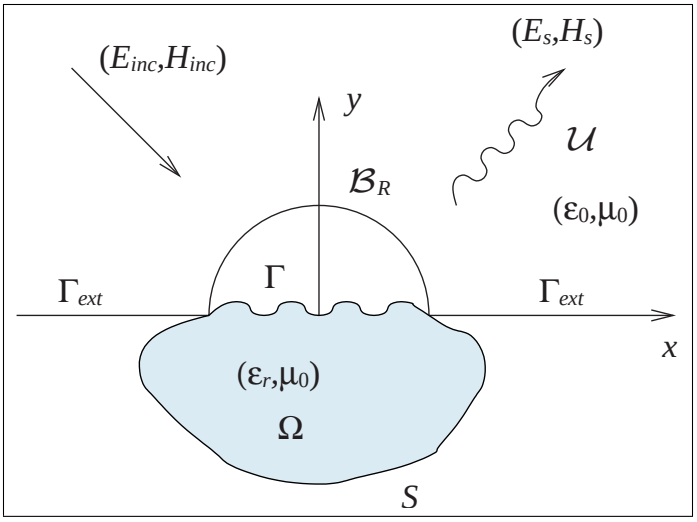


FIG. 1. *Cavity setting.*

3. Semidiscrete problems. First, let $\mathcal{B}_R$ be a semicircle of radius R large enough to completely enclose the overfilled portion of the cavity (see Figure 1). We denote the region bounded by $\mathcal{B}_R$ and the cavity wall S as Ω_R. Hence, this region Ω_R consists of the cavity and the homogeneous part between $\mathcal{B}_R$ and Γ. Let U_R be the homogeneous region outside of Ω_R, that is, $U_R = \{(r,\theta) : r > R, 0 < \theta < \pi\}$.

In this section, the TM and TE equations are discretized in time by using the Newmark time-marching scheme. The Newmark algorithm is a two-step algorithm and can be used in predictor-corrector form [31, 32]. At each time step, we construct a nonlocal boundary condition on the semicircle $\mathcal{B}_R$ to couple the solution in the infinite domain exterior to $\mathcal{B}_R$ and the solution in the bounded domain inside $\mathcal{B}_R$. The boundary condition enables us to reduce the infinite problem domain into a bounded one. The Newmark scheme is defined by the following.

Let $\mathcal{N}$ be a positive integer, $\Delta t = T/\mathcal{N}$ be the temporal step size, and $t_{n+1} = (n+1)\Delta t$ for $n = 0, 1, 2, \ldots, \mathcal{N}-1$. Denote u^{n+1}, $\dot{u}^{n+1}$, and $\ddot{u}^{n+1}$ as the approximations of u, $\frac{\partial u}{\partial t}$, and $\frac{\partial^2 u}{\partial t^2}$ at $t = t_{n+1}$, respectively. These approximations are related by

$$u^{n+1} = u^n + \Delta t \dot{u}^n + \frac{\Delta t^2}{2}\left[2\beta\ddot{u}^{n+1} + (1-2\beta)\ddot{u}^n\right], \quad 0 \le n \le \mathcal{N}-1,$$
$$\dot{u}^{n+1} = \dot{u}^n + \Delta t\left[\gamma\ddot{u}^{n+1} + (1-\gamma)\ddot{u}^n\right], \quad 0 \le n \le \mathcal{N}-1,$$

where γ and β are parameters to be determined to guarantee stability of the time-marching scheme.

3.1. TM polarization. For convenience, we denote u^i as the incident field E_z^i, u^r the reflected field E_z^r, u the total field E_z, and u^s the scattered field E_z^s. The semidiscrete problem is to find u^{n+1}, $n = 0, 1, \ldots, \mathcal{N}$ such that we have the following:

Prediction

$$\tilde{u}^{n+1} = u^n + \Delta t \dot{u}^n + \frac{(\Delta t)^2}{2}(1-2\beta)\ddot{u}^n, \tag{3.1}$$

$$\tilde{\dot{u}}^{n+1} = \dot{u}^n + \Delta t(1-\gamma)\ddot{u}^n, \tag{3.2}$$

Solution

$$\text{(3.3)}\qquad \begin{cases} -\Delta u^{n+1} + \alpha^2 \varepsilon_r u^{n+1} &= \alpha^2 \varepsilon_r \tilde{u}^{n+1} \quad \text{in } \Omega_R, \\ u^{n+1} &= 0 \quad \text{on } S, \\ u^{n+1} &= u^{s,n+1} + u^{i,n+1} + u^{r,n+1} \quad \text{on } \mathcal{B}_R, \end{cases}$$

Correction

$$\ddot{u}^{n+1} = \alpha^2 (u^{n+1} - \tilde{u}^{n+1}), \tag{3.4}$$

$$\dot{u}^{n+1} = \tilde{\dot{u}}^{n+1} + \Delta t \gamma \ddot{u}^{n+1}, \tag{3.5}$$

where $\alpha^2 = \frac{1}{\Delta t^2 \beta}$. The scattered field $u^{s,n+1}$ satisfies the following *exterior problem*:

$$\text{(3.6)}\qquad \begin{cases} -\Delta u^{s,n+1} + \alpha^2 u^{s,n+1} &= \alpha^2 \tilde{u}^{s,n+1} \quad \text{in } U_R, \\ u^{s,n+1}(R,\theta) &= g(R,\theta) \quad \text{on } \mathcal{B}_R, \\ u^{s,n+1} &= 0 \quad \text{on } \Gamma_{ext}, \end{cases}$$

where $g \overset{\text{def}}{=} u^{n+1} - u^{i,n+1} - u^{r,n+1}$, and

$$\lim_{r\to\infty} \sqrt{r} \left(\frac{\partial u^{s,n+1}}{\partial r} + \frac{1}{c} \dot{u}^{s,n+1} \right) = 0. \tag{3.7}$$

3.1.1. Exterior problem and transparent boundary condition. For brevity, we omit the superscript $n+1$ in u^{n+1} in the remainder of the section. To solve the exterior problem (3.6), we set, in the polar coordinates (r,θ),

$$u^s(r,\theta) = \sum_{m=1}^{\infty} A_m(r) \sin 2m\theta, \quad r \geq R,\ \theta \in [0,\pi],$$

where $A_m(r)$ is to be determined. We choose the sine series expansion because u^s vanishes on Γ_{ext}, i.e., $\theta = 0, \pi$ and $r \geq R$. By substituting the sine series into (3.6), we obtain the nonhomogeneous equation with a Dirichlet boundary condition

$$\text{(3.8)}\qquad \begin{cases} \left[\dfrac{d^2}{dr^2} + \dfrac{1}{r}\dfrac{d}{dr} - \left(\alpha^2 + \dfrac{4m^2}{r^2} \right) \right] A_m(r) &= \alpha^2 \tilde{f}_m^s(r), \quad r \in (R,\infty), \\ A_m(R) &= g_m(R), \end{cases}$$

where f_m^s and g_m are the coefficients of the sine series expansions of $\tilde{u}^s$ and g, respectively. The solution to (3.8) can be expressed as

$$A_m(r) = A_m^{\mathrm{h}}(r) + A_m^{\mathrm{p}}(r), \tag{3.9}$$

where $A_m^{\mathrm{h}}(r)$ is the solution of the homogeneous counterpart of (3.8) and $A_m^{\mathrm{p}}(r)$ is the solution of (3.8) with homogeneous boundary condition $A_m^{\mathrm{p}}(R) = 0$. To be more precise, A_m^{h} and A_m^{p} solve, respectively,

$$\text{(3.10)}\qquad \begin{cases} \left[\dfrac{d^2}{dr^2} + \dfrac{1}{r}\dfrac{d}{dr} - \left(\alpha^2 + \dfrac{4m^2}{r^2} \right) \right] A_m^{\mathrm{h}}(r) &= 0, \quad r \in (R,\infty), \\ A_m^{\mathrm{h}}(R) &= g_m(R) \end{cases}$$

and

$$\left\{ \begin{array}{rcl} \left[\dfrac{d^2}{dr^2}+\dfrac{1}{r}\dfrac{d}{dr}-\left(\alpha^2+\dfrac{4m^2}{r^2}\right)\right] A_m^{\mathrm{p}}(r) &=& \alpha^2 \tilde{f}_m^s(r), \quad r\in(R,\infty), \\ A_m^{\mathrm{p}}(R) &=& 0. \end{array} \right. \tag{3.11}$$

Equation (3.10) is a modified Bessel's equation of order $2m$. By the boundary condition and noting that the solution is bounded near infinity by the radiation condition, we have

$$A_m^{\mathrm{h}}(r) = \frac{g_m(R)}{K_{2m}(\alpha R)} K_{2m}(\alpha r), \quad r \geq R. \tag{3.12}$$

The solution to (3.11) can be obtained by applying the Green's function method as discussed in [10]. The Green's function associated to (3.11) is defined by

$$G_m(r,r') = \left\{ \begin{array}{ll} \dfrac{k_m(r')l_m(r)}{r'^2 W_m(r')}, & r \geq r', \\ \dfrac{l_m(r')k_m(r)}{r'^2 W_m(r')}, & r \leq r', \end{array} \right. \tag{3.13}$$

where

$$\begin{aligned} k_m(r') &= K_{2m}(\alpha r'), \\ l_m(r') &= I_{2m}(\alpha r')K_{2m}(\alpha R) - K_{2m}(\alpha r')I_{2m}(\alpha R), \\ W_m(r') &= l_m(r')k_m'(r') - k_m(r')l_m'(r'). \end{aligned}$$

Note that $G_m(R,r') = 0$ for all $r' \leq R$ and $G_m(r,R) = 0$ for $r \leq R$. Therefore, the solution to (3.11) is of the form

$$A_m^{\mathrm{p}}(r) = \alpha^2 \int_R^\infty G_m(r,r')\tilde{f}_m^s(r')dr', \quad r \geq R. \tag{3.14}$$

Consequently,

$$A_m(r) = \frac{g_m(R)}{K_{2m}(\alpha R)} K_{2m}(\alpha r) + \alpha^2 \int_R^\infty G_m(r,r')\tilde{f}_m^s(r')dr'.$$

Therefore, the scattered field u^s can be expressed explicitly by

$$u^s(r,\theta) = \sum_{m=1}^\infty \left[\frac{g_m(R)}{K_{2m}(\alpha R)} K_{2m}(\alpha r) + \alpha^2 \int_R^\infty G_m(r,r')\tilde{f}_m^s(r')dr' \right] \sin m\theta,$$

where

$$\begin{aligned} g_m(R) &= \frac{2}{\pi}\int_0^\pi (u - u^i - u^r)(R,\theta')\sin 2m\theta' d\theta', \\ f_m^s(r') &= \frac{2}{\pi}\int_0^\pi \tilde{u}^s(r',\theta')\sin 2m\theta' d\theta'. \end{aligned}$$

By using

$$\begin{aligned} \sin 2m\theta \sin 2m\theta' &= \frac{1}{2}[\cos 2m(\theta-\theta') - \cos 2m(\theta+\theta')] \\ &=: \frac{1}{2}C(\theta,\theta'), \end{aligned}$$

we can write

$$u^s(r,\theta) = \frac{2}{\pi}\sum_{m=1}^{\infty}\left\{\frac{K_{2m}(\alpha r)}{K_{2m}(\alpha R)}\int_0^{\pi} C(\theta,\theta')(u-u^i-u^r)(R,\theta')d\theta' + \alpha^2\int_R^{\infty} dr' G_m(r,r')\int_0^{\pi} d\theta' C(\theta,\theta')\tilde{u}^s(r',\theta')\right\}.$$

Taking the partial derivative $\frac{\partial}{\partial r}$ of u^s yields

$$\frac{\partial u^s}{\partial r} = \mathcal{T}(u-u^i-u^r) + \Phi(\tilde{u}^s), \quad r \geq R,\ \theta \in [0,\pi], \tag{3.15}$$

where $\mathcal{T}$ is a linear operator defined by

$$\mathcal{T}(v) = \frac{1}{\pi}\sum_{m=1}^{\infty}\alpha\frac{K'_{2m}(\alpha r)}{K_{2m}(\alpha R)}\int_0^{\pi} C(\theta,\theta')v(\theta')d\theta' \quad \forall v \in H^{1/2}(\mathcal{B}_R)$$

and

$$\Phi(\tilde{u}^s) = \frac{\alpha^2}{\pi}\sum_{m=1}^{\infty}\int_R^{\infty} dr'\frac{\partial G_m(r,r')}{\partial r}\int_0^{\pi} d\theta' C(\theta,\theta')\tilde{u}^s(r',\theta').$$

We now construct the nonlocal boundary condition on $\mathcal{B}_R$ using (3.15). Define the Sobolev spaces $H^{1/2}(\mathcal{B}_{\mathcal{R}})$ and $H^{-1/2}(\mathcal{B}_{\mathcal{R}})$, respectively, as

$$H^{1/2}(\mathcal{B}_R) = \left\{\phi : \sum_{m=0}^{\infty}\sqrt{1+m^2}|\phi_m|^2 < \infty\right\}, \tag{3.16}$$

$$H^{-1/2}(\mathcal{B}_R) = \left\{\phi : \sum_{m=0}^{\infty}\frac{1}{\sqrt{1+m^2}}|\phi_m|^2 < \infty\right\}, \tag{3.17}$$

where

$$\phi_m = \phi_m^c + i\phi_m^s = \frac{2}{\pi}\int_0^{\pi}\phi(\theta)e^{i2m\theta}d\theta.$$

Let $\mathcal{T}_R$ be the restriction of $\mathcal{T}$ to $\mathcal{B}_R$ so that $\mathcal{T}_R : H^{1/2}(\mathcal{B}_R) \to H^{-1/2}(\mathcal{B}_R)$, and let it be defined by

$$\mathcal{T}_R g(\theta) = \frac{1}{\pi}\sum_{m=1}^{\infty}\alpha\frac{K'_{2m}(\alpha R)}{K_{2m}(\alpha R)}\int_0^{\pi} C(\theta,\theta')g(\theta')d\theta', \tag{3.18}$$

where $\theta \in [0,\pi]$.

PROPOSITION 3.1. *The operator $\mathcal{T}_R$ is symmetric and bounded, that is,*

$$\langle \mathcal{T}_R g, \phi\rangle = \langle \mathcal{T}_R\phi, g\rangle,$$
$$\|\mathcal{T}_R g\|_{H^{-1/2}(\mathcal{B}_R)} \leq C\|g\|_{H^{1/2}(\mathcal{B}_R)} \quad \forall g \in H^{1/2}(\mathcal{B}_R),$$

where C is a constant.

Proof. By definition, we have

$$\|\mathcal{T}_R g\|_{H^{-1/2}(\mathcal{B}_R)} = \sup_{\phi\in H^{1/2}(\mathcal{B}_R)}\frac{|\langle \mathcal{T}_R g,\phi\rangle|}{\|\phi\|_{H^{1/2}(\mathcal{B}_R)}},$$

where

$$\langle \mathcal{T}_R g, \phi \rangle = \frac{1}{\pi} \sum_{m=1}^{\infty} \alpha \frac{K'_{2m}(\alpha R)}{K_{2m}(\alpha R)} \int_0^{\pi} d\theta \phi(\theta) \int_0^{\pi} d\theta' C(\theta, \theta') g(\theta').$$

By writing $\cos 2m(\theta \pm \theta') = \mathrm{Re}\{e^{i2m\theta} e^{\pm i2m\theta'}\}$, we have

$$\langle \mathcal{T}_R g, \phi \rangle = \frac{1}{2} \sum_{m=1}^{\infty} \alpha \frac{K'_{2m}(\alpha R)}{K_{2m}(\alpha R)} \phi_m^s g_m^s;$$

thus, $\mathcal{T}_R$ is symmetric and

$$|\langle \mathcal{T}_R g, \phi \rangle| \le \frac{1}{2} \sqrt{\sum_{m=1}^{\infty} \alpha \left| \frac{K'_{2m}(\alpha R)}{K_{2m}(\alpha R)} (\phi_m^s)^2 \right|} \sqrt{\sum_{m=1}^{\infty} \alpha \left| \frac{K'_{2m}(\alpha R)}{K_{2m}(\alpha R)} (g_m^s)^2 \right|}.$$

For sufficiently large m, $\alpha \frac{K'_{2m}(\alpha R)}{K_{2m}(\alpha R)} < \sqrt{1+m^2}$ (see [1]). Hence there exists a constant C independent of ϕ and g such that

$$|\langle \mathcal{T}_R g, \phi \rangle| \le C \|\phi\|_{H^{1/2}(\mathcal{B}_R)} \|g\|_{H^{1/2}(\mathcal{B}_R)}.$$

This proves the proposition. □

Hence we can easily see that on $\mathcal{B}_R$ the normal derivative of the total electric field satisfies the following continuity condition:

$$\begin{aligned} \left.\frac{\partial u}{\partial r}\right|_{r=R} &= \left.\frac{\partial u^i}{\partial r}\right|_{r=R} + \left.\frac{\partial u^r}{\partial r}\right|_{r=R} + \left.\frac{\partial u^s}{\partial r}\right|_{r=R} \\ &= \frac{\partial}{\partial r}(u^i + u^r)|_{r=R} + \mathcal{T}_R(u - u^i - u^r) + \Phi_R(\tilde{u}^s), \end{aligned} \tag{3.19}$$

where Φ_R is the restriction of Φ to $r = R$. Equation (3.19) is the transparent boundary condition on $\mathcal{B}_R$. It enables one to couple the total field in the infinite homogeneous domain U_R to the total field in the bounded domain Ω_R through the operator $\mathcal{T}_R$. We can rewrite the boundary value problem (3.3) as

$$\begin{cases} -\Delta u^{n+1} + \alpha^2 \varepsilon_r u^{n+1} = \alpha^2 \varepsilon_r \tilde{u}^{n+1} & \text{in } \Omega_R, \\ u^{n+1} = 0 \quad \text{on } S, \\ \dfrac{\partial u^{n+1}}{\partial r} - \mathcal{T}_R u^{n+1} = \dfrac{\partial u^{i,n+1}}{\partial r} + \dfrac{\partial u^{r,n+1}}{\partial r} \\ \qquad + \Phi_R(\tilde{u}^{s,n+1}) - \mathcal{T}_R(u^{i,n+1} + u^{r,n+1}) \quad \text{on } \mathcal{B}_R. \end{cases} \tag{3.20}$$

Next, we solve (3.20) by a variational method.

3.1.2. Variational formulation. Define the subspace V of $L^2(\Omega_R)$ by

$$V = \{v \in H^1(\Omega_R) : v|_S = 0\}$$

equipped with the H^1-norm, that is,

$$\|u\|_V = \|u\|_{H^1(\Omega_R)}.$$

The variational formulation of (3.3), or equivalently (3.20), is to find $u \in V$ such that

$$b_{TM}(u,v) = F(v) \quad \forall v \in V, \tag{3.21}$$

where

$$b_{TM}(u,v) = \int_\Omega (\nabla u \cdot \nabla v + \alpha^2 \varepsilon_r uv)\, dxdy - \int_{\mathcal{B}_R} \mathcal{T}_R(u) v\, d\theta, \tag{3.22}$$

$$F(v) = \alpha^2 \int_\Omega \varepsilon_r \tilde{u} v\, dxdy + \int_{\mathcal{B}_R} Hv\, d\theta + \int_{\mathcal{B}_R} \Phi_R(\tilde{u}^s) v d\theta, \tag{3.23}$$

with

$$H = \frac{\partial u^i}{\partial r}\Big|_{r=R} + \frac{\partial u^r}{\partial r}\Big|_{r=R} - \mathcal{T}_R(u^i|_{r=R} + u^r|_{r=R}).$$

We note that $F(v)$ in (3.23) contains the solution and its time derivatives in the previous time step through the term $\Phi_R(\tilde{u}^s)$.

THEOREM 3.2. *The variational problem* (3.21) *has a unique solution* $u \in V$ *and*

$$\|u\|_V \le C \left(\|\varepsilon_r \tilde{u}\|_{L^2(\Omega_R)} + \|u^i\|_V + \|u^r\|_V + \|\tilde{u}^s\|_V \right).$$

Proof. Since $\langle \mathcal{T}_R g, g \rangle \le 0$ and $\mathcal{T}_R : H^{1/2}(\mathcal{B}_R) \to H^{-1/2}(\mathcal{B}_R)$ is bounded, the symmetric bilinear form is coercive and bounded. Thus, by the Lax–Milgram lemma, (3.21) has a unique solution and

$$\|u\|_V \le C \|F\|_{V'}.$$

Standard trace theory then gives

$$\|F\|_{V'} \le C \left(\|\varepsilon_r \tilde{u}\|_{L^2(\Omega_R)} + \|u^i\|_V + \|u^r\|_V + \|\tilde{u}^s\|_V \right).$$

This proves the theorem. ∎

The Newmark time-marching algorithm for (3.20) can be described as follows:

1. Form the system matrix K from the bilinear forms b_{TM}.
   Time-loop: for $n = 0, 1, 2 \ldots$
   2. Compute the predicted values $\tilde{u}^{n+1}$, $\tilde{\dot{u}}^{n+1}$ in the interior Ω_R.
   3. Compute the predicted values $\tilde{u}^{n+1}$, $\tilde{\dot{u}}^{n+1}$ in the exterior region $\mathcal{U}_R$.
   4. Form the right-hand side vector F^{n+1}.
   5. Solve (3.21): $Ku^{n+1} = F^{n+1}$ (in Ω_R).
   6. Compute the solution u^{n+1} in the exterior $\mathcal{U}_R$.
   7. Correct $\ddot{u}^{n+1}$ and $\dot{u}^{n+1}$ in Ω_R.
   8. Correct $\ddot{u}^{n+1}$ and $\dot{u}^{n+1}$ in $\mathcal{U}_R$.

The term $\ddot{u}^0$ can be approximated as

$$\ddot{u}^0 = \frac{\dot{u}^0 - \dot{u}^{-1}}{\Delta t},$$

with $\dot{u}^{-1} = 0$; that is, all fields are assumed to be zero for $t < 0$.

3.2. TE polarization. As in the TM case, we denote u^i as the incident field H_z^i, u^r reflected field H_z^r, u total field H_z, and u^s scattered field H_z^s. It is known that

$$\frac{\partial u^i}{\partial y} + \frac{\partial u^r}{\partial y} = 0 \quad \text{on } \Gamma_{ext}.$$

The semidiscrete problem is to find u^{n+1}, $n = 0, 1, 2, \ldots, \mathcal{N}$, such that we have the following:

Prediction

$$\tilde{u}^{n+1} = u^n + \Delta t \dot{u}^n + \frac{(\Delta t)^2}{2}(1 - 2\beta)\ddot{u}^n, \tag{3.24}$$

$$\tilde{\dot{u}}^{n+1} = \dot{u}^n + \Delta t(1 - \gamma)\ddot{u}^n, \tag{3.25}$$

Solution

$$\begin{cases} -\nabla \cdot (\varepsilon_r^{-1} \nabla u^{n+1}) + \alpha^2 u^{n+1} &= \alpha^2 \tilde{u}^{n+1} \quad \text{in } \Omega_R, \\ \dfrac{\partial u^{n+1}}{\partial n} &= 0 \quad \text{on } S, \\ u^{n+1} &= u^{s,n+1} + u^{i,n+1} + u^{r,n+1} \quad \text{on } \mathcal{B}_R, \end{cases} \tag{3.26}$$

Correction

$$\ddot{u}^{n+1} = \alpha^2 (u^{n+1} - \tilde{u}^{n+1}), \tag{3.27}$$

$$\dot{u}^{n+1} = \tilde{\dot{u}}^{n+1} + \Delta t \gamma \ddot{u}^{n+1}, \tag{3.28}$$

where $\alpha^2 = \frac{1}{\Delta t^2 \beta}$.

The scattered field $u^{s,n+1}$ satisfies the exterior problem

$$\begin{cases} -\Delta u^{s,n+1} + \alpha^2 u^{s,n+1} &= \alpha^2 \tilde{u}^{s,n+1} \quad \text{in } U_R, \\ \dfrac{\partial u^{s,n+1}}{\partial r} &= 0 \quad \text{on } \Gamma_{ext}, \\ \dfrac{\partial u^{s,n+1}}{\partial r} &= h(R, \theta) \quad \text{on } \mathcal{B}_R, \end{cases} \tag{3.29}$$

where

$$h = \frac{\partial u^{n+1}}{\partial r} - \frac{\partial u^{i,n+1}}{\partial r} - \frac{\partial u^{r,n+1}}{\partial r},$$

and the radiation condition

$$\lim_{r \to \infty} \sqrt{r} \left(\frac{\partial u^{s,n+1}}{\partial r} + \frac{1}{c} \dot{u}^{s,n+1} \right) = 0.$$

In the next subsection, we construct the scattered field in U_R and then use it to construct the Dirichlet-to-Neumann mapping on $\mathcal{B}_R$. Again, the superscripts $n + 1$ are temporarily omitted for brevity.

3.2.1. Exterior problem and transparent boundary condition. We expand the scattered field in U_R as

$$u^s(R, \theta) = \sum_{m=0} B_m(r) \cos 2m\theta, \quad r \geq R, \ \theta \in [0, \pi], \tag{3.30}$$

where $B_m(r)$ is to be determined. Here the cosine series expansion is chosen because $\frac{\partial u^s}{\partial y}$ vanishes for $\theta = 0, \pi$ and $r \geq R$. By substituting (3.30) into (3.29), we obtain the nonhomogeneous equation with a Neumann boundary condition

$$\begin{cases} \left[\dfrac{d^2}{dr^2} + \dfrac{1}{r}\dfrac{d}{dr} - \left(\alpha^2 + \dfrac{4m^2}{r^2}\right)\right] B_m(r) &= \alpha^2 \tilde{f}_m^c(r), \quad r \in (R, \infty), \\ B_m'(R) &= h_m(R), \end{cases} \tag{3.31}$$

where f_m^c and h_m are the coefficients of the cosine series expansions of $\tilde{u}^s$ and h, respectively. The solution to (3.31) can be written as

$$B_m(r) = B_m^{\mathrm{h}}(r) + B_m^{\mathrm{p}}(r),$$

where $B_m^{\mathrm{h}}(r)$ and $B_m^{\mathrm{p}}(r)$ solve, respectively,

$$\begin{cases} \left[\dfrac{d^2}{dr^2} + \dfrac{1}{r}\dfrac{d}{dr} - \left(\alpha^2 + \dfrac{4m^2}{r^2}\right)\right] B_m^{\mathrm{h}}(r) &= 0, \quad r \in (R, \infty), \\ B_m'(R) &= h_m(R) \end{cases} \tag{3.32}$$

and

$$\begin{cases} \left[\dfrac{d^2}{dr^2} + \dfrac{1}{r}\dfrac{d}{dr} - \left(\alpha^2 + \dfrac{4m^2}{r^2}\right)\right] B_m^{\mathrm{p}}(r) &= \alpha^2 \tilde{f}_m^c(r), \quad r \in (R, \infty), \\ B_m'(R) &= 0. \end{cases} \tag{3.33}$$

Again, by applying the boundary and radiation conditions to the general solution of the modified Bessel's equation (3.32), we obtain

$$B_m^{\mathrm{h}}(r) = \frac{h_m(R)}{\alpha K_{2m}'(\alpha R)} K_{2m}(\alpha r). \tag{3.34}$$

Equation (3.33) can be solved by applying the Green's function method. The Green's function associated to this equation is defined by

$$G_m(r, \xi) = \begin{cases} \dfrac{k_m(\xi) l_m(r)}{\xi^2 W_m(\xi)}, & r \geq \xi, \\ \dfrac{l_m(\xi) k_m(r)}{\xi^2 W_m(\xi)}, & r \leq \xi, \end{cases} \tag{3.35}$$

where

$$\begin{aligned} k_m(\xi) &= K_{2m}(\alpha\xi), \\ l_m(\xi) &= \alpha[I_{2m}(\alpha\xi) K_{2m}'(\alpha R) - K_{2m}(\alpha\xi) I_{2m}'(\alpha R)], \\ W_m(\xi) &= l_m(\xi) k_m'(\xi) - k_m(\xi) l_m'(\xi). \end{aligned}$$

Hence the particular solution is of the form

$$B_m^{\mathrm{p}}(r) = \alpha^2 \int_R^\infty G_m(r, r') \tilde{f}_m^c(r') dr', \quad r \geq R.$$

Consequently, the scattered field u^s can be expressed as

$$u^s(r, \theta) = \sum_{m=0}^{\infty} \left[\frac{h_m(R)}{\alpha K_{2m}'(\alpha R)} K_{2m}(\alpha r) + \alpha^2 \int_R^\infty G_m(r, r') \tilde{f}_m^c(r') dr'\right] \cos 2m\theta,$$

where

$$h_m(R) = \frac{2}{\pi}\int_0^\pi \left(\frac{\partial u}{\partial r} - \frac{\partial u^i}{\partial r} - \frac{\partial u^r}{\partial r}\right)(R,\theta')\cos 2m\theta'\, d\theta',$$
$$\tilde{f}_m^c(r') = \frac{2}{\pi}\int_0^\pi \tilde{u}(r',\theta')\cos 2m\theta'\, d\theta'.$$

Using the identity $\cos 2m\theta \cos 2m\theta' = \frac{1}{2}[\sin 2m(\theta+\theta') - \sin 2m(\theta-\theta')] = \frac{1}{2}S(\theta,\theta')$ yields

$$u^s(r,\theta) = \mathcal{S}\left(\frac{\partial u}{\partial r} - \frac{\partial u^i}{\partial r} - \frac{\partial u^r}{\partial r}\right) + \Psi(\tilde{u}^s) \quad \text{in } U_R,$$

where $\mathcal{S}$ is the mapping defined by

$$\mathcal{S}g(\theta) = \frac{1}{\pi}\sum_{m=0}^{\infty}\left[\frac{K_{2m}(\alpha r)}{\alpha K'_{2m}(\alpha R)}\int_0^\pi S(\theta,\theta')g(\theta')\, d\theta'\right],$$

and

$$\Psi(\tilde{u}^s)(r,\theta) = \alpha^2\sum_{m=0}^{\infty}\int_R^\infty dr' G_m(r,r')\int_0^\pi d\theta' S(\theta,\theta')\tilde{u}^s(r',\theta').$$

Denote $\mathcal{S}_R$ as the restriction of $\mathcal{S}$ on $\mathcal{B}_R$ such that

$$\mathcal{S}_R g = \frac{1}{\pi}\sum_{m=0}^{\infty}\left[\frac{K_{2m}(\alpha R)}{\alpha K'_{2m}(\alpha R)}\int_0^\pi S(\theta,\theta')g(\theta')\, d\theta'\right]$$

for all $g \in H^{-1/2}(\mathcal{B}_R)$. We can show the following proposition.

PROPOSITION 3.3. *The operator $\mathcal{S}_R : H^{-1/2}(\mathcal{B}_R) \to H^{1/2}(\mathcal{B}_R)$ is symmetric and bounded.*

The proof of the proposition is similar to that of $\mathcal{T}_R$ and is omitted for brevity. Thus, the boundary condition on $\mathcal{B}_R$ can be defined as

$$u^s(R,\theta) = \mathcal{S}_R\left(\frac{\partial u}{\partial r}\right) - \mathcal{S}_R\left(\frac{\partial u^i}{\partial r}\right) - \mathcal{S}_R\left(\frac{\partial u^r}{\partial r}\right) + \Psi_R(\tilde{u}^s)(r,\theta),$$

where Ψ_R is the restriction of Ψ on $\mathcal{B}_R$. Consequently, by the continuity condition on $\mathcal{B}_R$ we have

$$\begin{aligned} u|_{r=R^-} &= u|_{r=R^+} = u^i|_{r=R} + u^r|_{r=R} + u^s|_{r=R} \\ &= u^i|_{r=R} + u^r|_{r=R} + \mathcal{S}_R\left(\frac{\partial u}{\partial r}\right) - \mathcal{S}_R\left(\frac{\partial u^i}{\partial r}\right) - \mathcal{S}_R\left(\frac{\partial u^r}{\partial r}\right) + \Psi_R(\tilde{u}^s). \end{aligned} \tag{3.36}$$

Thus, the boundary value problem for TE can be rewritten as

$$\begin{cases} -\nabla\cdot\left(\dfrac{1}{\varepsilon_r}\nabla u\right) + \alpha^2 u &= \alpha^2\tilde{u} \quad \text{in } \Omega_R, \\ \dfrac{\partial u}{\partial n} &= 0 \quad \text{on } S, \\ u - \mathcal{S}_R\left(\dfrac{\partial u}{\partial r}\right) &= u^i + u^r + \Psi_R(\tilde{u}^s) - \mathcal{S}_R\left(\dfrac{\partial u^i}{\partial r} + \dfrac{\partial u^r}{\partial r}\right) \quad \text{on } \mathcal{B}_R. \end{cases} \tag{3.37}$$

The next subsection is devoted to the variational formulation of (3.37).

3.2.2. Variational formulation. For convenience, we set

$$v = u - (u^i + u^r + \hat{\Psi}(\tilde{u}^s)) \quad \text{in } \Omega_R, \tag{3.38}$$

where $\hat{\Psi}(\tilde{u}^s)$ is the extension of $\Psi(\tilde{u}^s)$ into Ω_R. By direct computation, we can show that

$$\mathcal{S}_R\left(\frac{\partial \Psi_R}{\partial r}(\tilde{u}^s)\right) = 0 \quad \text{on } \mathcal{B}_R.$$

Hence the boundary value problem (3.37) can be written in terms of v as

$$\begin{cases} -\nabla\cdot\left(\dfrac{1}{\varepsilon_r}\nabla v\right) + \alpha^2 v &= \alpha^2\tilde{u} + \phi \quad \text{in } \Omega_R, \\ \dfrac{\partial v}{\partial n} &= -h \quad \text{on } S, \\ [1.5ex]v - \mathcal{S}_R\left(\dfrac{\partial v}{\partial r}\right) &= 0 \quad \text{on } \mathcal{B}_R, \end{cases} \tag{3.39}$$

where

$$\phi = \nabla\cdot\left[\frac{1}{\varepsilon_r}\nabla(u^i + u^r + \hat{\Psi}(\tilde{u}^s))\right] - \alpha^2(u^i + u^r + \hat{\Psi}(\tilde{u}^s)),$$

and

$$h = \frac{\partial(u^i + u^r + \hat{\Psi}(\tilde{u}^s))}{\partial n} \quad \text{on } S.$$

Hence we can define the variational space

$$W = \left\{ w \in H^1(\Omega_R) : w|_{\mathcal{B}_R} = \mathcal{S}_R\left(\frac{\partial w}{\partial r}\right) \text{ on } \mathcal{B}_R \right\} \tag{3.40}$$

with H^1-norm. The variational form of the boundary value equation (3.39) is to find $v \in W$ such that

$$b_{TE}(v, w) = F(w) \quad \forall w \in W, \tag{3.41}$$

where

$$b_{TE}(v, w) = \int_{\Omega_R}\left(\frac{1}{\varepsilon_r}\nabla v\cdot\nabla w + \alpha^2 uw\right)\, dxdy - \int_{B_R}\frac{\partial v}{\partial r}\mathcal{S}_R\left(\frac{\partial w}{\partial r}\right)\, d\theta, \tag{3.42}$$

and

$$F(w) = \int_{\Omega_R}(\alpha^2\tilde{u} + \phi)w\, dxdy - \int_S hw dS.$$

THEOREM 3.4. *The variational problem* (3.41) *has a unique solution in* W *defined by* (3.40) *and*

$$\|u\|_W \le C(\|\tilde{u}\|_{L^2(\Omega_R)} + \|u^i\|_W + \|u^r\|_W + \|\tilde{u}^s\|_W).$$

Proof. Since $\langle \mathcal{S}_R g, g\rangle \le 0$ and $\mathcal{S}_R : H^{-1/2}(\mathcal{B}_R) \to H^{1/2}(\mathcal{B}_R)$ is bounded, the symmetric bilinear form $a(u, w)$ in (3.41) is coercive and bounded. Hence the existence and uniqueness follow from the Lax–Milgram theorem. ∎

The time-marching procedure is the same as in the TM case.

4. Fully discrete problem. In this section, we consider the variational problems (3.21) and (3.41) by using finite element methods. Finite element error estimates and the stability of the Newmark scheme of the TM problem are presented. The arguments for the TE case, which are similar, are omitted here.

4.1. Finite element error analysis. Assume that Ω_R is covered by a family of quasi-uniform triangular meshes τ_h, where h is the mesh size, that is,

$$h = \max_{K\in\tau_h} h_K,$$

where h_K is the diameter of the element $K \in \tau_h$.

In the TM case, we consider the finite-dimensional subspace

$$V_h = \{v_h \in H^1(\Omega_R) : v_h|_K \text{ is linear }, K \in \tau_h\}.$$

We note that V_h is closed in V and $V_h \to V$ as $h \to 0$. The fully discrete problem is to find $u_h^n \in V_h$, $n = 1, 2, \ldots, \mathcal{N}$, such that

$$b(u_h^n, v_h) = F^n(v_h) \quad \forall v_h \in V_h, \tag{4.1}$$

where $b(u_h^n, v_h)$ and $F^n(v_h)$ are as defined in (3.22) and (3.23), respectively. Here, the subscript TM in $b(\cdot,\cdot)$ is omitted for brevity. We recall that the bilinear form b is coercive and continuous. Hence by Céa's lemma [8, pp. 36–69], the fully discrete problem (4.1) has a unique solution $u_h^n \in V_h$ and

$$\|u^n - u_h^n\|_V \le C \inf_{v_h\in V_h} \|u^n - v_h\|_V. \tag{4.2}$$

Since ε_r is discontinuous in Ω_R, the solution $u^n \notin H^2(\Omega_R)$, the inequality (4.2) does not yield a convergence rate in terms of h. In fact, since $V_h \to V$ for any $\epsilon > 0$, there is an $h_0 = h_0(\epsilon, u^n)$ such that for $0 < h < h_0$ there exists $v_h \in V_h$ satisfying

$$\|u^n - v_h\|_V \le \epsilon.$$

By (4.2), we have

$$\|u^n - u_h^n\|_V \le C\epsilon \quad \forall h < h_0(\epsilon, u^n).$$

Thus, the finite element solution u_h^n converges to u^n in V but not uniformly. We have the following.

THEOREM 4.1. *Let $u^n \in V$ and $u_h^n \in V_h$ be the solutions to* (3.21) *and* (4.1), *respectively, for $F^n \in V'$. Then given $\epsilon > 0$, there is an $h_0 = h_0(\epsilon)$ such that for all $0 < h < h_0$ we have*

$$\|u^n - u_h^n\|_{L^2(\Omega_R)} \le \epsilon \|u^n - u_h^n\|_V. \tag{4.3}$$

Furthermore, if $\varepsilon_r \in L^\infty(\Omega_R)$, hence $\varepsilon_r \tilde{u}^n \in L^2(\Omega_R)$, then there exists an $h_1 = h_1(\epsilon) > 0$ such that for all $0 < h < h_1$ we have

$$\|u^n - u_h^n\|_V \le C\epsilon \|F^n\|_{L^2(\Omega_R)}, \tag{4.4}$$

where C is a positive constant independent of h. Consequently, we have

$$\|u^n - u_h^n\|_{L^2(\Omega)} \le C\epsilon^2 \|F^n\|_{L^2(\Omega_R)}.$$

We first consider the following lemma.

LEMMA 4.2. *Let Λ be the set of solutions $w \in V$ to*

$$b(w, v) = (\psi, v) \quad \forall v \in V, \tag{4.5}$$

where $\|\psi\|_{L^2(\Omega_R)} = 1$. Then Λ is compact in V.

Proof. Since $w \in V$ is the solution to (4.5), it satisfies

$$\|w\|_V \leq C\|\psi\|_{L^2(\Omega_R)}.$$

Thus, the solution mapping $G : \psi \to G\psi = w$ is continuous from the dual space V' to $V \subset H^1(\Omega_R)$. Furthermore, the embedding, $I : \ L^2(\Omega_R) \subset V'$, is compact. It implies that $\Lambda \subset G \circ I(\{\psi \in L^2(\Omega_R) : \|\psi\|_{L^2(\Omega_R)} = 1\})$ is compact in V. ∎

We now prove the theorem.

Proof. By viewing $u^n - u_h^n$ as a linear functional in $L^2(\Omega_R)$, we have

$$\|u^n - u_h^n\|_{L^2(\Omega_R)} = \sup_{\|\psi\|_{L^2(\Omega_R)}=1} (u^n - u_h^n, \psi).$$

Let $w \in V$ be the solution to

$$b(v, \eta) = (\psi, \eta) \quad \forall \eta \in V.$$

Then

$$\|w\|_V \leq C\|\psi\|_{L^2(\Omega_R)}.$$

Thus, for $v_h \in V_h$, by the boundedness of the bilinear form $b(\cdot, \cdot)$, we have

$$\begin{aligned} |(u^n - u_h^n, \psi)| &= |b(u^n - u_h^n, w)| = |b(u^n - u_h^n, w - v_h)| \\ &\leq C\|u^n - u_h^n\|_V \|w - v_h\|_V. \end{aligned}$$

By the density property of V_h in V, we can choose v_h such that $\|w - v_h\|_V \leq \epsilon\|w\|_V$. We then obtain

$$|(u^n - u_h^n, \psi)| \leq C\epsilon\|u^n - u_h^n\|_V \|w\|_V \leq C\epsilon\|u^n - u_h^n\|_V \|\psi\|_{L^2(\Omega_R)}.$$

Thus

$$\|u^n - u_h^n\|_{L^2(\Omega_R)} \leq C\epsilon\|u^n - u_h^n\|_V.$$

This proves the estimate (4.3).

Next, we set

$$\hat{F}^n = \frac{F^n}{\|F^n\|_{L^2(\Omega_R)}}, \quad \hat{u}^n = \frac{u^n}{\|F^n\|_{L^2(\Omega_R)}}, \quad \hat{u}_h^n = \frac{u_h^n}{\|F^n\|_{L^2(\Omega_R)}}.$$

Then, we have

$$\begin{aligned} b(\hat{u}^n, v) &= \hat{F}^n(v) \quad \forall v \in V, \\ b(\hat{u}_h^n, v_h) &= \hat{F}^n(v_h) \quad \forall v_h \in V. \end{aligned}$$

By Céa's lemma,

$$\|\hat{u}^n - \hat{u}_h^n\|_V \leq C \inf_{v_h \in V_h} \|\hat{u}^n - v_h\|_V.$$

Since the set $\hat{\Lambda} = \{\hat{u}^n : b(\hat{u}^n, \phi) = \hat{F}^n(\phi), \|\hat{F}^n\|_{L^2(\Omega_R)} = 1\}$ is compact in V, we have, for $0 < h < h_0(\epsilon)$,

$$\inf_{v_h \in V_h} \|\hat{u}^n - v_h\|_V \leq \epsilon.$$

Thus

$$\|\hat{u}^n - \hat{u}_h^n\|_V \leq C\epsilon,$$

which implies that

$$\|u^n - u_h^n\|_V \leq C\epsilon \|F^n\|_{L^2(\Omega_R)}.$$

This completes the proof. ∎

4.2. Stability analysis. For stability analysis, we express the Newmark scheme in a three-step formulation. We start with

$$\begin{aligned} -\Delta u^{n+2} + \alpha^2 \varepsilon_r u^{n+2} &= \alpha^2 \varepsilon_r \tilde{u}^{n+2} \\ &= \alpha^2 \varepsilon_r \left[u^{n+1} + \Delta t \dot{u}^{n+1} + \frac{\Delta t^2}{2}(1-2\beta)\ddot{u}^{n+1} \right]. \end{aligned}$$

Using (3.5) to remove $\dot{u}^{n+1}$ and then (3.2) to remove $\tilde{\ddot{u}}^{n+1}$, we obtain

$$\begin{aligned} &-\Delta u^{n+2} + \alpha^2 \varepsilon_r u^{n+2} \\ &\quad = \alpha^2 \varepsilon_r \left[u^{n+1} + \Delta t(\dot{u}^n + \Delta t(1-\gamma)\ddot{u}^n) + \Delta t^2 \gamma + \frac{\Delta t^2}{2}(1-2\beta)\ddot{u}^{n+1} \right]. \end{aligned}$$

Using (3.1) to remove $\dot{u}^n$ and then (3.4) to remove $\ddot{u}^{n+1}$ and $\ddot{u}^n$ separately, and finally applying (3.3) to remove $\tilde{u}^{n+1}$ and $\tilde{u}^n$ separately, we obtain

$$\begin{aligned} &-\beta \Delta u^{n+2} - (\tfrac{1}{2} - 2\beta + \gamma)\Delta u^{n+1} - (\tfrac{1}{2} + \beta - \gamma)\Delta u^n \\ &+ \beta \alpha^2 \varepsilon_r (u^{n+2} - 2u^{n+1} + u^n) = 0. \end{aligned}$$

Adapting u_h^n for $u^n \in V_h$, we have the following variational form of the above equation:

$$\begin{aligned} &\tfrac{1}{\Delta t^2}\Big(\varepsilon_r(u_h^{n+2} - 2u_h^{n+1} + u_h^n), v_h\Big) \\ &+ a\Big(\beta u_h^{n+2} + (\tfrac{1}{2} - 2\beta + \gamma)u_h^{n+1} + (\tfrac{1}{2} + \beta - \gamma)u_h^n, v_h\Big) \\ &= \beta G^{n+2}(v_h) + (\tfrac{1}{2} - 2\beta + \gamma)G^{n+1}(v_h) + (\tfrac{1}{2} + \beta - \gamma)G^n(v_h) \quad \forall v_h \in V_h, \end{aligned} \tag{4.6}$$

where

$$a(u_h^n, v_h) = \int_{\Omega_R} \nabla u_h^n \cdot \nabla v_h \, dx - \int_{\mathcal{B}_R} \mathcal{T}_R(u_h^n) v_h \, d\theta,$$

and

$$G^n(v_h) = \int_{\mathcal{B}_R} H^n v_h d\theta - \int_{\mathcal{B}_R} \Phi_R(\tilde{u}_h^{s,n}) \, v_h \, d\theta.$$

It's clear that the bilinear form $a(u_h^n, v_h)$ is symmetric and coercive. Thus the eigenvalue problem

$$a(w_h, v_h) = \lambda_h(w_h, v_h) \quad \forall v_h \in V_h \tag{4.7}$$

has positive eigenvalues and corresponding orthonormal eigenvectors:

$$0 < \lambda_{h,1} \le \lambda_{h,2} \le \cdots \le \lambda_{h,M} < \infty,$$

$$w_{h,1}, w_{h,2}, \ldots, w_{h,M},$$

where $\dim V_h = M$.

Without confusion, we write $w_i = w_{h,i}$ and $\lambda_i = \lambda_{h,i}$. Substituting w_i for v_h in (4.6), noting that

$$a(u_h^n, w_i) = a(w_i, u_h^n) = \lambda_i(w_i, u_h^n) = \lambda_i(u_h^n, w_i)$$

yields

$$\begin{aligned}
&\tfrac{1}{\Delta t^2}(\varepsilon_r(u^{n+2} - 2u^{n+1} + u_h^n), w_i) \\
&+\lambda_i(\beta u_h^{n+1} + (\tfrac{1}{2} - 2\beta + \gamma)u_h^{n+1} + (\tfrac{1}{2} + \beta - \gamma)u_h^n, w_i) \\
&= \beta G^{n+2}(w_i) + (\tfrac{1}{2} - 2\beta + \gamma)G^{n+1}(w_i) + (\tfrac{1}{2} + \beta - \gamma)G^n(w_i) \\
&\equiv \Psi^n.
\end{aligned} \tag{4.8}$$

We observe that ε_r can be considered a constant and hence, without loss of generality, let $\varepsilon_r = 1$. Indeed, for $1 < \varepsilon_r \in L^\infty(\Omega_R)$, we may consider the weighted space $L^2(\Omega_R, \varepsilon_r)$ with the inner product

$$(u, v)_{\varepsilon_r} = (\varepsilon_r u, v) = (u, \varepsilon_r v).$$

We then consider the following eigenvalue equation: find λ_h and $u_h \in V_h$ such that

$$a(u_h, v_h) = \lambda_h(u_h, v_h)_{\varepsilon_r} \quad \forall v_h \in V_h. \tag{4.9}$$

Since $a(\cdot, \cdot)$ is symmetric and coercive, (4.9) has positive eigenvalues λ_i and corresponding orthonormal eigenvectors w_i, $i = 1, 2, \ldots, M$, such that

$$(w_i, w_j)_{\varepsilon_r} = \delta_{ij}.$$

Hence, by substituting $u_h^n = \sum_{i=1}^M u_i^n w_{h,i}$ into (4.8), we obtain for $i = 1, 2, \ldots, M$,

$$\tfrac{1}{\Delta t^2}(u_i^{n+2} - 2u_i^{n+1} + u_i^n) + \lambda_i\left(\beta u_i^{n+2} + \left(\tfrac{1}{2} - 2\beta + \gamma\right)u_i^{n+1} + \left(\tfrac{1}{2} + \beta - \gamma\right)u_i^n\right) = 0, \tag{4.10}$$

that is,

$$\begin{aligned}
u_i^{n+2} &= \frac{\frac{2}{\Delta t^2} - \lambda_i(\frac{1}{2} - 2\beta + \gamma)}{\frac{1}{\Delta t^2} + \lambda_i\beta}u_i^{n+1} - \frac{\frac{1}{\Delta t^2} + \lambda_i(\frac{1}{2} + \beta - \gamma)}{\frac{1}{\Delta t^2} + \lambda_i\beta}u_i^n \\
&\equiv \eta u_i^{n+1} - \kappa u_i^n.
\end{aligned}$$

Thus (4.10) can be written in a matrix form as

$$\begin{pmatrix} u_i^{n+2} \\ u_i^{n+1} \end{pmatrix} = \begin{pmatrix} \eta & -\kappa \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u_i^{n+1} \\ u_i^n \end{pmatrix} \equiv B \begin{pmatrix} u_i^{n+1} \\ u_i^n \end{pmatrix}.$$

By denoting

$$X_i^n = \begin{pmatrix} u_i^{n+1} u_i^n \end{pmatrix}$$

for $n = 1, 2, \ldots, N$, $i = 1, 2, \ldots, M$, we obtain the recursive relation

$$X_i^{n+1} = B(\lambda_i) X_i^n,$$

or equivalently

$$X_i^{n+1} = B^n X_i^0. \tag{4.11}$$

For stability analysis we wish to establish conditions on β, γ, and Δt such that $|X_i^n| = (|u_i^{n+1}|^2 + |u_i^n|^2)^{1/2}$ for all i, and hence $|u^n| = (\sum_{i=1}^M |u_i^n|^2)^{1/2}$, is bounded independent of n.

We observe that, if B is diagonalizable with the spectral radius $\rho(B) \leq 1$, then

$$|X^n| = |B^n X^0| \leq \|G^{-1}\| \rho^n(B) \|G\| |X^0| \leq C$$

for some matrix G. For simplicity, we seek conditions on β, γ, and Δt such that B has distinct eigenvalues (hence diagonalizable) of lengths less than 1.

We shall assume that $\beta \geq 0$. We consider the characteristic equation of B:

$$\det(\mu I - B) = \mu^2 - \mu\eta + \kappa = 0.$$

The solutions μ_1, μ_2 are

$$\mu_{1,2} = \frac{\eta \pm \sqrt{\eta^2 - 4\kappa}}{2}.$$

We consider the following two cases.

Case 1. Suppose $\Delta = \eta^2 - 4\kappa < 0$. Then $\mu_1 = \bar{\mu}_2$ and $|\mu_1| = |\mu_2| = \sqrt{\kappa}$. Thus we require $\kappa \leq 1$, which implies that $\gamma \geq \frac{1}{2}$. We have

$$\Delta = \left[2 - \lambda_i \Delta t^2 \left(\frac{1}{2} - 2\beta + \gamma\right)\right]^2 - 4(1 + \lambda_i \Delta t^2 \beta)\left[1 + \lambda_i \Delta t^2 \left(\frac{1}{2} + \beta - \gamma\right)\right] < 0,$$

which is

$$-4\lambda_i \Delta t^2 + (\lambda_i \Delta t^2)^2 [(1+\gamma)^2 - 4\beta] < 0,$$

or

$$\lambda_i \Delta t^2 \left[\frac{1}{4}\left(\frac{1}{2} + \gamma\right)^2 - \beta\right] < 1,$$

or equivalently

$$\frac{1}{4}\left(\frac{1}{2} + \gamma\right)^2 - \beta < \frac{1}{\lambda_i \Delta t^2} \quad \forall i = 1, 2, \ldots, M.$$

Case 2. Suppose $\Delta > 0$, that is,

$$\frac{1}{4}\left(\frac{1}{2} + \gamma\right)^2 - \beta > \frac{1}{\lambda_i \Delta t^2} \quad \forall i = 1, 2, \ldots, M. \tag{4.12}$$

Without loss of generality, $\mu_1 < \mu_2$. Let

$$-1 \le \mu_1 = \frac{\eta}{2} - \frac{\sqrt{\Delta}}{2} < \frac{\eta}{2} + \frac{\sqrt{\Delta}}{2} = \mu_2 \le 1.$$

The inequality $\mu_1 \ge -1$ implies $1 + \eta + \kappa \ge 0$, or

$$\frac{\gamma}{2} - \beta \le \frac{1}{\lambda_i \Delta t^2}. \tag{4.13}$$

$\mu_2 \le 1$ implies $1 - \eta + \kappa \ge 0$. So we require $\kappa \ge -1$, which implies

$$\frac{1}{2}\left(\gamma - \frac{1}{2} - 2\beta\right) \le \frac{1}{\lambda_i \Delta t^2}. \tag{4.14}$$

By combining (4.13) and (4.14), we have

$$\frac{\gamma}{2} - \beta - \frac{1}{4} \le \frac{1}{\lambda_i \Delta t^2}. \tag{4.15}$$

However, the inequalities (4.12) and (4.15) are inconsistent, so we ignore Case 2.

Thus, $X^n = B^n X^0$ is stable if

$$\gamma \ge \frac{1}{2} \quad \text{and} \quad \frac{1}{4}\left(\frac{1}{2} + \gamma\right)^2 - \beta < \frac{1}{\lambda_i \Delta t^2}, \quad i = 1, 2, \ldots, M. \tag{4.16}$$

We summarize the above analysis in the following theorem.

THEOREM 4.3. *The Newmark scheme for the TM variational problem is stable if* $\gamma \ge \frac{1}{2}$, $\beta \ge 0$, *and*

$$\frac{1}{4}\left(\frac{1}{2} + \gamma\right)^2 - \beta < \frac{1}{\lambda_i \Delta t^2}, \quad i = 1, 2, \ldots, M, \tag{4.17}$$

where $\lambda_{h,i}$ *are the eigenvalues of* $a(w, v_h) = \lambda_h (w, v_h)$ *for all* $v_h \in V_h$.

Remark 4.4. By assuming the value $\gamma = \frac{1}{2}$, one achieves an $\mathcal{O}(\Delta t^2)$ error rate, but this is not always the best value to use. The finite element discretization of the problem tends to create a stiff system of ordinary differential equations. Standard theory in finite elements indicate that the modes corresponding to the higher frequencies become more and more inaccurate as we move up the spectrum (see [24, pp. 244–256] and [31, pp. 63–65]). In practice, a value of γ larger than $\frac{1}{2}$ is often used in order to damp out the higher frequencies while preserving the more accurate lower ones.

5. Conclusion. In this paper, the two-dimensional time-dependent TM and TE scattering problems of an overfilled cavity in the infinite ground plane are considered. In each case, the problem was first discretized in time by the Newmark time-marching scheme. At each time step $t = t_n$, the partial differential equations defined in an infinite domain are solved. Transparent boundary conditions are constructed using the pseudodifferential operators $\mathcal{T}_R$ for TM and $\mathcal{S}_R$ for TE to reduce the computational domain to the bounded region Ω_R. Variational formulations for both polarizations are derived. Existence and uniqueness of both the semidiscrete and the fully discrete solutions, u^n and u_h^n, are obtained. Error estimates in both the H^1-norm, $\|u^n - u_h^n\|_{H^1}$, and L^2-norm, $\|u^n - u_h^n\|_{L^2}$, $n = 0, 1, \ldots, \mathcal{N}$, are achieved. Stability criteria for

the time-marching scheme are established. In particular, the scheme is shown to be unconditionally stable if $\gamma \geq \frac{1}{2}$ and

$$\beta > \frac{1}{4}\left(\frac{1}{2} + \gamma\right)^2.$$

REFERENCES

[1] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 1972.

[2] H. Ammari, G. Bao, and A. Wood, *A cavity problem for Maxwell's equations*, Meth. Math. Appl., 9 (2002), pp. 249–260.

[3] H. Ammari, G. Bao, and A. Wood, *An integral equation method for the electromagnetic scattering from cavities*, Math. Methods Appl. Sci., 23 (2000), pp. 1057–1072.

[4] H. Ammari, G. Bao, and A. Wood, *Analysis of the electromagnetic scattering from a cavity*, Japan J. Indust. Appl. Math, 19 (2002), pp. 301–308.

[5] A. Barka, P. Soudais, and D. Volpert, *Scattering from 3-d cavities with a plug and play numerical scheme combining ie, pde, and modal techniques*, IEEE Trans. Antennas Propagation, 48 (2000), pp. 704–712.

[6] R. J. Burkholder, *Two ray shooting methods for computing the EM scattering by large open-ended cavities*, Comput. Phys. Comm., 68 (1991), pp. 353–365.

[7] T. T. Chia, R. J. Burkholder, and R. Lee, *The application of fdtd in hybrid methods for cavity scattering analysis*, IEEE Trans. Antennas Propagation, 43 (1995), pp. 1082–1090.

[8] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, 1978.

[9] T. B. Hansen and A. D. Yaghjian, *Low-frequency scattering from two-dimensional perfect conductors*, IEEE Trans. Antennas Propagation, 40 (1992), pp. 1389–1402.

[10] F. B. Hildebrand, *Advanced Calculus for Applications*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1976.

[11] D. Jiao, A. Ergin, B. Shanker, E. Michelssen, and J. Jin, *A fast higher-order time-domain finite element boundary integral method for 3-D electromagnetic scattering analysis*, IEEE Trans. Antennas Propagation, 50 (2002), pp. 1192–1202.

[12] D. Jiao, M. Lu, E. Michielssen, and J. Jin, *A fast time-domain finite element-boundary integral method for electromagnetic analysis*, IEEE Trans. Antennas Propagation, 49 (2001), pp. 1453–1461.

[13] J. M. Jin, *Electromagnetic scattering from large, deep, and arbitrarily-shaped open cavities*, Electromagnetics, 18 (1998), pp. 3–34.

[14] J. M. Jin, S. Ni, and S. W. Lee, *Hybridization of sbr and fem for scattering by large bodies with cracks and cavities*, IEEE Trans. Antennas Propagation, 43 (1995), pp. 1130–1139.

[15] J. M. Jin and J. L. Volakis, *A hybrid finite element method for scattering and radiation by microstrip patch antennas and arrays residing in a cavity*, IEEE Trans. Antennas Propagation, 39 (1991), pp. 1598–1604.

[16] H. Ling, R.-C. Chou, and S.-W. Lee, *Shooting and bouncing rays: Calculating the RCS of an arbitrarily shaped cavity*, IEEE Trans. Antennas Propagation, 37 (1989), pp. 194–205.

[17] J. Liu and J. M. Jin, *Scattering analysis of a large body with deep cavities*, IEEE Antennas and Propagation Society International Symposium, San Antonio, TX, 2002.

[18] J. Liu and J. M. Jin, *A special higher order finite-element method for scattering by deep cavities*, IEEE Trans. Antennas Propagation, 48 (2000), pp. 694–703.

[19] C. C. Lu and W. C. Chew, *A near-resonance decoupling approach (nrda) for scattering solution of near-resonant structures*, IEEE Trans. Antennas Propagation, 45 (1997), pp. 1857–1862.

[20] P. K. Murthy, K. C. Hill, and G. A. Thiele, *A hybrid-iterative method for scattering problems*, IEEE Trans. on Antennas Propagation, 34 (1986), pp. 1173–1180.

[21] D. C. Ross, J. L. Volakis, and H. T. Anastassiu, *Hybrid finite element-modal analysis of jet engine inlet scattering*, IEEE Trans. Antennas Propagation, 43 (1995), pp. 277–285.

[22] D. C. Ross, J. L. Volakis, and H. T. Anastassiu, *Three-dimensional edge-based finite element analysis for discrete bodies of revolution*, IEEE Trans. Antennas Propagation, 45 (1997), pp. 1160–1165.

[23] D. H. Reuster and G. A. Thiele, *A field iterative method for computing the scattered electric fields at the apertures of large perfectly conducting cavities*, IEEE Trans. Antennas Propagation, 43 (1995), pp. 286–290.

[24] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, 2nd ed., Wellesley–Cambridge Press, New York, 1988.
[25] T. Van and A. W. Wood, *Analysis of time-domain Maxwell's equations for* 3-*d cavities*, Adv. Comput. Math., 16 (2002), pp. 211–228.
[26] T. Van and A. W. Wood, *Time-domain finite element method for Helmholtz equations*, J. Comput. Phys., 183 (2002), pp. 486–507.
[27] T. Van and A. W. Wood, *A time-marching finite element method for an electromagnetic scattering problem*, Math Methods Appl. Sci., 26 (2003), pp. 1025–1045.
[28] T. Van and A. W. Wood, *Finite element analysis of electromagnetic scattering from a cavity*, IEEE Trans. Antennas Propagation, 51 (2003), pp. 130–137.
[29] T. M. Wang and H. Ling, *Electromagnetic scattering from three-dimensional cavities via a connection scheme*, IEEE Trans. Antennas Propagation, 39 (1991), pp. 1505–1513.
[30] W. D. Wood and A. W. Wood, *Development and numerical solution of integral equations for electromagnetic scattering from a trough in a ground plane*, IEEE Trans. Antennas Propagation, 47 (1999), pp. 1318–1322.
[31] W. L. Wood, *Practical Time-Stepping Schemes*, Clarendon Press, London, 1990.
[32] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method. Solid and Fluid Mechanics. Dynamics and Non-linearity*, Vol. II, 4th ed., McGraw–Hill, New York, 1991.

SIAM J. APPL. MATH.
Vol. 64, No. 2, pp. 709–723

THE DETERMINATION OF THE SURFACE IMPEDANCE OF A PARTIALLY COATED OBSTACLE FROM FAR FIELD DATA*

FIORALBA CAKONI† AND DAVID COLTON†

Abstract. A variational method is given for determining the essential supremum of the surface impedance of a partially coated perfect conductor from a knowledge of the far field pattern of the time-harmonic electric field at fixed frequency. It is assumed that the shape of the scatterer has been determined (e.g., by solving the far field equation and using the linear sampling method). Numerical examples are given for the scalar case with constant surface impedance.

Key words. inverse scattering problem, impedance boundary condition, electromagnetic waves

AMS subject classifications. 35P25, 35R30, 78A45

DOI. 10.1137/S0036139903424254

1. Introduction. In order to avoid detection by radar, hostile objects are often partially coated by a material designed to reduce the radar cross section of the scattered wave. From the point of view of target identification a key question to answer is, given the shape of a scattering obstacle (which can be determined, for example, by the linear sampling method [2], [3]), is the obstacle coated or not and if so what are the electrical properties of the coating? The simplest example of such a problem is the case of a perfect conductor that is partially coated by a dielectric. In this case the direct scattering problem is a mixed boundary value problem for Maxwell's equations where on the coated part of the boundary the electromagnetic field satisfies an impedance boundary condition [9], [12] and on the remaining part of the boundary the tangential component of the total electric field vanishes. The inverse problem of determining whether or not the obstacle is coated, and, if so, what the values of the surface impedance are, is complicated by the fact that the extent of the coating (if indeed the object is coated at all!) is not known a priori.

In this paper we will provide a variational method for determining the essential supremum of the surface impedance (which may be zero if the scatterer is not coated!) from a knowledge of the far field pattern of the scattered electric field corresponding to a time-harmonic incident plane wave at fixed frequency. In the special case where the surface impedance is a constant, this of course yields this constant. However, in neither case does our method provide information on how much of the scattering obstacle is coated. (In particular, there could be no coating at all or the entire obstacle could be coated!) Our analysis is based on our recent investigations of the inverse scattering problem for partially coated obstacles where the aim was to determine the shape of the scattering obstacle with unknown boundary condition from a knowledge of the electric far field pattern [2], [3]. As we show in this paper, the far field equation that was used in [2] and [3] to determine the shape can also be used in conjunction with a variational method to determine the essential supremum of the surface impedance on the coated portion of the boundary. Although for the sake of exposition we assume in

*Received by the editors March 12, 2003; accepted for publication (in revised form) July 30, 2003; published electronically January 30, 2004. The authors gratefully acknowledge the support of their research by the Air Force Office of Scientific Research under grants F49620-02-1-0071 and F49620-02-1-0353.

http://www.siam.org/journals/siap/64-2/42425.html

†Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (cakoni@math.udel.edu, colton@math.udel.edu).

this paper that we have full-aperture far field data, we point out at the end of section 3 how all of our results remain valid for the practical case of limited-aperture data.

Given the shape of the scattering obstacle, the problem of determining lower bounds for the surface impedance in the scalar case when the obstacle is completely coated has previously been considered by Colton and Kress [6] (full-aperture scattering data) and Colton and Piana [8] (limited-aperture scattering data). In particular, the paper of Colton and Piana has had a strong influence on the approach used in the present paper. We also draw the reader's attention to a recent paper of Akduman and Kress [1], where a potential theoretic method is given for determining the surface impedance in the case when the shape of the scatterer is known and the obstacle is completely coated.

The plan of our paper is as follows. We first consider the scattering of time-harmonic plane waves by a partially coated infinite cylinder (which in fact can be totally coated, partially coated or not coated at all). This leads to the investigation of a mixed boundary value problem for the two-dimensional Helmholtz equation in the exterior of a bounded domain D with Lipschitz boundary Γ. Assuming the surface impedance $\lambda = \lambda(x)$ on the coated portion Γ_I of Γ is in $L_\infty(\Gamma_I)$, we derive a variational method for determining ess sup $\lambda(x)$ from a knowledge of the far field pattern of the scattered wave. We then extend this result to the case of Maxwell's equations in $\mathbb{R}^3$. In the final section of our paper we consider several numerical examples in the scalar case when the surface impedance is a constant.

2. The scalar case. We consider the scattering of an electromagnetic time harmonic plane wave by a perfectly conducting infinite cylinder that is (partially) coated by an inhomogeneous dielectric material. This leads to a mixed boundary value problem for the Helmholtz equation [2]. In particular let $D \subset \mathbb{R}^2$ be an open bounded region with Lipschitz boundary Γ such that $\mathbb{R}^2 \setminus \overline{D}$ is connected. We assume that the boundary Γ has a Lipschitz dissection $\Gamma = \Gamma_D \cup \Pi \cup \Gamma_I$, where Γ_D and Γ_I are disjoint, relatively open subsets of Γ, having Π as their common boundary in Γ (see e.g., [10]). Furthermore, boundary conditions of Dirichlet and impedance type with the surface impedance a bounded measurable function $\lambda \in L_\infty(\Gamma_I)$ are specified on Γ_D and Γ_I, respectively. We assume that the surface impedance is positive and uniformly bounded, i.e., $\lambda(x) \geq \lambda_0 > 0$ for $x \in \Gamma_I$. Let ν denote the unit outward normal vector defined almost everywhere on $\Gamma_D \cup \Gamma_I$. The total field $u = u^s + e^{ikx\cdot d}$ given as the sum of the unknown scattered wave and incident plane wave satisfies

$$\Delta u + k^2 u = 0 \qquad \text{in} \qquad \mathbb{R}^2 \setminus \overline{D}, \tag{2.1a}$$

$$u = 0 \qquad \text{on} \qquad \Gamma_D, \tag{2.1b}$$

$$\frac{\partial u}{\partial \nu} + i\lambda(x)u = 0 \qquad \text{on} \qquad \Gamma_I, \tag{2.1c}$$

where $k > 0$ is the wave number and d is a unit vector describing the incident direction. Moreover, the scattered field u^s satisfies the Sommerfeld radiation condition

$$\lim_{r\to\infty} \sqrt{r}\left(\frac{\partial u^s}{\partial r} - iku^s\right) = 0 \tag{2.2}$$

uniformly in $\hat{x} = x/|x|$ with $r = |x|$.

The well-posedness of the exterior mixed boundary value problem is established in [2] (in [2] λ was assumed to be constant, but all the results remain valid if $\lambda = \lambda(x) \in L_\infty(\Gamma_I)$). In particular it is shown that the direct scattering problem (2.1a)–(2.2) has a unique solution $u \in H_{loc}(D_e)$.

It is easy to see [5] that the scattered field has the asymptotic behavior

$$u^s(x) = \frac{e^{ikr}}{\sqrt{r}} u_\infty(\hat{x}, d) + O(r^{-3/2}), \tag{2.3}$$

where u_∞ is the *far field pattern* of the scattered wave. The far field pattern defines the far field operator $F : L^2(\Omega) \to L^2(\Omega)$ by

$$(Fg)(\hat{x}) := \int_\Omega u_\infty(\hat{x}, d) g(d) ds(d), \qquad g \in L^2(\Omega). \tag{2.4}$$

The corresponding interior mixed boundary value problem is also studied in [2]. In particular we consider the following problem: find $u_z \in H^1(D)$ that satisfies

$$\Delta u_z + k^2 u_z = 0 \qquad \text{in} \qquad D, \tag{2.5a}$$

$$u_z = -\Phi(\cdot, z) \qquad \text{on} \qquad \Gamma_D, \tag{2.5b}$$

$$\frac{\partial u_z}{\partial \nu} + i\lambda(x) u_z = -\frac{\partial \Phi(\cdot, z)}{\partial \nu} - i\lambda(x)\Phi(\cdot, z) \qquad \text{on} \qquad \Gamma_I \tag{2.5c}$$

for a fixed $z \in D$, where Φ is the fundamental solution to the Helmholtz equation defined by

$$\Phi(x, z) := \frac{i}{4} H_0^{(1)}(k|x - z|) \tag{2.6}$$

with $H_0^{(1)}$ being a Hankel function of the first kind of order zero. Then in [2] it is shown that (2.5a)–(2.5c) has a unique solution $u_z \in H^1(D)$ provided $\Gamma_I \neq \emptyset$ and $\lambda \neq 0$.

Next we introduce the far field equation

$$(Fg)(\hat{x}) = \gamma e^{-ik\hat{x}\cdot z}, \qquad g \in L^2(\Omega), \quad z \in D, \tag{2.7}$$

where $\gamma = \frac{e^{i\pi/4}}{\sqrt{8\pi k}}$ and $\gamma e^{-ik\hat{x}\cdot z}$ is the far field pattern of $\Phi(x, z)$.

A Herglotz wave function with kernel $g \in L^2(\Omega)$ is an entire solution of the Helmholtz equation defined by

$$v_g(x) = \int_\Omega e^{ikx\cdot d} g(d) ds(d), \qquad x \in \mathbb{R}^2.$$

The following theorem is proved in [2].

THEOREM 2.1. *Let $\epsilon > 0$, $z \in D$, and u_z be the unique solution of* (2.5a)–(2.5c). *Then there exists a Herglotz wave function $v_{g_\epsilon^z}$ with kernel $g_\epsilon^z \in L^2(\Omega)$ such that*

$$\|u_z - v_{g_\epsilon^z}\|_{H^1(D)} \leq \epsilon. \tag{2.8}$$

Moreover, there exists a positive constant $c > 0$ independent of ϵ such that

$$\|(Fg_\epsilon^z)(\hat{x}) - \gamma e^{-ik\hat{x}\cdot z}\|_{L^2(\Omega)} \leq c\epsilon. \tag{2.9}$$

Now let us define w_z by

$$w_z := u_z + \Phi(\cdot, z). \tag{2.10}$$

In particular, since $u_z \in H^1(D)$ and $z \in D$, we have that $w_z|_\Gamma \in H^{\frac{1}{2}}(\Gamma)$, $\frac{\partial w_z}{\partial \nu}|_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$, and

$$w_z|_{\Gamma_D} = 0 \qquad \text{and} \qquad \left(\frac{\partial w_z}{\partial \nu} + i\lambda w_z\right)|_{\Gamma_I} = 0 \tag{2.11}$$

interpreted in the sense of the trace theorem.

LEMMA 2.2. *For every two points z_1 and z_2 in D we have that*

$$2\int_{\Gamma_I} w_{z_1}\lambda(x)\overline{w}_{z_2}\,ds = -4k\pi|\gamma|^2 J_0(k|z_1 - z_2|) + i\left(u_{z_1}(z_2) - \overline{u}_{z_2}(z_1)\right), \tag{2.12}$$

where u_{z_1}, w_{z_1} and u_{z_2}, w_{z_2} are defined by (2.5a)–(2.5c) *and* (2.10)*, respectively, and J_0 is a Bessel function of order zero.*

Proof. Let z_1 and z_2 be two points in D and u_{z_1}, w_{z_1} and u_{z_2}, w_{z_2} the corresponding functions defined by (2.5a)–(2.5c) and (2.10). From (2.11) we have that

$$\begin{aligned}
2i\int_{\Gamma_I} w_{z_1}\lambda(x)\overline{w}_{z_2}\,ds &= \int_\Gamma \left(w_{z_1}\frac{\partial \overline{w}_{z_2}}{\partial \nu} - \overline{w}_{z_2}\frac{\partial w_{z_1}}{\partial \nu}\right) ds \\
&= \int_\Gamma \left(\Phi(\cdot,z_1)\frac{\partial \overline{\Phi(\cdot,z_2)}}{\partial \nu} - \overline{\Phi(\cdot,z_2)}\frac{\partial \Phi(\cdot,z_1)}{\partial \nu}\right) ds \\
&+ \int_\Gamma \left(u_{z_1}\frac{\partial \overline{\Phi(\cdot,z_2)}}{\partial \nu} - \overline{\Phi(\cdot,z_2)}\frac{\partial u_{z_1}}{\partial \nu}\right) ds \\
&+ \int_\Gamma \left(\Phi(\cdot,z_1)\frac{\partial \overline{u}_{z_2}}{\partial \nu} - \overline{u}_{z_2}\frac{\partial \Phi(\cdot,z_1)}{\partial \nu}\right) ds.
\end{aligned}$$

From Green's theorem applied to the radiating solution $\Phi(\cdot,z)$ of the Helmholtz equation in D_e and the uniformity of the asymptotic relation (2.3) we have (see [7])

$$\begin{aligned}
&\int_\Gamma \left(\Phi(\cdot,z_1)\frac{\partial \overline{\Phi(\cdot,z_2)}}{\partial \nu} - \overline{\Phi(\cdot,z_2)}\frac{\partial \Phi(\cdot,z_1)}{\partial \nu}\right) ds = -2ik\int_\Omega \Phi_\infty(\cdot,z_1)\overline{\Phi_\infty(\cdot,z_2)}ds \\
&= -2ik\int_\Omega |\gamma|^2 e^{-ik\hat{x}\cdot z_1}e^{ik\hat{x}\cdot z_2}\,ds = -4ik\pi|\gamma|^2 J_0(k|z_1 - z_2|).
\end{aligned}$$

Now from the representation formula for u_{z_1} and u_{z_2} we obtain

$$2i\int_{\Gamma_I} w_{z_1}\lambda(x)\overline{w}_{z_2}\,ds = -4ik\pi|\gamma|^2 J_0(k|z_1 - z_2|) + \overline{u}_{z_2}(z_1) - u_{z_1}(z_2).$$

Finally, dividing both sides of the above relation by i yields the result. ∎

In the following let us consider a ball $B_r \subset D$ of radius r contained in D and denoted by

$$\mathcal{W} := \left\{ f \in L^2(\Gamma_I) : \begin{array}{c} f = w_z|_{\Gamma_I} \text{ with } w_z = u_z + \Phi(\cdot,z), \\ z \in B_r \text{ and } u_z \text{ the solution of (2.5a)–(2.5c)} \end{array} \right\}.$$

Now we are ready to prove the main result of this section.

THEOREM 2.3. *Let $\lambda \in L_\infty(\Gamma_I)$ be the surface impedance of the scattering problem* (2.1a)–(2.2)*. Then*

$$\|\lambda\|_{L_\infty(\Gamma_I)} = \sup_{\substack{z_i\in B_r\\ \alpha_i\in\mathbb{C}}} \frac{\sum_{i,j}\alpha_i\overline{\alpha_j}\left[-4\pi k|\gamma|^2 J_0(k|z_i-z_j|)+i\left(u_{z_i}(z_j)-\overline{u}_{z_j}(z_i)\right)\right]}{2\|\sum_i \alpha_i\left(u_{z_i}+\Phi(\cdot;z_i)\right)\|^2_{L^2(\Gamma)}}, \tag{2.13}$$

where u_z is the solution to (2.5a)–(2.5c) *and the sums are arbitrary finite sums.*

Proof. First we show that $\mathcal{W}$ is complete in $L^2(\Gamma_I)$. To this end let φ be a function in $L^2(\Gamma_I)$ such that for every $z \in B_r$

$$\int_{\Gamma_I} w_z\varphi\, ds = 0.$$

Construct $v \in H^1(D)$ as the unique solution of the interior mixed boundary value problem [2]

$$\begin{aligned} \Delta v + k^2 v &= 0 \quad \text{in} \quad D,\\ v &= 0 \quad \text{on} \quad \Gamma_D,\\ \frac{\partial v}{\partial \nu} + i\lambda(x)v &= \varphi \quad \text{on} \quad \Gamma_I. \end{aligned}$$

Then for every $z \in B_r$, using the boundary conditions and the integral representation formula, we have that

$$\begin{aligned} 0 = \int_{\Gamma_I} w_z\varphi\, ds &= \int_{\Gamma_I} w_z\left(\frac{\partial v}{\partial\nu}+i\lambda v\right) ds = \int_{\Gamma} w_z\left(\frac{\partial v}{\partial\nu}+i\lambda v\right) ds\\ &= \int_\Gamma \left(u_z\frac{\partial v}{\partial\nu}+i\lambda u_z v+\Phi(\cdot,z)\frac{\partial v}{\partial\nu}+i\lambda\Phi(\cdot,z)v\right) ds\\ &= \int_\Gamma \left[u_z\frac{\partial v}{\partial\nu}+v\left(-\frac{\partial u_z}{\partial\nu}-\frac{\partial\Phi(\cdot,z)}{\partial\nu}-i\lambda\Phi(\cdot,z)\right)\right] ds\\ &+ \int_\Gamma \left(\Phi(\cdot,z)\frac{\partial v}{\partial\nu}+i\lambda v\Phi(\cdot,z)\right) ds = v(z). \end{aligned}$$

Now the unique continuation principle implies that $v(z) = 0$ for all $z \in D$, whence from the trace theorem $\varphi = 0$.

We now show that

$$\|\lambda\|_{L_\infty(\Gamma_I)} := \operatorname{ess\,sup}\lambda = \sup_{f\in L^2(\Gamma_I)} \frac{1}{\|f\|^2_{L^2(\Gamma_I)}}\int_{\Gamma_I}\lambda(x)|f|^2 ds.$$

The theorem then follows from Lemma 2.2 and the denseness of $\mathcal{W}$ in $L^2(\Gamma_I)$ by fixing first z_2 and then z_1 and considering linear combinations of w_z for different $z \in B_r$ together with the fact that $\|w_z\|_{L^2(\Gamma)} = \|w_z\|_{L^2(\Gamma_I)}$. (Note that w_{z_1} and w_{z_2} are not orthogonal with respect to $\lambda(x)$ and hence two different points are needed.) To prove the above identity, let $C = \operatorname{ess\,sup}\lambda > 0$. Obviously,

$$\frac{1}{\|f\|^2_{L^2(\Gamma_I)}}\int_{\Gamma_I}\lambda(x)|f|^2 ds \le C \qquad \forall f \in L^2(\Gamma_I).$$

Now for every $0 < \epsilon < C$ the set $M_\epsilon = \{x \in \Gamma_I : |\lambda(x)| \geq C - \epsilon\}$ has a positive measure and for an $f_\epsilon \in L^2(\Gamma_I)$ supported in M_ϵ we have

$$\frac{1}{\|f_\epsilon\|^2_{L^2(\Gamma_I)}} \int_{\Gamma_I} \lambda(x)|f_\epsilon|^2 ds \geq (C - \epsilon),$$

which ends the proof. □

Given that D is known (for example, by using the far field equation and the linear sampling method as discussed in [2]), u_z in the right-hand side of (2.13) still cannot be computed since it depends on the unknown function λ. However, from Theorem 2.1, we can use in (2.13) an approximation to u_z given by the Herglotz wave function v_{g^z} with kernel g^z being the (regularized) solutions of the far field equation (2.7).

In the particular case where the surface impedance is a positive constant $\lambda > 0$ we can further simplify the formula (2.13). In particular, fix an arbitrary point $z_0 \in B_r$ and consider $z_1 = z_2 = z_0$. Then (2.12) simply becomes

$$\lambda = \frac{-2k\pi|\gamma|^2 - \operatorname{Im}(u_{z_0}(z_0))}{\|u_{z_0} + \Phi(\cdot; z_0)\|^2_{L^2(\Gamma)}}. \tag{2.14}$$

Note that the expressions on the right-hand sides of (2.13) and (2.14) can be used as a target signature to detect if an obstacle is coated or not. In particular an object is coated if and only if the numerator is nonzero.

3. The vector case. We now turn our attention to the electromagnetic scattering problem for a (partially) coated perfect conductor in $\mathbb{R}^3$. In particular let $D \subset \mathbb{R}^3$ be a bounded region with boundary Γ such that $D_e := \mathbb{R}^3 \setminus \overline{D}$ is connected. Each simply connected piece of D is assumed to be a Lipschitz curvilinear polyhedron. Moreover, we assume that the boundary $\Gamma = \Gamma_D \cup \Pi \cup \Gamma_I$ is split into two disjoint parts Γ_D and Γ_I having Π as their possible common boundary in Γ and that each part Γ_D and Γ_I can be written as the union of a finite number of open smooth faces $(\Gamma_D^j)_{j=1,\ldots,N_D}$ and $(\Gamma_I^j)_{j=1,\ldots,N_I}$, respectively, where e_{ij} denotes the common edge of two adjacent faces Γ^i and Γ^j. Let ν denote the unit outward normal defined almost everywhere on Γ.

The direct scattering problem for the scattering of a time-harmonic electromagnetic plane wave by a partially coated obstacle D is to find an electric field E and a magnetic field $H := \frac{1}{ik}\operatorname{curl} E$ such that

$$\operatorname{curl}\operatorname{curl} E - k^2 E = 0 \quad \text{in} \quad \mathbb{R}^3 \setminus \overline{D}, \tag{3.1a}$$

$$\nu \times E = 0 \quad \text{on} \quad \Gamma_D, \tag{3.1b}$$

$$\nu \times \operatorname{curl} E - i\lambda(x)(\nu \times E) \times \nu = 0 \quad \text{on} \quad \Gamma_I, \tag{3.1c}$$

where the surface impedance $\lambda \in L_\infty(\Gamma_I)$ satisfies $\lambda(x) \geq \lambda_0 > 0$. The total electric field E is given by

$$E = E^i + E^s, \tag{3.2}$$

where E^s is the scattered field satisfying the Silver–Müller radiation condition

$$\lim_{r\to\infty} (\operatorname{curl} E^s \times x - ikrE^s) = 0 \tag{3.3}$$

uniformly in $\hat{x} = x/|x|$, where $r = |x|$ and the incident field E^i is given by

$$E^i(x) := \frac{i}{k}\operatorname{curl}\operatorname{curl} p e^{ikx\cdot d} = ik(d \times p) \times d e^{ikx\cdot d}, \tag{3.4}$$

where $k > 0$ is the wave number, d is a unit vector giving the direction of propagation, and p is the polarization vector. The well-posedness of the direct problem is established in [3] (in [3] λ was assumed to be constant, but all the results remain valid if $\lambda = \lambda(x) \in L_\infty(\Gamma_I)$). In particular it is shown that there exists a unique solution E, and $H = \frac{1}{ik}\operatorname{curl} E$ of (3.1a)–(3.4), and, moreover, $E \in X(D_e \cap B_R, \Gamma_I)$ for every ball of radius R containing D, where $X(D_e \cap B_R, \Gamma_I)$ is the Sobolev space defined by

$$X(D_e \cap B_R, \Gamma_I) := \{u \in H(\operatorname{curl}, D_e \cap B_R) \,:\, \nu \times u|_{\Gamma_I} \in L_t^2(\Gamma_I)\}$$

with

$$H(\operatorname{curl}, D_e \cap B_R) := \{u \in (L^2(D_e \cap B_R))^3 \,:\, \operatorname{curl} u \in (L^2(D_e \cap B_R))^3\},$$

$$L_t^2(\Gamma_I) := \{u \in (L^2(\Gamma_I))^3 \,:\, \nu \cdot u = 0 \quad \text{on} \quad \Gamma_I\}.$$

The scattered electric field E^s has the asymptotic behavior [5]

$$E^s(x) = \frac{e^{ik|x|}}{|x|}\left\{E_\infty(\hat{x}, d, p) + O\left(\frac{1}{|x|}\right)\right\}$$

as $|x| \to \infty$, where E_∞ is a tangential vector field defined on the unit sphere Ω and known as the *electric far field pattern*. The electric far field operator $F : L_t^2(\Omega) \to L_t^2(\Omega)$ is then defined by

$$(Fg)(\hat{x}) := \int_\Omega E_\infty(\hat{x}, d, g(d)) ds(d), \qquad \hat{x} \in \Omega, \tag{3.5}$$

for $g \in L_t^2(\Omega)$. Note that by superposition Fg is the electric far field pattern of the exterior mixed boundary value problem corresponding to the electromagnetic Herglotz pair with kernel ikg as incident field. An *electromagnetic Herglotz pair* is defined to be a pair of vector fields of the form

$$E_g(x) = \int_\Omega e^{ikx\cdot d} g(d) ds(d), \qquad H_g(x) = \frac{1}{ik}\operatorname{curl} E_g(x), \tag{3.6}$$

where $g \in L_t^2(\Omega)$. It is easily seen that E_g, H_g is a solution of Maxwell's equations $\operatorname{curl} E - ikH = 0$, $\operatorname{curl} H + ikE = 0$ in $\mathbb{R}^3$. Now let us consider the electric dipole with polarization q defined by

$$E_e(x, z, q) := \frac{i}{k}\operatorname{curl}_x \operatorname{curl}_x q\, \Phi(x, z), \qquad H_e(x, z, q) := \operatorname{curl}_x q\, \Phi(x, z), \tag{3.7}$$

where Φ is the fundamental solution of the Helmholtz equation in $\mathbb{R}^3$ defined by

$$\Phi(x, z) := \frac{1}{4\pi}\frac{e^{ik|x-z|}}{|x-z|}, \quad x \neq z \quad \text{and} \quad x, z \in \mathbb{R}^3.$$

If $z \in D$, then $E_e(x,z,q)$ and $H_e(x,z,q)$ satisfy Maxwell's equations in $\mathbb{R}^3 \setminus \overline{D}$, and the corresponding electric far field pattern $E_{e,\infty}(\hat{x},z,q)$ is given by

$$E_{e,\infty}(\hat{x},z,q) = \frac{ik}{4\pi}(\hat{x}\times q)\times \hat{x}\, e^{-ik\hat{x}\cdot z}. \tag{3.8}$$

As in the scalar case, we also need the interior mixed boundary value problem corresponding to the scattering problem which is studied in detail in [3]. (For the case when either $\Gamma_I = \emptyset$ or $\Gamma_D = \emptyset$, see [11].) In particular, let $E_z \in X(D,\Gamma_I)$ be the unique solution of

$$\operatorname{curl}\operatorname{curl} E_z - k^2 E_z = 0 \quad \text{in} \quad D, \tag{3.9a}$$

$$\nu \times [E_z + E_e(\cdot,z,q)] = 0 \quad \text{on} \quad \Gamma_D, \tag{3.9b}$$

$$\nu \times \operatorname{curl}(E_z + E_e(\cdot,z,q)) - i\lambda[\nu\times(E_z+E_e(\cdot,z,q))]\times\nu = 0 \quad \text{on } \Gamma_I \tag{3.9c}$$

for a fixed but arbitrary $z \in D$. Define

$$W_z := E_z + E_e(\cdot,z,q) \tag{3.10}$$

and let $u_T := (\nu\times u)\times\nu$ be the tangential component of a function $u \in H(\operatorname{curl},D)$. Note that $(W_z)_T|_{\Gamma_I} \in L_t^2(\Gamma_I)$ and that W_z depends on the artificial polarization q as well. We now look for a solution to the far field equation

$$Fg(\hat{x}) = E_{e,\infty}(\hat{x},z,q), \qquad z \in D, \tag{3.11}$$

where F is given by (3.5). We have the following result (see [3, Thm. 3.2]).

THEOREM 3.1. *For every $\epsilon > 0$ and $z \in D$ there exists an electric Herglotz wave function $E_{g_\epsilon^z}$ with kernel $g_\epsilon^z \in L_t^2(\Omega)$ such that*

$$\|E_z - ikE_{g_\epsilon^z}\|_{X(D,\Gamma_I)} \le \epsilon, \tag{3.12}$$

where E_z is the solution of (3.9a)–(3.9c). *Moreover, there exists a positive constant $c > 0$ independent of ϵ such that*

$$\|(Fg_\epsilon^z)(\hat{x}) - E_{e,\infty}(\hat{x},z,q)\|_{L_t^2(\Omega)} \le c\epsilon. \tag{3.13}$$

Our next aim is to find a relation that connects the surface impedance λ with E_z.

LEMMA 3.2. *For every two points z_1 and z_2 in D and polarization $q \in \mathbb{R}^3$ we have that*

$$2\int_{\Gamma_I} (W_{z_1})_T \cdot \lambda(\overline{W}_{z_2})_T\, ds = -\|q\|^2 A(z_1,z_2,k,q) + k\left(q\cdot E_{z_1}(z_2) + q\cdot \overline{E}_{z_2}(z_1)\right),$$

where E_{z_1}, E_{z_2} and W_{z_1}, W_{z_2} are defined by (3.9a)–(3.9c) *and* (3.10), *respectively, and $A(z_1,z_2,k,q)$ is a computable number depending only on z_1, z_2, k, and q.*

Proof. By applying the second vector Green's formula and using the boundary conditions for E_{z_1} and E_{z_2} on Γ we obtain

$$\begin{aligned}
2i\int_{\Gamma_I} (W_{z_1})_T \cdot \lambda(\overline{W}_{z_2})_T\, ds &= \int_\Gamma \left(\nu\times W_{z_1}\cdot\operatorname{curl}\overline{W}_{z_2} - \nu\times\overline{W}_{z_2}\cdot\operatorname{curl} W_{z_1}\right) ds \\
&= \int_\Gamma \left(\nu\times E_e(\cdot,z_1,q)\cdot\operatorname{curl}\overline{E_e(\cdot,z_2,q)} - \nu\times\overline{E_e(\cdot,z_2,q)}\cdot\operatorname{curl} E_e(\cdot,z_1,q)\right) ds \\
&\quad + \int_\Gamma \left(\nu\times E_{z_1}\cdot\operatorname{curl}\overline{E_e(\cdot,z_2,q)} - \nu\times\overline{E_e(\cdot,z_2,q)}\cdot\operatorname{curl} E_{z_1}\right) ds \\
&\quad + \int_\Gamma \left(\nu\times E_e(\cdot,z_1,q)\cdot\operatorname{curl}\overline{E}_{z_2} - \nu\times\overline{E}_{z_2}\cdot\operatorname{curl} E_e(\cdot,z_1,q)\right) ds.
\end{aligned} \tag{3.14}$$

One can easily see that if $E \in H(\mathrm{curl}, D)$ and $H = \frac{1}{ik}\mathrm{curl}\, E$ is a solution of Maxwell's equations and $z \in D$, we have

$$\begin{aligned}&\nu \times E_e(y,z,q) \cdot \mathrm{curl}_y\, \overline{E}(y) = -\frac{i}{k}(-ik)\mathrm{curl}_z\, \mathrm{curl}_z\, q\Phi(y,z) \cdot (\nu \times \overline{H}(y))\\ &= -q \cdot \mathrm{curl}_z\, \mathrm{curl}_z\, \Phi(y,z)(\nu \times \overline{H}(y))\end{aligned}$$

and

$$\nu \times \overline{E}(y) \cdot \mathrm{curl}_y\, E_e(y,z,q) = ik\nu \times \overline{E}(y) \cdot H_e(y,z,q) = ikq \cdot \mathrm{curl}_z \Phi(y,z)(\nu \times \overline{E}(y)),$$

and therefore from the Stratton–Chu formula

$$\int_\Gamma \left(\nu \times E_e(y,z,q) \cdot \mathrm{curl}_y\, \overline{E}(y) - \nu \times \overline{E}(y) \cdot \mathrm{curl}_y\, E_e(y,z,q)\right) = ikq \cdot \overline{E}(z). \tag{3.15}$$

Moreover (see [7]),

$$\begin{aligned}&\int_\Gamma \left(\nu \times E_e(\cdot,z_1,q) \cdot \mathrm{curl}\, \overline{E_e(\cdot,z_2,q)} - \nu \times \overline{E_e(\cdot,z_2,q)} \cdot \mathrm{curl}\, E_e(\cdot,z_1,q)\right) ds\\ &\quad = -2ik \int_\Omega E_{e,\infty}(\cdot,z_1,q) \cdot \overline{E_{e,\infty}(\cdot,z_2,q)} ds\\ &\quad = -\frac{ik^3}{8\pi^2} \int_\Omega ((\hat{x} \times q) \times \hat{x}) \cdot ((\hat{x} \times q) \times \hat{x})\, e^{-ik\hat{x}\cdot(z_1-z_2)} ds \qquad (3.16)\\ &\quad = -\frac{ik^3}{8\pi^2} \int_\Omega (\|q\|^2 - (\hat{x} \cdot q)^2)\, e^{-ik\hat{x}\cdot(z_1-z_2)} ds := -i\|q\|^2 A(z_1,z_2,k,q),\end{aligned}$$

where by straightforward calculations

$$A(z_1,z_2,k,q) = \frac{k^3}{6\pi}\left[2j_0(k|z_1-z_2|) + j_2(k|z_1-z_2|)(3\cos^2\phi - 1)\right] \tag{3.17}$$

with j_0 and j_2 being spherical Bessel functions of order 0 and 2, respectively, and ϕ is the angle between $(z_1 - z_2)$ and q. Hence using (3.15) and (3.16) in (3.14) and dividing both sides of (3.14) by i yield the result. ∎

Next we consider a subset $\mathcal{E}$ of $L_t^2(\Gamma_I)$ defined by

$$\mathcal{E} := \left\{ f \in L_t^2(\Gamma_I) : \begin{array}{c} f = (W_z)_T|_{\Gamma_I} \text{ with } W_z = E_z + E_e(\cdot,z,q), \\ z \in B_r,\ E_z \text{ the solution of (3.9a)–(3.9c) and } q \in \mathbb{R}^3 \end{array} \right\},$$

where B_r is a ball of radius r contained in D.

LEMMA 3.3. *$\mathcal{E}$ is complete in $L_t^2(\Gamma_I)$.*

Proof. Let $\varphi \in L_t^2(\Gamma_I)$ such that for every $z \in B_r$

$$\int_{\Gamma_I} (W_z)_T \cdot \varphi\, ds = 0.$$

Let $E \in X(D, \Gamma_I)$ be the solution of the interior mixed boundary value problem [3]

$$\begin{aligned}&\mathrm{curl}\, \mathrm{curl}\, E - k^2 E = 0 \quad \text{in} \quad D,\\ &\nu \times E = 0 \quad \text{on} \quad \Gamma_D,\\ &\nu \times \mathrm{curl}\, E - i\lambda E_T = \varphi \quad \text{on} \quad \Gamma_I.\end{aligned}$$

Then for $z \in B_r$ and $q \in \mathbb{R}^3$, using the fact that $(W_z)_T = E_T = 0$ on Γ_D, the second vector Green's formula, and (3.15), we have that

$$
\begin{aligned}
0 &= \int_{\Gamma_I} (W_z)_T \cdot \varphi \, ds = \int_{\Gamma} W_z \cdot (\nu \times \operatorname{curl} E - i\lambda E_T) \, ds \\
&= \int_{\Gamma} [E_z \cdot (\nu \times \operatorname{curl} E) - i\lambda E_z \cdot E_T + E_e(\cdot, z, q) \cdot (\nu \times \operatorname{curl} E) - i\lambda E_e(\cdot, z, q) \cdot E_T] \, ds \\
&= \int_{\Gamma} [E_z \cdot (\nu \times \operatorname{curl} E) - E \cdot (\nu \times \operatorname{curl} E_z)] \, ds \\
&+ \int_{\Gamma} [-E \cdot (\nu \times \operatorname{curl} E_e(\cdot, z, q)) + i\lambda E_T \cdot E_e(\cdot, z, q)] \, ds \\
&+ \int_{\Gamma} [E_e(\cdot, z, q) \cdot (\nu \times \operatorname{curl} E) - i\lambda E_e(\cdot, z, q) \cdot E_T] \, ds \\
&= \int_{\Gamma} [E_e(\cdot, z, q) \cdot (\nu \times \operatorname{curl} E) - E \cdot (\nu \times \operatorname{curl} E_e(\cdot, z, q))] \, ds \\
&= -\int_{\Gamma} [(\nu \times E_e(\cdot, z, q)) \cdot \operatorname{curl} E - (\nu \times E) \cdot \operatorname{curl} E_e(\cdot, z, q)] \, ds = ikq \cdot E(z).
\end{aligned}
$$

Thus $q \cdot E(z) = 0$ holds for all polarizations $q \in \mathbb{R}^3$ and $z \in B_r$, and hence $E(z) = 0$ for $z \in B_r$. By the unique continuation principle for the solution of Maxwell's equations in D we now see that $E \equiv 0$ in D, whence, by the trace theorem, $\varphi \equiv 0$, which proves the lemma. □

Combining Lemmas 3.2 and 3.3, we can prove in the same way as in the last part of the proof of Theorem 2.3 the main result of this section.

THEOREM 3.4. *Let $\lambda \in L_\infty(\Gamma_I)$ be the surface impedance of the scattering problem* (3.1a)–(3.4). *Then*

$$
\begin{aligned}
&\|\lambda\|_{L_\infty(\Gamma_I)} \\
&= \sup_{\substack{z_i \in B_r,\, q \in \mathbb{R}^3 \\ \alpha_i \in \mathbb{C}}} \frac{\sum_{i,j} \alpha_i \overline{\alpha_j} \left[-\|q\|^2 A(z_i, z_j, k, q) + k \left(q \cdot E_{z_i}(z_j) + q \cdot \overline{E}_{z_j}(z_i) \right) \right]}{2 \| \sum_i \alpha_i (W_{z_i})_T \|^2_{L_t^2(\Gamma)}},
\end{aligned} \tag{3.18}
$$

where $W_z = E_z + E_e(\cdot, z, q)$ with E_z being the solution to (3.9a)–(3.9c), *$A(z_i, z_j, k, q)$ is given by* (3.17), *and the sums are arbitrary finite sums.*

In the particular case where λ is a positive constant and setting $z_1 = z_2 = z_0 \in B_r$, we obtain the following formula for constant surface impedance:

$$
\lambda = \frac{-\frac{k^2}{6\pi} \|q\|^2 + k \operatorname{Re} (q \cdot E_{z_0})}{\|(W_{z_0})_T\|^2_{L_t^2(\Gamma)}}, \tag{3.19}
$$

where $W_{z_0} = E_{z_0} + E_e(\cdot, z_0, q)$ with E_{z_0} being the solution of (3.9a)–(3.9c) corresponding to $z_0 \in B_r$.

In both cases (3.18) and (3.19) E_z cannot be computed since λ appears in the boundary conditions. However, from Theorem 3.1 we can approximate E_z by the electric field ikE_{g^z} of the Herglotz electromagnetic pair with kernel ikg^z, where g^z is a (regularized) solution of the far field equation (3.11) for $z \in B_r \subset D$ and E_∞ is the measured far field data (we again assume that D is known by using the far field equation (3.11) and the linear sampling method as discussed in [3]). We note that, as

in the scalar case, the numerator on the right-hand side of (3.18) and (3.19) can be used as a target signature to detect whether or not a object is coated.

We conclude this section by remarking that, in both scalar and vector cases, it suffices to know only the far field data for a limited-aperture $\Omega_0 \subset \Omega$. In particular, in sections 2.3 and 3.2 of [4] it is proved that a Herglotz wave function and an electromagnetic Herglotz pair and their first derivatives can be approximated uniformly on a compact subset of a disk B_R of radius R by a Herglotz wave function and an electromagnetic Herglotz pair, respectively, with kernel supported in a subset $\Omega_0 \subset \Omega$. The kernel of this new Herglotz wave function can now be used in place of g_ϵ^z in Theorems 2.1 and 3.1, and therefore the corresponding $v_{g_\epsilon^z}$ and $E_{g_\epsilon^z}$ can be used as approximations of u_z and E_z, respectively, in the above formulas.

4. Numerical examples. In this section we give some results of numerical experiments performed in the scalar case when the surface impedance λ is a constant. As shown in section 2, an approximation for λ is given by

$$\frac{-2k\pi|\gamma|^2 - \operatorname{Im}\left(v_{g^z}(z)\right)}{\|v_{g^z} + \Phi(\cdot; z)\|^2_{L^2(\Gamma)}}, \qquad z = (x, y) \in D, \tag{4.1}$$

where $v_{g^z} = \int_0^{2\pi} g^z(d) e^{ik(x\cos\theta + y\sin\theta)} d\theta$, $d = (\cos\theta, \sin\theta)$, and the kernel g^z is the solution of the far field equation

$$\int_0^{2\pi} u_\infty(d, \hat{x}) g^z(d) d\theta = \gamma e^{-ik\hat{x}\cdot z}, \qquad z \in B_r \subset D.$$

The far field data is generated by the method of integral equations and is corrupted by random noise. We fix $k = 3$, select a domain D, boundaries Γ_D and Γ_I (in most of our examples $\Gamma_D = \emptyset$), and a constant λ and then solve the corresponding forward problem. We compute the far field pattern for 100 incident directions and observation directions equally distributed on the unit circle and add random noise of 1% or 10% to the Fourier coefficients of the far field pattern. Tikhonov regularization and the Morozov discrepancy principle are then used to solve the ill-posed discrete far field equation (see section 4 of [2] for details). We choose the sampling points z on a uniform grid of 101×101 points in the square region $[-5,\ 5]^2$ and compute the corresponding g^z. To visualize the obstacle we plot the level curves of the inverse of the discrete ℓ_2 norm of g (note that by the linear sampling method the boundary of the obstacle is characterized as the set of points where the L^2-norm of g starts to become large; see [2]). Then we compute (4.1) at the sampling points in the disk centered at the origin with radius 0.5 (in our examples this circle is always inside D). Although (4.1) is theoretically a constant, because of the ill-posed nature of the far field equation we evaluated (4.1) at all the grid points z in the disk and exhibit the maximum, the average, and the median of the computed values of (4.1). In all tested cases there are some outliers for the minimum value but this is not the case for the maximum. The average and median of the numbers obtained by evaluating (4.1) at the sampling points show that these numbers accumulate near the maximum value and that the average, median, and maximum each provides a reasonable approximation to the true impedance.

For our examples we select two scatterers shown in Figure 4.1 (the kite and the peanut).

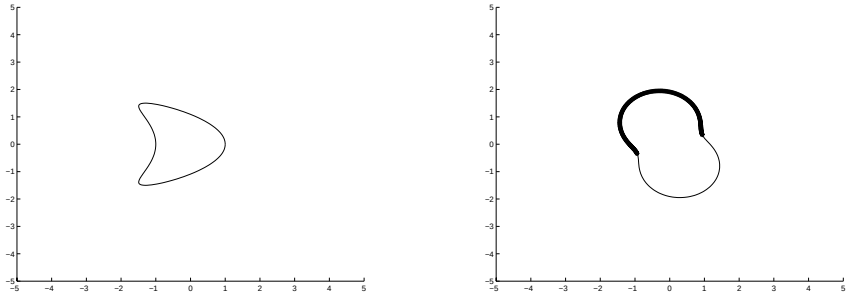


FIG. 4.1. *The boundary of the scatterers used in this study: kite/peanut. When a mixed condition is used for the peanut, the thicker portion of the boundary is* Γ_D.

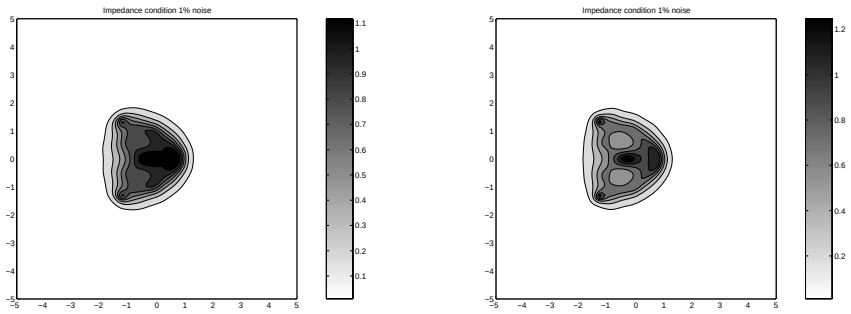


FIG. 4.2. *These figures show the reconstruction of a kite with impedance boundary condition with* 1% *noise: on the left with* $\lambda = 5$ *and on the right with* $\lambda = 9$.

We have obviously left open a number of interesting numerical questions, e.g., what is observed when $\lambda = 0$, what is the dependence of the algorithm on the wave number k, etc. In particular, the examples given here are preliminary in nature. Note that only in the example of the peanut do we consider an object that is really partially coated.

4.1. The kite. We consider the impedance boundary value problem for the kite described by the equation (the left curve in Figure 4.1)

$$x(t) = (1.5\sin(t), \cos(t) + 0.65\cos(2t) - 0.65), \qquad 0 \le t \le 2\pi,$$

with impedance $\lambda = 2$, $\lambda = 5$, and $\lambda = 9$. In Figure 4.2 we show two examples of the reconstructed kite (the reconstructions for the other tested cases look similar). Note that the reconstruction of the boundary is quite accurate so one obtains a good guess for the equation of the boundary Γ of the scatterer. In the numerical results for the reconstructed λ shown in Tables 4.1 and 4.2 we use the exact boundary Γ when we compute the $L^2(\Gamma)$-norm that appears in the denominator of (4.1).

4.2. The peanut. Next we consider a peanut described by the equation (the right curve in Figure 4.1)

$$x(t) = \left(\sqrt{\cos^2(t) + 4\sin^2(t)}\,\cos(t), \quad \sqrt{\cos^2(t) + 4\sin^2(t)}\,\sin(t), \qquad 0 \le t \le 2\pi\right)$$

rotated by $\pi/9$. Here we choose the surface impedance $\lambda = 2$ and $\lambda = 5$ and consider the case of a totally coated peanut (i.e., impedance boundary value problem) as well as

TABLE 4.1
The reconstruction of the surface impedance λ for the kite with 1% noise.

	Maximum	Average	Median
λ=2	2.050	1.975	1.982
λ=5	4.976	4.679	4.787
λ=9	8.883	8.342	8.403

TABLE 4.2
The reconstruction of the surface impedance λ for the kite with 10% noise.

	Maximum	Average	Median
λ=2	2.043	1.960	1.957
λ=5	4.858	4.513	4.524
λ=9	9.0328	8.013	7.992

of a partially coated peanut (i.e., mixed Dirichlet-impedance boundary value problem with Γ_I being the lower half of the peanut as shown in Figure 4.1). Two examples of the reconstructed peanut are presented in Figure 4.3 where, as expected, one notices that for the mixed case the Dirichlet portion of the boundary is more visible. In practice the exact boundary is not available to compute the $L^2(\Gamma)$-norm in (4.1). As suggested by the reconstruction of the peanut, the natural guess for the boundary of the scatterer is the ellipse shown by dashed line in Figure 4.4. So we also examine the sensitivity of our formula on the approximation of the boundary by using this ellipse for computing $\|v_{g^z} + \Phi(\cdot; z)\|_{L^2(\Gamma)}$ in (4.1). The recovered values of λ for our experiments are shown in Tables 4.3 and 4.4.

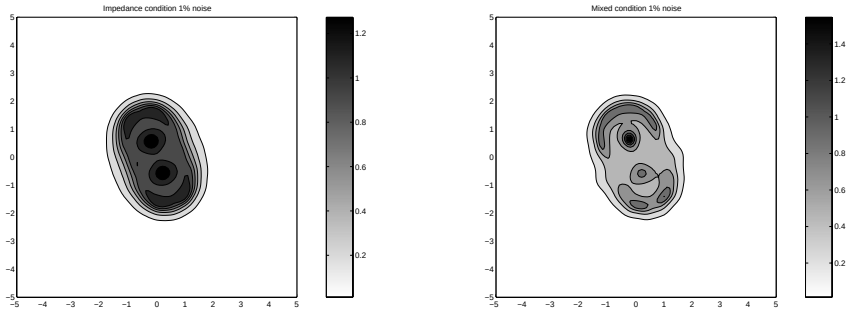


FIG. 4.3. *The figure on the left shows the reconstruction of a peanut with impedance boundary condition with $\lambda = 5$. The figure on the right shows the reconstruction of a peanut with mixed condition with $\lambda = 5$ on the impedance part. Both examples are for $k = 3$ with 1% noise.*

4.3. Conclusions. We have presented the results of some numerical experiments for the scalar case with constant surface impedance. The only a priori information we use is that the coating is homogeneous. Our results suggest that the maximum, median, and average values obtained by evaluating (4.1) at a set of sampling points in a disk closely approximate the true value of λ. We have further shown that even if the boundary of the scatterer is not known exactly, reasonable approximations to the impedance can still be obtained. Numerical experiments need to be done in $\mathbb{R}^3$ and for the nonhomogeneous coating where the scheme is a variational problem. This will be the subject of a forthcoming work.

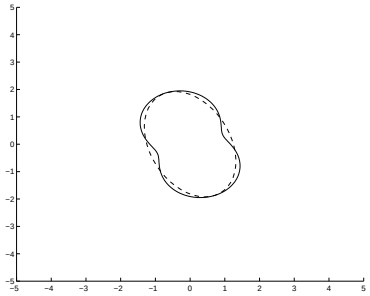


FIG. 4.4. *The dashed line is the approximated boundary we use for computing* $\|v_{g^z} + \Phi(\cdot; z)\|_{L^2(\Gamma)}$ *in* (4.1) *in the case of a peanut with impedance boundary condition.*

TABLE 4.3
Reconstruction of λ *for the peanut with* 1% *noise.*

	Maximum	Average	Median
λ=2 impedance	2.192	1.992	1.979
λ=2 imped., approx. bound.	2.395	1.823	1.886
λ=2 mixed conditions	2.595	2.207	2.257
λ=5 impedance	5.689	4.950	5.181
λ=5 imped., approx. bound.	5.534	4.412	4.501
λ=5 mixed conditions	5.689	4.950	5.180

TABLE 4.4
Reconstruction of λ *for the peanut with* 10% *noise.*

	Maximum	Average	Median
λ=2 impedance	2.297	1.985	1.978
λ=2 imped., approx. bound.	2.301	1.828	1.853
λ=2 mixed conditions	2.681	2.335	2.374
λ=5 impedance	5.335	4.691	4.731
λ=5 imped., approx. bound.	5.806	4.231	4.313
λ=5 mixed conditions	5.893	4.649	4.951

Acknowledgment. We would like to thank Professor Michele Piana and Professor Peter Monk for making their codes available to us so that we could construct the examples in section 4.

REFERENCES

[1] I. AKDUMAN AND R. KRESS, *The direct and inverse scattering problems for inhomogeneous impedance cylinders of arbitrary shape*, Radio Sci., 38 (2003), 1055.
[2] F. CAKONI, D. COLTON, AND P. MONK, *The direct and inverse scattering problems for partially coated obstacles*, Inverse Problems, 17 (2001), pp. 1997–2015.
[3] F. CAKONI, D. COLTON, AND P. MONK, *The electromagnetic inverse scattering problem for partially coated Lipschitz domains*, to appear.
[4] F. CAKONI AND D. COLTON, *Combined far field operators in electromagnetic inverse scattering theory*, Math. Methods Appl. Sci., 26 (2003), pp. 413–429.
[5] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory,* 2nd ed., Springer-Verlag, New York, 1998.
[6] D. COLTON AND R. KRESS, *Eigenvalues of the far field operator for the Helmholtz equation in an absorbing medium*, SIAM J. Appl. Math., 55 (1995), pp. 1724–1735.
[7] D. COLTON AND R. KRESS, *Eigenvalues of the far field operator and inverse scattering theory*, SIAM J. Math. Anal., 26 (1995), pp. 601–615.

[8] D. Colton and M. Piana, *Inequalities for inverse scattering problems in absorbing media*, Inverse Problems, 17 (2001), pp. 597–605.
[9] D. J. Hoppe and Y. Rahmat-Samii, *Impedance Boundary Conditions in Electromagnetics*, Taylor & Francis, New York, 1995.
[10] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[11] P. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, Oxford, UK, 2003.
[12] T. B. A. Senior and J. L. Volakis, *Approximate Boundary Conditions in Electromagnetics*, IEE, London, 1995.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 725–745

ANALYSIS OF THE RESPONSE MATRIX FOR AN EXTENDED TARGET*

HONGKAI ZHAO†

Abstract. In this paper we study the response matrix obtained from the interelement response of an active array of transducers that can send out signals and record reflected signals. In particular we analyze the eigenvalues and eigenvectors of the response matrix corresponding to the acoustic field reflected by an extended target, the size of which is comparable to the wavelength. We show that the eigenvalues are not well separated for a single extended target in general. However, when both the size of the target and the size of the active array are small compared to the distance from the array to the target, it is shown that the eigenvalues are well separated and that the leading eigenvalues and eigenvectors can be characterized in terms of the location and dimension of the target. Numerical experiments are presented to verify the analysis.

Key words. active array, response matrix, time reversal, Green's function

AMS subject classifications. 74J20, 74J25, 41A60

DOI. 10.1137/S0036139902415282

1. Introduction.

Active arrays have been used and studied in many applications such as medical imaging, nondestructive testing, seismic imaging, and target detection and recognition for sonar or radar systems. The types of signals (or wave fields) and devices vary by application. For example, in medical imaging ultrasound is mostly used, in sonar systems or underwater communications acoustic waves are typically used, in radar systems or wireless communications electromagnetic waves are used, and in other applications optics or lasers may be used. The typical setup of an active array is illustrated in Figure 1.1. The most important function of the active array is that each element in the array can both send out a signal and record the reflected signal. Such an active array can be used to probe a medium by sending out waves to illuminate reflective targets. Information about the targets can be extracted from the reflected signal. One of the key observations, which is explained in the next section, is that the reflected signal recorded at the array is related to the output signal of the array by a matrix, the response matrix. In many applications the response matrix can be obtained from the interelement response, i.e., the response received at one transducer corresponding to an impulse sent out from another transducer. Moreover, the product of the response matrix and its adjoint corresponds to the time reversal operator that has been studied extensively in [15, 14, 13, 12, 16, 17, 4, 3, 6, 1]. Understanding the structure of the response matrix, such as its eigenvalues and eigenvectors, is crucial for applications using active arrays.

In [15, 14, 13] an iterative time reversal procedure is proposed and analyzed for detecting and selectively focusing on targets. After recording the reflected signal, reversing it in time, and then sending it out to the medium for a few iterations, the wave field will automatically focus on the "strongest" scatterer. The whole physical procedure can be viewed as a power method for finding the leading eigenvector of the

*Received by the editors September 25, 2002; accepted for publication (in revised form) August 4, 2003; published electronically March 11, 2004.
http://www.siam.org/journals/siap/64-3/41528.html
†Department of Mathematics, University of California, Irvine, CA 92697-3875 (zhao@math.uci.edu). The research was supported by ONR grant N00014-02-1-0090, DARPA grant N00014-02-1-0603 and the Sloan Fellowship Foundation.

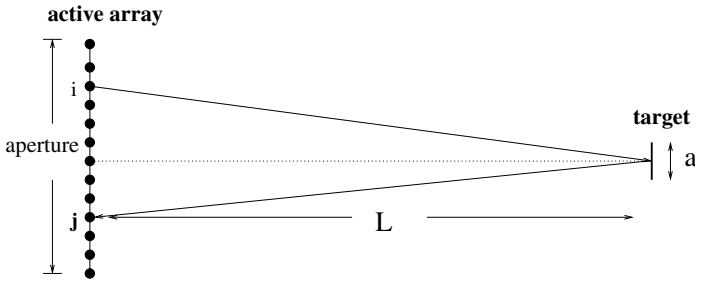


FIG. 1.1. *The setup of an active array.*

response matrix. For well-resolved point scatterers, whose sizes are small compared to those of the wavelength, it can be shown that each eigenvector of the response matrix corresponds to the wave field at the array due to a point source located at one of the scatterers. Instead of a physical iterated time reversal procedure, which can generate a physical wave field that focuses on a selective target, we can also analyze the response matrix on a computer for the detection or imaging of targets. For example, singular value decomposition (SVD) of the response matrix and subspace projection were used in [6] for detecting the locations of targets. In all these methods and their analyses the scatterers are considered as point-like scatterers so that the response matrix can be cleanly decomposed as the tensor product of the Green's function corresponding to each scatterer. It should be pointed out that in many applications the eigenvalues and eigenvectors of the response matrix may also depend on material properties. In [4, 3] it was shown that the compressibility and density contrast causes different scattered waves and generates more than one eigenstate even for a small spherical scatterer. However, the geometry of a general extended scatterer was not taken into account in an explicit way.

In this paper we will study the eigenvalues and eigenvectors of the response matrix corresponding to a single extended target. We show that in some asymptotic regimes, the leading eigenvalues and eigenvectors are well separated into groups and can be characterized in terms of the location and geometry of the target. Numerical tests match well with our analysis and show that our formulas can work well in more general situations. In the future we will study multiple targets, different designs of the arrays, imaging procedures for both locations and sizes of extended targets, and the effect of random inhomogeneity in the medium and self-averaging in the time domain.

Here is the outline of this paper. First, the response matrix and its basic properties for point scatterers are briefly reviewed in section 2. In section 3 the eigenvalues and eigenvectors of the response matrix corresponding to an extended scatterer are analyzed in some asymptotic regimes. In section 4 we study the effect of alignment between the active array and the target. Finally we show numerical experiments to verify our analysis in section 5.

2. The response matrix of an active array. Define the interelement response $p_{ij}(t)$ to be the reflected signal at the jth transducer corresponding to an impulse sent out from the ith transducer. For an array consisting of N transducers, the matrix $P(t) = [p_{ij}(t)]_{N\times N}$ is called the response matrix. If the medium is static, we have $p_{ij}(t) = p_{ji}(t)$ due to spatial reciprocity. If we assume the medium and the array response are linear, for an output signal $\vec{e}(t) = [e_1(t), e_2(t), \ldots, e_N(t)]^T$, where $e_i(t)$ is the output signal at the ith transducer and T means transpose, the reflected signal

at the array is

$$\vec{r}(t) = [r_1(t), r_2(t), \ldots, r_N(t)]^T = P(t) * \vec{e}(t).$$

Here $*$ denotes convolution in time. The convolution in time domain becomes the multiplication in frequency domain,

$$\vec{r}(\omega) = P(\omega)\vec{e}(\omega),$$

where ω is the frequency and $P(\omega)$ is the Fourier transform of $P(t)$. We follow the derivations in [14, 6] to illustrate the basic structure of the response matrix $P(\omega)$ for point scatterers in the medium. Denote $G(\boldsymbol{\xi}, \boldsymbol{x})$ to be Green's function of the medium for frequency ω, which represents the wave field at $\boldsymbol{x}$ for a point source located at $\boldsymbol{\xi}$. Due to the spatial reciprocity, $G(\boldsymbol{x}, \boldsymbol{\xi}) = G(\boldsymbol{\xi}, \boldsymbol{x})$. Here we suppress the dependence of Green's function on the frequency.

Assuming there are M point scatterers located at $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M$ in the medium with reflectivity $\tau_1, \tau_2, \ldots, \tau_M$, if we neglect the multiple scattering among the scatterers, then for a signal $\vec{e}(\omega) = [e_1(\omega), e_2(\omega), \ldots, e_N(\omega)]^T$ sent out from the active array, the reflected signal at the jth transducer is

$$r_j(\omega) = \sum_{k=1}^{M} \sum_{i=1}^{N} G(\boldsymbol{\xi}_j, \boldsymbol{x}_k)\tau_k G(\boldsymbol{\xi}_i, \boldsymbol{x}_k) e_i(\omega),$$

where $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_N$ are the locations of the transducers. If we define the illumination vectors, $\vec{g}_k, k = 1, 2, \ldots, M$, to be

$$\vec{g}_k = [G(\boldsymbol{\xi}_1, \boldsymbol{x}_k), G(\boldsymbol{\xi}_2, \boldsymbol{x}_k), \ldots, G(\boldsymbol{\xi}_N, \boldsymbol{x}_k)]^T,$$

i.e., the wave field at the array of transducers corresponding to a point source at the kth scatterer, we have

$$P(\omega) = \sum_{k=1}^{M} \tau_k \vec{g}_k \vec{g}_k^T \quad \text{and} \quad \vec{r}(\omega) = P(\omega)\vec{e}(\omega). \tag{2.1}$$

Due to the spatial reciprocity, $P(\omega)$ is symmetric. If we do time reversal, which is phase conjugation in frequency domain, for the reflected signal and send it back to the medium, the new reflected signal is $P(\omega)\overline{P(\omega)\vec{e}(\omega)}$, where $\bar{}$ denotes complex conjugation. Another phase conjugation gives the second time reversed output signal $\overline{P(\omega)}P(\omega)\vec{e}(\omega)$ in terms of the original output signal $\vec{e}(\omega)$. So $R(\omega) = \overline{P(\omega)}P(\omega) = P^*(\omega)P(\omega)$ is called the time reversal matrix (operator), where $*$ denotes the adjoint. $R(\omega)$ is a Hermitian matrix and from (2.1) we have

$$R(\omega) = \sum_{k=1}^{M} \overline{\tau_k \vec{g}_k \vec{g}_k^T} \sum_{k'=1}^{M} \tau_{k'} \vec{g}_{k'} \vec{g}_{k'}^T = \sum_{k'=1}^{M} \sum_{k=1}^{M} \Lambda_{k,k'} \bar{\vec{g}}_k \vec{g}_{k'}^T, \tag{2.2}$$

where

$$\Lambda_{k,k'} = \overline{\tau}_k \tau_{k'} \langle \vec{g}_k, \vec{g}_{k'} \rangle = \overline{\tau}_k \tau_{k'} \bar{\vec{g}}_k^T \vec{g}_{k'}.$$

All medium properties are embedded in Green's function in the above formulations. From representations (2.1) and (2.2), we can easily see that both the response matrix $P(\omega)$ and the time reversal matrix $R(\omega)$ are of rank M, if $M < N$, and that

any eigenvector corresponding to a nonzero eigenvalue is a linear combination of the illumination vectors $\vec{g}_k, k = 1, 2, \ldots, M$. We then define the point spread function

$$\Gamma(\boldsymbol{x}', \boldsymbol{x}) = \sum_{i=1}^{N} \overline{G(\boldsymbol{\xi}_i, \boldsymbol{x}')} G(\boldsymbol{\xi}_i, \boldsymbol{x}). \tag{2.3}$$

$\Gamma(\boldsymbol{x}', \boldsymbol{x})$ is exactly the wave field at point $\boldsymbol{x}$ after phase conjugation of the signal received at the active array corresponding to a point source at point $\boldsymbol{x}'$ and sending it back to the medium. The scatterers are well resolved by the active array means

$$\Gamma(\boldsymbol{x}_k, \boldsymbol{x}_{k'}) = \overline{\vec{g}_k^T} \vec{g}_{k'} \approx 0 \text{ if } k \neq k';$$

i.e., the wave field corresponding to the time reversal of a point source at one scatterer is almost zero at all other scatterers. Then $\vec{g}_k$ ($\overline{\vec{g}_k}$) are the left (right) singular vectors for $P(\omega)$ with singular values $|\tau_k| \|\vec{g}_k\|^2$ since

$$P(w)\overline{\vec{g}_k} = \tau_k \|\vec{g}_k\|^2 \vec{g}_k, \quad P^*(w)\vec{g}_k = \overline{\tau}_k \|\vec{g}_k\|^2 \overline{\vec{g}_k}.$$

Similarly it can be shown that $\overline{\vec{g}_k}$ is the eigenvector for the Hermitian matrix $R(\omega)$ with eigenvalue $|\tau_k|^2 \|\vec{g}_k\|^4$. In a homogeneous medium, the focusing property of the point spread function $\Gamma(\boldsymbol{x}', \boldsymbol{x})$ is dictated by the diffraction limit, which is proportional to wavelength and propagation distance and is inversely proportional to the size (aperture) of the active array. However, if the medium is inhomogeneous and random, the resolution of time reversal can beat the diffraction limit. The superresolution phenomenon is both observed in experiments [7, 5, 8, 9, 11] and theoretically analyzed in [7, 2]. It is shown in [2] that the effective aperture can be much larger than the physical size of the array due to multipathing in an inhomogeneous medium, and that the superresolution for time reversal is statistically stable in the time domain due to self-averaging of different frequencies in a broadband signal.

In [15, 14, 13], a physical iterative time reversal procedure (D.O.R.T.) was used to focus selectively on reflective targets in a real medium. The procedure is equivalent to the power method for finding the leading eigenvector of the time reversal matrix. Since physical time reversal is used, we do not need to know the medium. However, for selective focusing, the targets have to be well resolved by the active array. This procedure is useful for automatic target detection/destruction in practice. In [6] an algorithm for imaging point targets in the medium on computers using an active array was developed. However all these formulations and analyses assume the targets are point scatterers so that the response matrix and time reversal matrix have the simple structure in (2.1) and (2.2), respectively. In this paper we will analyze the eigenvalues and eigenvectors of the response matrix corresponding to an extended target.

3. The SVD of the response matrix for an extended target. In general the response matrix for an extended target does not have a simple decomposition as in the case for point targets. To simplify the analysis, we assume that each transducer of the active array can be viewed as a point source and the target is a perfect reflector with a normal reflectivity that is equal to 1. In this case the scattered field can be represented as an integral over the illuminated surface. So the response matrix can be written as

$$P_{ij}(\omega) = \int_{\Omega} G(\boldsymbol{\xi}_i, \boldsymbol{x}) G(\boldsymbol{\xi}_j, \boldsymbol{x}) \tau(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\xi}_j) d\boldsymbol{x}, \tag{3.1}$$

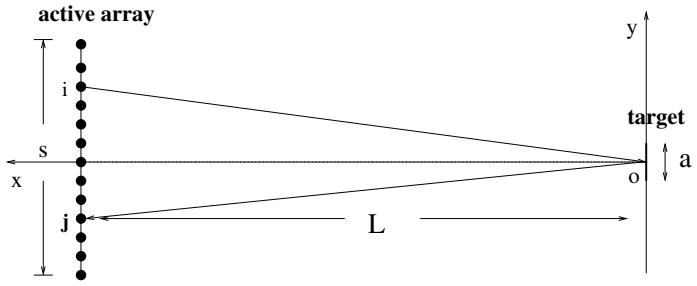


FIG. 3.1.

where Ω is the part of the surface that can be illuminated by the active array, and $\tau(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\xi}_j)$ is a reflectivity kernel that depends on the incidence and outgoing angle, i.e., the angle between the normal of the surface at $\boldsymbol{x}$ and the vectors $\boldsymbol{\xi}_i-\boldsymbol{x}$ and $\boldsymbol{\xi}_j-\boldsymbol{x}$, respectively.

In many applications, such as target detections using a sonar or radar system, wireless or underwater communications, and geophysics imaging, the distance between the target and the active array is much larger than the wavelength or sizes of the target and array. Let L be the distance between the array and the target, s be the size of the array, a be the size of the target, and $k=\frac{\omega}{c}$ be the wave number, where c is the wave speed. In our study we consider the case where

- the wavelength is comparable to the size of the array and the size of the extended target, i.e., $ka\sim O(1),\ ks\sim O(1)$;
- the wavelength is small compared to the distance between the array and the target, i.e., $\frac{1}{kL}\sim\frac{s}{L}\sim\frac{a}{L}\sim o(1)$.

In this case the wave from a transducer is almost planar when it reaches the target. Furthermore we assume the target is a planar target and lies in a plane that is parallel to the plane of the array. Since the size of the array and the size of the target are much smaller than the propagation distance, both the incidence and outgoing angles are small. We first neglect the reflectivity kernel and approximate the response matrix by

$$P_{ij}(k)=\int_{\Omega}G(\boldsymbol{\xi}_i,\boldsymbol{x})G(\boldsymbol{\xi}_j,\boldsymbol{x})d\boldsymbol{x}. \tag{3.2}$$

We will put in a reflectivity kernel later.

For simplicity we start with a one-dimensional target and array in a homogeneous medium as illustrated in Figure 3.1. We expand Green's function $G(\boldsymbol{\xi}_i,\boldsymbol{x})$ in free space at a point $\boldsymbol{x}=(0,y)$ on the target in powers of $\frac{1}{L}$. The expansion actually involves powers of ka, ks, which are $O(1)$, and powers of $\frac{a}{L},\frac{s}{L}$, which are $o(1)$.

$$\begin{aligned}G(\boldsymbol{\xi}_i,\boldsymbol{x})&=\frac{e^{ik|\boldsymbol{\xi}_i-\boldsymbol{x}|}}{4\pi|\boldsymbol{\xi}_i-\boldsymbol{x}|}=\frac{e^{ik\sqrt{L^2+(\eta_i-y)^2}}}{4\pi\sqrt{L^2+(\eta_i-y)^2}}\\&=\tilde{G}(\boldsymbol{\xi}_i,\boldsymbol{o})\frac{e^{\frac{ik(-2\eta_i y+y^2)}{2L}+O(\frac{1}{L^3})}}{1-\frac{2\eta_i y-y^2}{2L^2+\eta_i^2}+O(\frac{1}{L^4})}=\tilde{G}(\boldsymbol{\xi}_i,\boldsymbol{o})\frac{e^{\frac{ik(-2\eta_i y+y^2)}{2L}+O(\frac{1}{L^3})}}{1-\frac{2\eta_i y-y^2}{2L^2}+O(\frac{1}{L^4})},\end{aligned} \tag{3.3}$$

where $\boldsymbol{o}=(0,0)$ is the center of the target, $\boldsymbol{\xi}_i=(L,\eta_i)$ is the location of the ith trans-

ducer, and $\tilde{G}(\boldsymbol{\xi}_i, \boldsymbol{o}) = \frac{e^{ikL(1+\frac{\eta_i^2}{2L^2})}}{4\pi L(1+\frac{\eta_i^2}{2L^2})}$ is the parabolic approximation of Green's function $G(\boldsymbol{\xi}_i, \boldsymbol{o})$. Here we can use the two-dimensional Green's function, which is a Hankel function of the first kind, but instead we use the three-dimensional Green's function for simplicity and consistency with later analysis. Further expanding in $\frac{1}{L}$, we have

$$G(\boldsymbol{\xi}_i, \boldsymbol{x}) = \tilde{G}(\boldsymbol{\xi}_i, \boldsymbol{o}) \left[1 + \frac{ik(y^2 - 2\eta_i y)}{2L} - \frac{k^2(y^2 - 2\eta_i y)^2}{8L^2} - \frac{y^2 - 2\eta_i y}{2L^2} + O(\frac{1}{L^3}) \right]. \tag{3.4}$$

Now the response matrix becomes

$$\begin{aligned}
P_{ij}(k) &= \int_{-\frac{a}{2}}^{\frac{a}{2}} G(\boldsymbol{\xi}_i, \boldsymbol{x}) G(\boldsymbol{\xi}_j, \boldsymbol{x}) dy \\
&= \tilde{G}(\boldsymbol{\xi}_i, \boldsymbol{o}) \tilde{G}(\boldsymbol{\xi}_j, \boldsymbol{o}) \int_{-\frac{a}{2}}^{\frac{a}{2}} \left[1 + \frac{iky^2}{L} - \frac{k^2(\eta_i + \eta_j - y)^2 y^2}{2L^2} + \frac{(\eta_i + \eta_j - y)y}{L^2} + O\left(\frac{1}{L^3}\right) \right] dy \\
&= \tilde{G}(\boldsymbol{\xi}_i, \boldsymbol{o}) \tilde{G}(\boldsymbol{\xi}_j, \boldsymbol{o}) \int_{-\frac{a}{2}}^{\frac{a}{2}} \left(1 + \frac{iky^2}{L} - \frac{k^2 y^4}{2L^2} - \frac{y^2}{L^2} \right) dy \\
&\quad - \frac{k^2}{L^2} \tilde{G}(\boldsymbol{\xi}_i, \boldsymbol{o}) \tilde{G}(\boldsymbol{\xi}_j, \boldsymbol{o}) \eta_i \eta_j \int_{-\frac{a}{2}}^{\frac{a}{2}} y^2 dy - \frac{k^2}{2L^2} \tilde{G}(\boldsymbol{\xi}_i, \boldsymbol{o}) \tilde{G}(\boldsymbol{\xi}_j, \boldsymbol{o}) (\eta_i^2 + \eta_j^2) \int_{-\frac{a}{2}}^{\frac{a}{2}} y^2 dy \\
&\quad + O\left(\frac{1}{L^3}\right).
\end{aligned} \tag{3.5}$$

Denote $\alpha(a, k, L) = \int_{-\frac{a}{2}}^{\frac{a}{2}} (1 + \frac{iky^2}{L} - \frac{k^2 y^4}{2L^2} - \frac{y^2}{L^2}) dy \approx a$. So the response matrix can be decomposed as

$$P(k) = \alpha(a, k, L) \vec{\tilde{g}} \vec{\tilde{g}}^T - \frac{k^2 a^3}{12L^2} \vec{\tilde{g}}_1 \vec{\tilde{g}}_1^T - \frac{k^2 a^3}{24L^2} \left[\vec{\tilde{g}}_2 \vec{\tilde{g}}^T + \vec{\tilde{g}} \vec{\tilde{g}}_2^T \right] + O\left(\frac{1}{L^3}\right), \tag{3.6}$$

where

$$\begin{aligned}
\vec{\tilde{g}} &= [\tilde{G}(\boldsymbol{\xi}_1, \boldsymbol{o}), \tilde{G}(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \tilde{G}(\boldsymbol{\xi}_N, \boldsymbol{o})]^T, \\
\vec{\tilde{g}}_1 &= [\eta_1 \tilde{G}(\boldsymbol{\xi}_1, \boldsymbol{o}), \eta_2 \tilde{G}(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \eta_N \tilde{G}(\boldsymbol{\xi}_N, \boldsymbol{o})]^T, \\
\vec{\tilde{g}}_2 &= [\eta_1^2 \tilde{G}(\boldsymbol{\xi}_1, \boldsymbol{o}), \eta_2^2 \tilde{G}(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \eta_N^2 \tilde{G}(\boldsymbol{\xi}_N, \boldsymbol{o})]^T.
\end{aligned} \tag{3.7}$$

The leading term $a \vec{\tilde{g}} \vec{\tilde{g}}^T$ corresponds to a point scatterer at the center of the extended target with a total reflectivity proportional to the size of the target. Moreover, if the center of the array is aligned with the center of the target, $\vec{\tilde{g}}$ and $\vec{\tilde{g}}_2$ are even in η while $\vec{\tilde{g}}_1$ is odd in η, so we have

$$\vec{\tilde{g}}_1^T \vec{\tilde{g}} = \vec{\tilde{g}}_1^T \vec{\tilde{g}}_2 = 0.$$

From the orthogonality condition we can separate the second term from the other terms in the matrix decomposition formula (3.6). Moreover, since the size of array s is small compared to the distance L, the parabolic factor $\frac{\eta^2}{L^2}$ is very flat, and $\vec{g}_2$ is approximately aligned with $\vec{g}$. Most of the term $\frac{k^2 a^3}{24L^2} [\vec{\tilde{g}}_2 \vec{\tilde{g}}^T + \vec{\tilde{g}} \vec{\tilde{g}}_2^T]$ is absorbed in

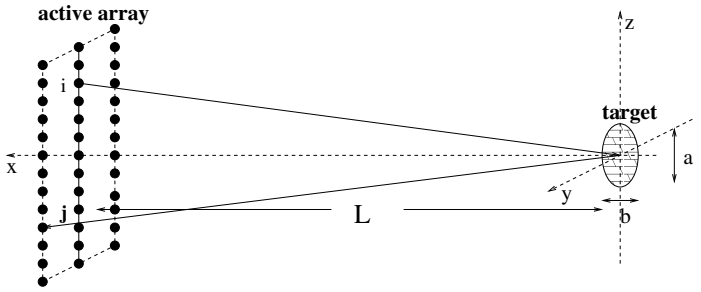


FIG. 3.2.

the leading term as a small perturbation. In particular the contribution of this term together with other high-order terms in $\alpha(a,k,L)$ tends to make the amplitude of the first eigenvalue smaller. This will be verified by numerical experiments in section 5.

From the analysis, we see that

$$P\bar{\vec{g}} \approx a\|\vec{g}\|^2\vec{g}, \quad P^*\vec{g} \approx a\|\vec{g}\|^2\bar{\vec{g}},$$
$$P\bar{\vec{g}}_1 \approx -\tfrac{a^3k^2}{12L^2}\|\vec{g}_1\|^2\vec{g}_1, \quad P^*\vec{g}_1 \approx -\tfrac{a^3k^2}{12L^2}\|\vec{g}_1\|^2\bar{\vec{g}}_1,$$

and

- the dominant singular value λ_1 has a magnitude $|\lambda_1| \approx a\|\vec{g}\|^2$, and the associated left singular vector is the illumination vector $\bar{\vec{g}}$;
- the second dominant singular value λ_2 has a magnitude $|\lambda_2| \approx \frac{a^3k^2}{12L^2}\|\vec{g}_1\|^2$, and the associated left singular vector is $\bar{\vec{g}}_1$.

When we extend the above analysis to a two-dimensional planar target and array as illustrated in Figure 3.2, the calculation is similar but messier. First pick a point $\boldsymbol{o}$ on the target as the origin and choose two orthogonal directions in the plane of the target as y and z axes. Let $\boldsymbol{x} = (0, y, z)$ denote a point on the target and $\boldsymbol{\xi}_i = (L, \eta_i, \zeta_i)$ denote the coordinates of the ith transducer. The situation becomes more complicated due to the coupling of different directions. We have

$$G(\boldsymbol{\xi}_i, \boldsymbol{x}) = \frac{e^{ik\sqrt{L^2+(\eta_i-y)^2+(\zeta_i-z)^2}}}{4\pi\sqrt{L^2+(\eta_i-y)^2+(\zeta_i-z)^2}} = \tilde{G}(\boldsymbol{\xi}_i, \boldsymbol{o})$$
$$\times\left[1+\frac{ik(y^2-2\eta_i y+z^2-2\zeta_i z)}{2L}-\frac{k^2(y^2-2\eta_i y+z^2-2\zeta_i z)^2}{8L^2}\right.$$
$$\left.-\frac{y^2-2\eta_i y+z^2-2\zeta_i z}{2L^2}+O\left(\frac{1}{L^3}\right)\right],$$

where $\tilde{G}(\boldsymbol{\xi}_i, \boldsymbol{o}) = \frac{e^{ikL(1+\frac{\eta_i^2+\zeta_i^2}{2L^2})}}{4\pi L(1+\frac{\eta_i^2+\zeta_i^2}{2L^2})}$ is again the parabolic approximation of Green's function, and

$$G(\boldsymbol{\xi}_i, \boldsymbol{x})G(\boldsymbol{\xi}_j, \boldsymbol{x}) = \tilde{G}(\boldsymbol{\xi}_i, \boldsymbol{o})\tilde{G}(\boldsymbol{\xi}_j, \boldsymbol{o}) \times \left\{1+\frac{ik[y^2-(\eta_i+\eta_j)y+z^2-(\zeta_i+\zeta_j)z]}{L}\right.$$
$$\left.-\frac{k^2[y^2-(\eta_i+\eta_j)y+z^2-(\zeta_i+\zeta_j)z]^2}{2L^2}-\frac{y^2-(\eta_i+\eta_j)y+z^2-(\zeta_i+\zeta_j)z}{L^2}+O\left(\frac{1}{L^3}\right)\right\}.$$

So the response matrix is

$$
\begin{aligned}
&P_{ij}(k) = \int_\Omega G(\boldsymbol{\xi}_i, \boldsymbol{x}) G(\boldsymbol{\xi}_j, \boldsymbol{x}) dydz = \tilde{G}(\boldsymbol{\xi}_i, \boldsymbol{o}) \tilde{G}(\boldsymbol{\xi}_j, \boldsymbol{o}) \\
&\times \left[\int_\Omega \left(1 + \frac{ik(y^2+z^2)}{L} - \frac{k^2(y^2+z^2)^2}{2L^2} - \frac{y^2+z^2}{L^2} \right) dydz \right. \\
&-\eta_i\eta_j \int_\Omega \frac{k^2y^2}{L^2} dydz - \zeta_i\zeta_j \int_\Omega \frac{k^2z^2}{L^2} dydz \\
&-(\eta_i^2+\eta_j^2) \int_\Omega \frac{k^2y^2}{2L^2} dydz - (\zeta_i^2+\zeta_j^2) \int_\Omega \frac{k^2z^2}{2L^2} dydz \\
&+(\eta_i+\eta_j) \int_\Omega \left(-\frac{iky}{L} + \frac{k^2(y^2+z^2)y}{L^2} + \frac{y}{L^2} \right) dydz \\
&+(\zeta_i+\zeta_j) \int_\Omega \left(-\frac{ikz}{L} + \frac{k^2(y^2+z^2)z}{L^2} + \frac{z}{L^2} \right) dydz \\
&\left. -(\eta_i\zeta_i + \eta_i\zeta_j + \eta_j\zeta_i + \eta_j\zeta_j) \int_\Omega \frac{k^2yz}{L^2} dydz + O\left(\frac{1}{L^3}\right) \right].
\end{aligned}
$$

Denote

$$
\begin{aligned}
\vec{\tilde{g}} &= [\tilde{G}(\boldsymbol{\xi}_1, \boldsymbol{o}), \tilde{G}(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \tilde{G}(\boldsymbol{\xi}_N, \boldsymbol{o})]^T, \\
\vec{\tilde{g}}_{1y} &= [\eta_1\tilde{G}(\boldsymbol{\xi}_1, \boldsymbol{o}), \eta_2\tilde{G}(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \eta_N\tilde{G}(\boldsymbol{\xi}_N, \boldsymbol{o})]^T, \\
\vec{\tilde{g}}_{1z} &= [\zeta_1\tilde{G}(\boldsymbol{\xi}_1, \boldsymbol{o}), \zeta_2\tilde{G}(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \zeta_N\tilde{G}(\boldsymbol{\xi}_N, \boldsymbol{o})]^T, \\
\vec{\tilde{g}}_{2y} &= [\eta_1^2\tilde{G}(\boldsymbol{\xi}_1, \boldsymbol{o}), \eta_2^2\tilde{G}(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \eta_N^2\tilde{G}(\boldsymbol{\xi}_N, \boldsymbol{o})]^T, \\
\vec{\tilde{g}}_{2z} &= [\zeta_1^2\tilde{G}(\boldsymbol{\xi}_1, \boldsymbol{o}), \zeta_2^2\tilde{G}(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \zeta_N^2\tilde{G}(\boldsymbol{\xi}_N, \boldsymbol{o})]^T, \\
\vec{\tilde{g}}_{2yz} &= [\eta_1\zeta_1\tilde{G}(\boldsymbol{\xi}_1, \boldsymbol{o}), \eta_2\zeta_2\tilde{G}(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \eta_N\zeta_N\tilde{G}(\boldsymbol{\xi}_N, \boldsymbol{o})]^T.
\end{aligned} \tag{3.8}
$$

The response matrix can be decomposed as

$$
\begin{aligned}
&P(k) = \vec{\tilde{g}}\vec{\tilde{g}}^T \int_\Omega \left(1 + \frac{ik(y^2+z^2)}{L} - \frac{k^2(y^2+z^2)^2}{2L^2} - \frac{y^2+z^2}{L^2} \right) dydz \\
&-\vec{\tilde{g}}_{1y}\vec{\tilde{g}}_{1y}^T \int_\Omega \frac{k^2y^2}{L^2} dydz - \vec{\tilde{g}}_{1z}\vec{\tilde{g}}_{1z}^T \int_\Omega \frac{k^2z^2}{L^2} dydz \\
&- \left(\vec{\tilde{g}}_{2y}\vec{\tilde{g}}^T + \vec{\tilde{g}}\vec{\tilde{g}}_{2y}^T \right) \int_\Omega \frac{k^2y^2}{2L^2} dydz - \left(\vec{\tilde{g}}_{2z}\vec{\tilde{g}}^T + \vec{\tilde{g}}\vec{\tilde{g}}_{2z}^T \right) \int_\Omega \frac{k^2z^2}{2L^2} dydz \\
&+ \left(\vec{\tilde{g}}_{1y}\vec{\tilde{g}}^T + \vec{\tilde{g}}\vec{\tilde{g}}_{1y}^T \right) \int_\Omega \left(-\frac{iky}{L} + \frac{k^2(y^2+z^2)y}{L^2} + \frac{y}{L^2} \right) dydz \\
&+ \left(\vec{\tilde{g}}_{1z}\vec{\tilde{g}}^T + \vec{\tilde{g}}\vec{\tilde{g}}_{1z}^T \right) \int_\Omega \left(-\frac{ikz}{L} + \frac{k^2(y^2+z^2)z}{L^2} + \frac{z}{L^2} \right) dydz \\
&- \left(\vec{\tilde{g}}_{1y}\vec{\tilde{g}}_{1z}^T + \vec{\tilde{g}}_{1z}\vec{\tilde{g}}_{1y}^T + \vec{\tilde{g}}_{2yz}\vec{\tilde{g}}^T + \vec{\tilde{g}}\vec{\tilde{g}}_{2yz}^T \right) \int_\Omega \frac{k^2yz}{L^2} dydz + O\left(\frac{1}{L^3}\right).
\end{aligned} \tag{3.9}
$$

The above formula shows the decomposition of the response matrix in any coordinate system. Again the leading term corresponds to a point scatterer. The decomposition of the response matrix is not obvious from this general expression since the two directions and the center are coupled together. However, if we choose $\boldsymbol{o}$ to be the mass center of the target, we have

$$\int_\Omega y dydz, \qquad \int_\Omega z dydz = 0.$$

Now we have an extra freedom of rotation of the y and z axes around $\boldsymbol{o}$. Denote the integral $R(\theta) = \int_\Omega yzdydz$ as a function of rotation angle θ. Since $R(\theta) = -R(\theta \pm \pi/2)$, we must have at least two θ such that $\int_\Omega yzdydz = 0$. Due to this symmetry and cancellations, the following terms are usually small:

$$\int_\Omega y^p z^q dydz \approx 0 \text{ for } (p,q) \in \{(1,2),(2,1),(0,3),(3,0)\}.$$

Hence the decomposition of the response matrix can be simplified as

$$\begin{aligned} P(k) \approx{}& \vec{\tilde{g}}\vec{\tilde{g}}^T \int_\Omega dydz - \frac{k^2}{L^2}\vec{\tilde{g}}_{1y}\vec{\tilde{g}}_{1y}^T \int_\Omega y^2 dydz \\ &- \frac{k^2}{L^2}\vec{\tilde{g}}_{1z}\vec{\tilde{g}}_{1z}^T \int_\Omega z^2 dydz - \frac{k^2}{2L^2}\left[\vec{\tilde{g}}_{2y}\vec{\tilde{g}}^T + \vec{\tilde{g}}\vec{\tilde{g}}_{2y}^T\right] \int_\Omega y^2 dydz \\ &- \frac{k^2}{2L^2}\left[\vec{\tilde{g}}_{2z}\vec{\tilde{g}}^T + \vec{\tilde{g}}\vec{\tilde{g}}_{2z}^T\right] \int_\Omega z^2 dydz + O\left(\frac{1}{L^3}\right). \end{aligned} \tag{3.10}$$

Again, if the array of transducers is symmetric with respect to the y and z axes, we have the following orthogonality properties defined in the sense of complex inner product as in the one-dimensional case:

$$\vec{\tilde{g}}_{1y} \perp \vec{\tilde{g}},\ \vec{\tilde{g}}_{2y},\ \vec{\tilde{g}}_{2z}, \quad \vec{\tilde{g}}_{1z} \perp \vec{\tilde{g}},\ \vec{\tilde{g}}_{2y},\ \vec{\tilde{g}}_{2z}, \quad \vec{\tilde{g}}_{1y} \perp \vec{\tilde{g}}_{1z}.$$

In this case we have

- the dominant singular value λ_1 has a magnitude $|\lambda_1| \approx \|\vec{\tilde{g}}\|^2 \int_\Omega dydz$ with the illumination $\vec{\tilde{g}}$ as the singular vector;
- the next two dominant singular values have magnitudes $|\lambda_2| \approx \frac{k^2}{L^2}\|\vec{\tilde{g}}_{1y}\|^2 \int_\Omega y^2 dydz$ and $|\lambda_3| \approx \frac{k^2}{L^2}\|\vec{\tilde{g}}_{1z}\|^2 \int_\Omega z^2 dydz$, respectively, and the corresponding singular vectors are $\vec{\tilde{g}}_{1y}$ and $\vec{\tilde{g}}_{1z}$, respectively.

The two symmetric axes of the target are intrinsic and are independent of the artificial y and z axes we choose. The directions of the two symmetric axes with respect to the artificial y and z axes we choose are embedded in the response matrix and can be extracted from the leading singular vectors, as will be shown later from the numerical experiments in section 5. Essentially these formulas suggest that we can find both the location and size of an extended target as well as the symmetric axes and a few moments with respect to these axes using the leading singular values and singular vectors of the response matrix. In practice, ratios of the singular values are more robust and can be used to determine aspect ratios.

For a three-dimensional scatterer for which the Born approximation is valid, the integration over the scatterer will be in three dimensions. Using similar analysis

for the response matrix, we will get a third singular value and singular vector that corresponds to the third dimension in the second group of leading singular values and singular vectors. In practice, the depth information is often relatively weak due to a small glancing aperture and is more difficult to capture from the response matrix. Especially when the distance between the active array and the target is long, we would see an effective planar shape of the target. In [17], careful numerical simulations are done to analyze the eigenvalues and eigenvectors of the response matrix and illustrate very similar behavior to our results here.

For a medium with weak random inhomogeneity, it is shown that for a long propagation distance, Green's function is a modification of the homogeneous Green's function with an attenuation factor due to multiple scattering [10]. So the behaviors of the response matrix and its eigenvalues and eigenvectors should be similar. We will analyze these situations in a future study.

Now we take into account the variation of reflectivity due to different incident and outgoing angles. In particular we choose a special reflectivity kernel

$$\tau(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\xi}_j)=\cos\theta_i(\boldsymbol{x})\cos\theta_j(\boldsymbol{x}),$$

where $\theta_i(\boldsymbol{x})$ is the angle between $\boldsymbol{x}-\boldsymbol{\xi}_i$ and the normal at $\boldsymbol{x}$. At a point $\boldsymbol{x}$ on a one-dimensional target,

$$\cos\theta_i(\boldsymbol{x})=\frac{L}{\sqrt{L^2+(\eta_i-y)^2}}=\frac{1}{1+\frac{\eta_i^2}{2L^2}}\left(1-\frac{y^2-2\eta_i y}{2L^2}\right)+O\left(\frac{1}{L^4}\right).$$

If we plug this asymptotic expansion back into the response matrix expression (3.1) and denote

$$\breve{G}(\boldsymbol{\xi}_i,\boldsymbol{o})=\tilde{G}(\boldsymbol{\xi}_i,\boldsymbol{o})\frac{1}{1+\frac{\eta_i^2}{2L^2}},$$

we have

$$\begin{aligned}P_{ij}(k)&=\breve{G}(\boldsymbol{\xi}_i,\boldsymbol{o})\breve{G}(\boldsymbol{\xi}_j,\boldsymbol{o})\int_{-\frac{a}{2}}^{\frac{a}{2}}\left[1+\frac{iky^2}{L}-\frac{k^2(\eta_i+\eta_j-y)^2y^2}{2L^2}+\frac{2(\eta_i+\eta_j-y)y}{L^2}+O\left(\frac{1}{L^3}\right)\right]dy\\&=\breve{G}(\boldsymbol{\xi}_i,\boldsymbol{o})\breve{G}(\boldsymbol{\xi}_j,\boldsymbol{o})\int_{-\frac{a}{2}}^{\frac{a}{2}}\left(1+\frac{iky^2}{L}-\frac{k^2y^4}{2L^2}-\frac{2y^2}{L^2}\right)dy\\&\quad-\frac{k^2}{L^2}\breve{G}(\boldsymbol{\xi}_i,\boldsymbol{o})\breve{G}(\boldsymbol{\xi}_j,\boldsymbol{o})\eta_i\eta_j\int_{-\frac{a}{2}}^{\frac{a}{2}}y^2dy-\frac{k^2}{2L^2}\breve{G}(\boldsymbol{\xi}_i,\boldsymbol{o})\breve{G}(\boldsymbol{\xi}_j,\boldsymbol{o})(\eta_i^2+\eta_j^2)\int_{-\frac{a}{2}}^{\frac{a}{2}}y^2dy\\&\quad+O\left(\frac{1}{L^3}\right),\end{aligned}$$

which is almost exactly the same as (3.5). Now if we define the new illumination vectors to be

$$\begin{aligned}\vec{\breve{g}}&=[\breve{G}(\boldsymbol{\xi}_1,\boldsymbol{o}),\breve{G}(\boldsymbol{\xi}_2,\boldsymbol{o}),\ldots,\breve{G}(\boldsymbol{\xi}_N,\boldsymbol{o})]^T,\\\vec{\breve{g}}_1&=[\eta_1\breve{G}(\boldsymbol{\xi}_1,\boldsymbol{o}),\eta_2\breve{G}(\boldsymbol{\xi}_2,\boldsymbol{o}),\ldots,\eta_N\breve{G}(\boldsymbol{\xi}_N,\boldsymbol{o})]^T,\\\vec{\breve{g}}_2&=[\eta_1^2\breve{G}(\boldsymbol{\xi}_1,\boldsymbol{o}),\eta_2^2\breve{G}(\boldsymbol{\xi}_2,\boldsymbol{o}),\ldots,\eta_N^2\breve{G}(\boldsymbol{\xi}_N,\boldsymbol{o})]^T,\end{aligned}\tag{3.11}$$

and $\alpha(a,k,L)=\int_{-\frac{a}{2}}^{\frac{a}{2}}(1+\frac{iky^2}{L}-\frac{k^2y^4}{2L^2}-\frac{2y^2}{L^2})dy$, we have exactly the same decomposition of the response matrix as in (3.6). For a more general reflectivity kernel,

$$\tau(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\xi}_j)=f(\theta_i(\boldsymbol{x}),\theta_j(\boldsymbol{x}))=\tilde{f}\left(\frac{1}{\sqrt{1+\left(\frac{\eta_i-y}{L}\right)^2}},\frac{1}{\sqrt{1+\left(\frac{\eta_j-y}{L}\right)^2}}\right),$$

we can use power expansion in $\frac{1}{L}$ to get the explicit formulas. The extension to a two-dimensional target is exactly the same.

4. Alignment of the array and the target. In the above analysis, the geometric decomposition of the response matrix for an extended target utilizes the symmetry and alignment of the active array with the target. In most applications, neither the geometry nor the location of the target is known. Hence the alignment of the active array with the target and how it affects the decomposition of the response matrix are important questions in practice. For a general two-dimensional target there are two alignments; one is the alignment of the center and the other is the alignment of the lines of symmetry. The center and symmetry of the target are intrinsic while the center and symmetry of the active array can be maneuvered. In fact, if the geometry of the active array is designed properly, such as in the shape of a disc, the array is symmetric with respect to any orthogonal coordinate system whose origin is at the center of the array. We will see from numerical tests in section 5 that the two symmetric directions are automatically embedded in the singular vectors of the response matrix and can be found easily. Now the only issue becomes the alignment of the center of the array and the center of the target.

We study the simple case of a one-dimensional target illustrated in Figure 4.1. Our previous decomposition of the response matrix (3.6) is not changed. Hence the leading term in the decomposition is still approximately $a\vec{g}\vec{g}^T$, where a is the size of the target and $\vec{g}$ is the corresponding illumination vector. Now we analyze how much the orthogonality property is violated if the shift in the alignment of the center of the array and the center of the target is small compared to the size of the active array. We will also verify this using numerical experiments in section 5.

When the distance between the target and the active array L is large compared to the size of the array s, i.e., when the aperture $\frac{s}{L}$ is small, we have $G(\boldsymbol{\xi}_i,\boldsymbol{o})\approx G(\tilde{\boldsymbol{o}},\boldsymbol{o})$,

FIG. 4.1.

where $\tilde{\boldsymbol{o}}$ is the center of the array and $\boldsymbol{o}$ is the center of the target. So

$$\vec{\tilde{g}} = [G(\boldsymbol{\xi}_1, \boldsymbol{o}), G(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, G(\boldsymbol{\xi}_N, \boldsymbol{o})]^T \approx G(\tilde{\boldsymbol{o}}, \boldsymbol{o})[1, 1, \ldots, 1]^T,$$

$$\vec{\tilde{g}}_1 = [\eta_1 G(\boldsymbol{\xi}_1, \boldsymbol{o}), \eta_2 G(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \eta_N G(\boldsymbol{\xi}_N, \boldsymbol{o})]^T \approx G(\tilde{\boldsymbol{o}}, \boldsymbol{o})[\eta_1, \eta_2, \ldots, \eta_N]^T,$$

$$\vec{\tilde{g}}_2 = [\eta_1^2 G(\boldsymbol{\xi}_1, \boldsymbol{o}), \eta_2^2 G(\boldsymbol{\xi}_2, \boldsymbol{o}), \ldots, \eta_N^2 G(\boldsymbol{\xi}_N, \boldsymbol{o})]^T \approx G(\tilde{\boldsymbol{o}}, \boldsymbol{o})[\eta_1^2, \eta_2^2, \ldots, \eta_N^2]^T.$$

Let ds be the separation space between two adjacent transducers and $|G| = |G(\tilde{\boldsymbol{o}}, \boldsymbol{o})|$; then

$$\|\vec{\tilde{g}}\| \approx \sqrt{\int_{-s/2-\delta}^{s/2-\delta} |G|^2 dy/ds} = |G|\sqrt{\frac{s}{ds}},$$

$$\|\vec{\tilde{g}_1}\| \approx \sqrt{\int_{-s/2-\delta}^{s/2-\delta} |G|^2 y^2 dy/ds} = |G|\sqrt{\frac{(s/2-\delta)^3 + (s/2+\delta)^3}{3ds}},$$

$$\|\vec{\tilde{g}_2}\| \approx \sqrt{\int_{-s/2-\delta}^{s/2-\delta} |G|^2 y^4 dy/ds} = |G|\sqrt{\frac{(s/2-\delta)^5 + (s/2+\delta)^5}{5ds}},$$

and

$$\left|\left\langle \frac{\vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}, \frac{\vec{\tilde{g}}_1}{\|\vec{\tilde{g}}_1\|} \right\rangle\right| \approx \frac{|\int_{-s/2-\delta}^{s/2-\delta} |G|^2 y dy/ds|}{\|\vec{\tilde{g}}\|\|\vec{\tilde{g}}_1\|} = \frac{[(s/2+\delta)^2 - (s/2-\delta)^2]/2}{\sqrt{[(s/2-\delta)^3 + (s/2+\delta)^3]s/3}} \le 2\sqrt{3}\frac{\delta}{s},$$

where

$$\alpha^3 + \beta^3 = (\alpha+\beta)(\alpha^2 - \alpha\beta + \beta^2),$$
$$\alpha^2 - \alpha\beta + \beta^2 \ge (\alpha+\beta)^2/4$$

is used. Similarly we have

$$\left|\left\langle \frac{\vec{\tilde{g}}_1}{\|\vec{\tilde{g}}_1\|}, \frac{\vec{\tilde{g}}_2}{\|\vec{\tilde{g}}_2\|} \right\rangle\right| \approx \frac{[(s/2+\delta)^4 - (s/2-\delta)^4]/4}{\sqrt{[(s/2-\delta)^3+(s/2+\delta)^3]/3}\sqrt{[(s/2-\delta)^5+(s/2+\delta)^5]/5}} \le 2\sqrt{15}\frac{\delta}{s},$$

where

$$\alpha^5 + \beta^5 = (\alpha+\beta)(\alpha^4 - \alpha^3\beta + \alpha^2\beta^2 - \alpha\beta^3 + \beta^4),$$
$$\alpha^4 - \alpha^3\beta + \alpha^2\beta^2 - \alpha\beta^3 + \beta^4 \ge (\alpha^2+\beta^2)^2/4$$

is used. So the orthogonality condition deteriorates approximately linearly in $\frac{\delta}{s}$. The larger the size of the active array is, the more robust the decomposition of the response matrix is. Also it appears from the numerical experiments that the even part of $\vec{\tilde{g}}_1$ is absorbed in the first singular vector as a perturbation and the odd part of $\vec{\tilde{g}}_1$ becomes the second singular vector. Since the first term, $a\vec{\tilde{g}}\vec{\tilde{g}}^T$, in the response matrix decomposition is dominant and more robust to the center shift, we can first use the leading term to estimate the center of the target and then adjust the center of the array toward the estimated center of the target to get a better estimation of the size in imaging.

5. Numerical experiments. In this section we use numerical examples to verify our analysis on the leading eigenvalues and eigenvectors of the response matrix for an extended target. In particular we will demonstrate the relation between the eigenvectors and the illumination vectors and verify the formulas for the leading eigenvalues. Some of the numerical examples also show that our analysis is quite accurate even for more general setups. In our numerical examples, the response matrix of an extended target was formed using the integral

$$P_{ij} = \int_{\Omega} G(\boldsymbol{\xi}_i, \boldsymbol{x}) G(\boldsymbol{\xi}_j, \boldsymbol{x}) d\boldsymbol{x},$$

where $\boldsymbol{\xi}_i, \boldsymbol{\xi}_j$ are the positions of the ith and jth transducers, respectively, and

$$G(\boldsymbol{\xi}, \boldsymbol{x}) = \frac{e^{ik|\boldsymbol{\xi}-\boldsymbol{x}|}}{4\pi|\boldsymbol{\xi}-\boldsymbol{x}|}$$

is the three-dimensional Green's function for a homogeneous medium. We use a simple quadrature for the integral on the target, denoted by Ω, with a grid size h that resolves both the wavelength and the target. In all of our numerical setups we use fixed wavelength $\lambda = 0.5m$ and wavenumber $k = \frac{2\pi}{\lambda} = 4\pi$. We vary the size of the active array s, the propagation distance L, and the size of the target relative to the wavelength, i.e., ka, to verify our analysis and formulas. The SVD of the response matrix is done by MATLAB. Note that the singular vectors computed by SVD in MATLAB (1) are always normalized to have a unit L_2 norm, and (2) have an arbitrary phase shift. Also the phase plot is up to a 2π shift.

In the following, we present numerical examples of one-dimensional arrays and targets in 5.1 as well as two-dimensional arrays and targets in 5.2.

5.1. One-dimensional arrays and targets. In this section we present numerical results in one dimension. We show the spectrum of the response matrix and the asymptotic formulas for the top two singular values as well as the top two singular vectors and their relations to the illumination vectors.

Example 1. In this example we show the spectrum of the response matrix for a single one-dimensional extended target. Figure 5.1 shows the loglog plot of the magnitudes of all singular values of the response matrix corresponding to a different target size a and propagation distance L. The size of the array is $s = 10m$ and the transducers are placed a half-wavelength apart; i.e., there are $\frac{2s}{\lambda} = 40$ transducers. We see that the top two singular values are well separated from each other and from the other singular values in the asymptotic regime which we discussed earlier.

Example 2. In this example we demonstrate the relation between the top two singular vectors of the response matrix and the corresponding illumination vectors. We also verify the formulas for the top two singular values numerically. The basic setup is illustrated in Figure 3.1 with $L = 500m, k = 4\pi, ka = 16, s = 40m$. The transducers are a half-wavelength apart. In Figure 5.2, we plot the magnitude and phase for each component of the top two singular vectors and compare them to the illumination vectors. Figure 5.2(a) plots the magnitude and phase of the first singular vector against the illumination vector $\vec{\tilde{g}}$ defined in (3.7). Figure 5.2(b) plots the magnitude and phase of the second singular vector against the illumination vector $\vec{\tilde{g}}_1$ defined in (3.7). The correspondence and pattern similarities are striking.

Denoting $\vec{v}_1, \vec{v}_2$ to be the top two singular vectors with singular values λ_1, λ_2

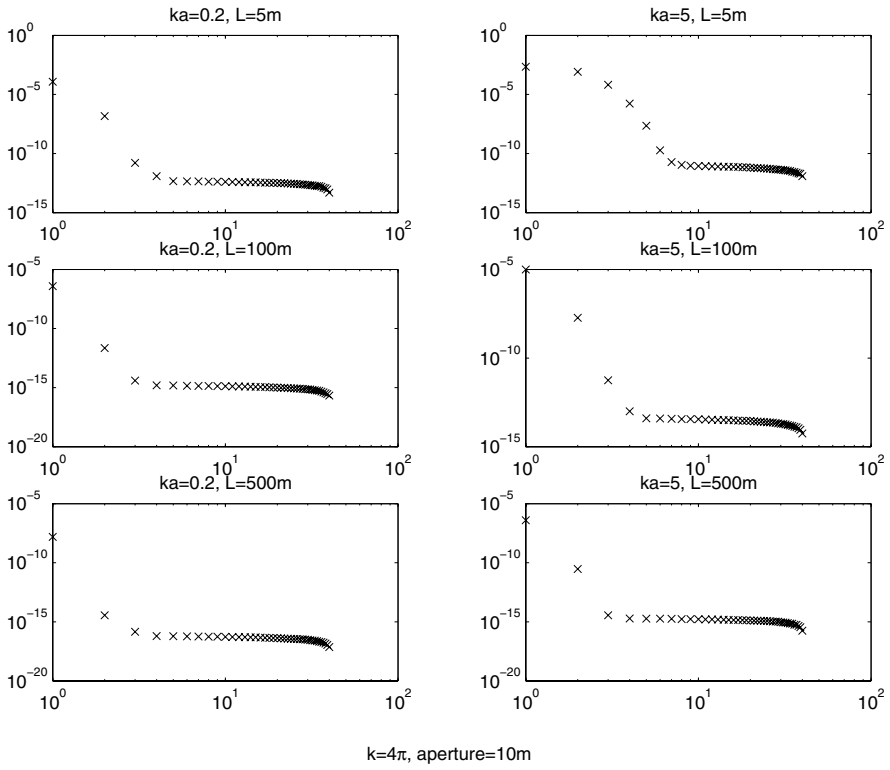


FIG. 5.1. *Plots of the magnitudes of eigenvalues.*

computed by MATLAB, we have numerically

$$\begin{bmatrix} \left|\frac{\bar{\vec{v}}_1^T \vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\bar{\vec{v}}_1^T \vec{g}_1}{\|\vec{g}_1\|}\right| \\ \left|\frac{\bar{\vec{v}}_2^T \vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\bar{\vec{v}}_2^T \vec{g}_1}{\|\vec{g}_1\|}\right| \end{bmatrix} = \begin{bmatrix} 1 & 6.2058 \times 10^{-16} \\ 1.0124 \times 10^{-15} & 1 \end{bmatrix},$$

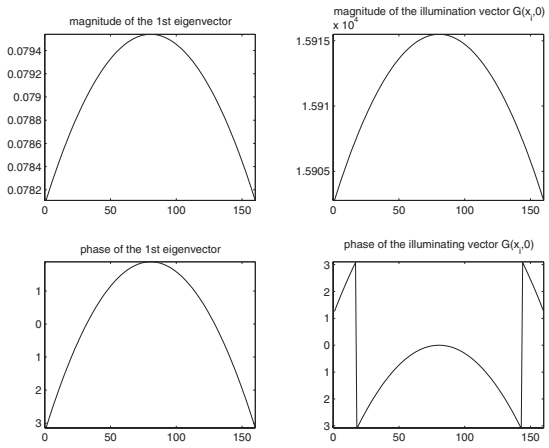
$$|\lambda_1| = 5.0848 \times 10^{-6}, \quad a\|\vec{\tilde{g}}\|^2 = 5.1575 \times 10^{-6},$$

$$|\lambda_2| = 5.9370 \times 10^{-8}, \quad \tfrac{a^3 k^2}{12L^2}\|\vec{\tilde{g}}_1\|^2 = 5.9393 \times 10^{-8}.$$

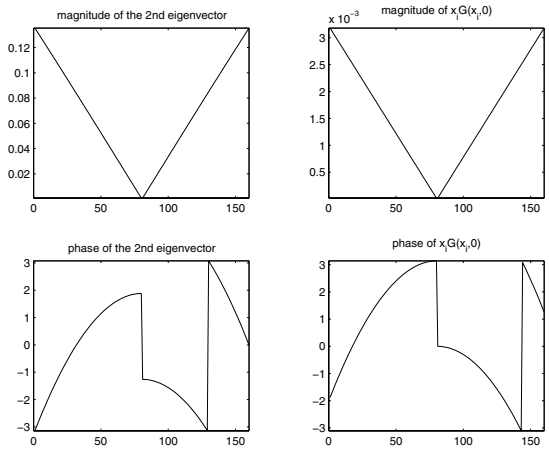
We see an almost perfect orthogonality condition up to machine accuracy. Our asymptotic formulas for the leading singular values are also very accurate. Moreover, the illumination vector $\vec{\tilde{g}}_2$ defined in (3.7) has the following relation with the first three singular vectors of the response matrix:

$$\left|\frac{\bar{\vec{v}}_1^T \vec{\tilde{g}}_2}{\|\vec{\tilde{g}}_2\|}\right| = 0.7418, \quad \left|\frac{\bar{\vec{v}}_2^T \vec{\tilde{g}}_2}{\|\vec{\tilde{g}}_2\|}\right| = 6.6258 \times 10^{-14}, \quad \left|\frac{\bar{\vec{v}}_3^T \vec{\tilde{g}}_2}{\|\vec{\tilde{g}}_2\|}\right| = 0.6706;$$

i.e., $\vec{\tilde{g}}_2$ is mostly absorbed in $\vec{v}_1$ as a small perturbation. We also see that the asymptotic formula $a\|\vec{\tilde{g}}\|^2$ overestimates $|\lambda_1|$ a little bit due to higher order perturbations, as was explained in section 3. From the following examples we can see that the overestimation tends to be more when ka becomes larger and less when L becomes larger.



(a)



(b)

FIG. 5.2. (a) *The phase and magnitude of the first eigenvector and* (b) *the phase and magnitude of the second eigenvector.*

Now we vary the setup by changing one parameter at a time. First we increase the size of the target $ka = 50$, and we have

$$\begin{bmatrix} \left|\frac{\vec{v}_1^T \vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\vec{v}_1^T \vec{g}_1}{\|\vec{g}_1\|}\right| \\ \left|\frac{\vec{v}_2^T \vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\vec{v}_2^T \vec{g}_1}{\|\vec{g}_1\|}\right| \end{bmatrix} = \begin{bmatrix} 0.9988 & 1.1532 \times 10^{-16} \\ 1.3101 \times 10^{-16} & 0.9996 \end{bmatrix},$$

$$|\lambda_1| = 1.4434 \times 10^{-5}, \quad a\|\vec{g}\|^2 = 1.6117 \times 10^{-5},$$

$$|\lambda_2| = 1.6083 \times 10^{-6}, \quad \frac{a^3 k^2}{12L^2}\|\vec{g}_1\|^2 = 1.8125 \times 10^{-6}.$$

Next we increase the distance between the array and the target L to 10,000m,

and we have

$$\begin{bmatrix} \left|\frac{\vec{v}_1^T\vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\vec{v}_1^T\vec{g}_1}{\|\vec{g}_1\|}\right| \\ \left|\frac{\vec{v}_2^T\vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\vec{v}_2^T\vec{g}_1}{\|\vec{g}_1\|}\right| \end{bmatrix} = \begin{bmatrix} 1 & 2.9737\times 10^{-17} \\ 9.8603\times 10^{-14} & 1 \end{bmatrix},$$

$$|\lambda_1| = 1.2867\times 10^{-8}, \quad a\|\vec{\tilde{g}}\|^2 = 1.2901\times 10^{-8},$$

$$|\lambda_2| = 3.7649\times 10^{-13}, \quad \tfrac{a^3k^2}{12L^2}\|\vec{\tilde{g}}_1\|^2 = 3.7157\times 10^{-13}.$$

Finally, when we increase the separation distance between the transducers to $2\lambda = 1m$, we have

$$\begin{bmatrix} \left|\frac{\vec{v}_1^T\vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\vec{v}_1^T\vec{g}_1}{\|\vec{g}_1\|}\right| \\ \left|\frac{\vec{v}_2^T\vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\vec{v}_2^T\vec{g}_1}{\|\vec{g}_1\|}\right| \end{bmatrix} = \begin{bmatrix} 1 & 1.3157\times 10^{-17} \\ 5.0919\times 10^{-16} & 1 \end{bmatrix},$$

$$|\lambda_1| = 1.2706\times 10^{-6}, \quad a\|\vec{\tilde{g}}\|^2 = 1.2893\times 10^{-6},$$

$$|\lambda_2| = 1.5401\times 10^{-8}, \quad \tfrac{a^3k^2}{12L^2}\|\vec{\tilde{g}}_1\|^2 = 1.5415\times 10^{-8}.$$

In our numerical tests it seems that we can further increase the separation between transducers.

Example 3. In this example we test how the alignment of the array and the target affects our formula for the response matrix decomposition. Figure 5.3 shows plots of the top two singular vectors when the center of the target is not aligned with the center of the array with a shift of $3m$. The setup is the same as the above example with $L = 500m,\ s = 40m$. We see asymmetries in the plots. We have numerically

$$\begin{bmatrix} \left|\frac{\vec{v}_1^T\vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\vec{v}_1^T\vec{g}_1}{\|\vec{g}_1\|}\right| \\ \left|\frac{\vec{v}_2^T\vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\vec{v}_2^T\vec{g}_1}{\|\vec{g}_1\|}\right| \end{bmatrix} = \begin{bmatrix} 1 & 0.2440 \\ 0.0059 & 0.9698 \end{bmatrix},$$

$$|\lambda_1| = 5.0770\times 10^{-6}, \quad a\|\vec{\tilde{g}}\|^2 = 5.1573\times 10^{-6},$$

$$|\lambda_2| = 5.9377\times 10^{-8}, \quad \tfrac{a^3k^2}{12L^2}\|\vec{\tilde{g}}_1\|^2 = 6.3343\times 10^{-8}.$$

When we increase the shift of the center to $6m$, we have

$$\begin{bmatrix} \left|\frac{\vec{v}_1^T\vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\vec{v}_1^T\vec{g}_1}{\|\vec{g}_1\|}\right| \\ \left|\frac{\vec{v}_2^T\vec{g}}{\|\vec{g}\|}\right| & \left|\frac{\vec{v}_2^T\vec{g}_1}{\|\vec{g}_1\|}\right| \end{bmatrix} = \begin{bmatrix} 0.9999 & 0.4479 \\ 0.0119 & 0.8940 \end{bmatrix},$$

$$|\lambda_1| = 5.0536\times 10^{-6}, \quad a\|\vec{\tilde{g}}\|^2 = 5.1567\times 10^{-6},$$

$$|\lambda_2| = 5.9398\times 10^{-8}, \quad \tfrac{a^3k^2}{12L^2}\|\vec{\tilde{g}}_1\|^2 = 7.5192\times 10^{-8}.$$

As was analyzed in section 4, the orthogonality condition does not hold anymore. The sensitivity depends linearly on the shift relative to the size of the array which

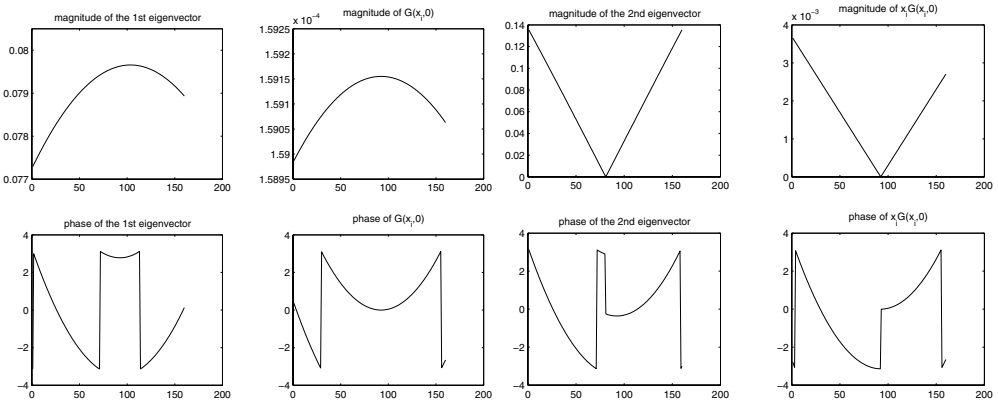


FIG. 5.3. (a) *The phase and magnitude of the first eigenvector and* (b) *the phase and magnitude of the second eigenvector.*

agrees with the numerical results very well. From the tests we can also see that the first singular vector is quite robust and agrees with the illumination vector $\vec{\tilde{g}}$ fairly well. The second singular vector is more sensitive to the shift. It appears that the second singular vector contains only the asymmetric part of $\vec{\tilde{g}}_1$.

Example 4. In this example, we show that we can get subwavelength information about the size of a target. The setup is: $L = 10m, k = 4\pi, ka = 0.1, s = 2m$. There are only eight transducers in the array. Numerically we have

$$\begin{bmatrix} \left|\frac{\vec{\overline{v}}_1^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\overline{v}}_1^T \vec{\tilde{g}}_1}{\|\vec{\tilde{g}}_1\|}\right| \\ \left|\frac{\vec{\overline{v}}_2^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\overline{v}}_2^T \vec{\tilde{g}}_1}{\|\vec{\tilde{g}}_1\|}\right| \end{bmatrix} = \begin{bmatrix} 1 & 5.0849 \times 10^{-17} \\ 4.2743 \times 10^{-10} & 1 \end{bmatrix},$$

$$|\lambda_1| = 3.7834 \times 10^{-6}, \quad a\|\vec{\tilde{g}}\|^2 = 4.0143 \times 10^{-6},$$

$$|\lambda_2| = 1.8716 \times 10^{-11}, \quad \frac{a^3k^2}{12L^2}\|\vec{\tilde{g}}_1\|^2 = 1.4290 \times 10^{-11}.$$

The actual size of the target is $a = 0.008m$. From our formula, the estimation of the size is: $\frac{|\lambda_1|}{\|\vec{\tilde{g}}\|^2} = 0.0075m$, or, $(\frac{12L^2|\lambda_2|}{k^2\|\vec{\tilde{g}}_1\|^2})^{1/3} = 0.0087m$. Clearly we achieve subwavelength accuracy from the experiment.

5.2. Two-dimensional arrays and targets. In this section we present numerical experiments for two-dimensional extended targets. The harmonic wave used is as before with fixed wavelength $\lambda = 0.5m$ and wavenumber $k = 4\pi$.

Example 1. In this example, we use a square array, each side of which is $10m$ long. The transducers are placed on a rectangular grid whose grid size is one wavelength; i.e., there are 20 rows and 20 columns of transducers and the total number of transducers is 400. So the size of the response matrix is 400×400. The distance between the array and the target is $L = 500m$, and the target is an ellipse with two major axes which are $2a = 1.2732m(ka = 8)$ and $2b = 0.6366m(kb = 4)$ long, respectively.

In the first test, the center and two sides of the square array are aligned with the center and two major axes of the ellipse. We define the two sides of the array as y and

z axes, respectively, with the y axis parallel to the longer major axis of the ellipse. Figure 5.4(a) is the plot of the magnitudes of the singular values of the response matrix. We see the top three singular values are well separated from the other ones. Figure 5.5 compares the amplitude and phase of the top two singular vectors with the two illumination vectors defined in (3.8). We skip the plot of the third singular vector and the corresponding illumination vector since it is similar to the plot of the second singular vector. We see that for each row or column of the transducers, the plot is very similar to the previous plots for one-dimensional arrays and targets.

Let $\vec{v}_1, \vec{v}_2, \vec{v}_3$ be the top three eigenvectors with eigenvalues $\lambda_1, \lambda_2, \lambda_3$, and $\vec{\tilde{g}}, \vec{\tilde{g}}_{1y}, \vec{\tilde{g}}_{1z}$ be the three illumination vectors defined in (3.8). We have numerically

$$\begin{bmatrix} \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \\ \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \\ \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \end{bmatrix} = \begin{bmatrix} 1 & 3.9885 \times 10^{-17} & 2.1706 \times 10^{-17} \\ 7.1030 \times 10^{-17} & 1 & 3.3406 \times 10^{-15} \\ 2.4696 \times 10^{-17} & 3.3394 \times 10^{-15} & 1 \end{bmatrix},$$

$$|\lambda_1| = 6.0744 \times 10^{-6}, \quad |\Omega| \|\vec{\tilde{g}}\|^2 = \pi ab \|\vec{\tilde{g}}\|^2 = 6.4498 \times 10^{-6},$$

$$|\lambda_2| = 3.5652 \times 10^{-9}, \quad \tfrac{k^2}{L^2} \|\vec{\tilde{g}}_{1y}\|^2 \textstyle\int_\Omega y^2 = \frac{\pi a^3 b k^2}{4L^2} \|\vec{\tilde{g}}_{1y}\| = 3.8019 \times 10^{-9},$$

$$|\lambda_3| = 8.8290 \times 10^{-10}, \quad \tfrac{k^2}{L^2} \|\vec{\tilde{g}}_{1z}\|^2 \textstyle\int_\Omega z^2 = \frac{\pi a b^3 k^2}{4L^2} \|\vec{\tilde{g}}_{1z}\| = 9.5047 \times 10^{-10}.$$

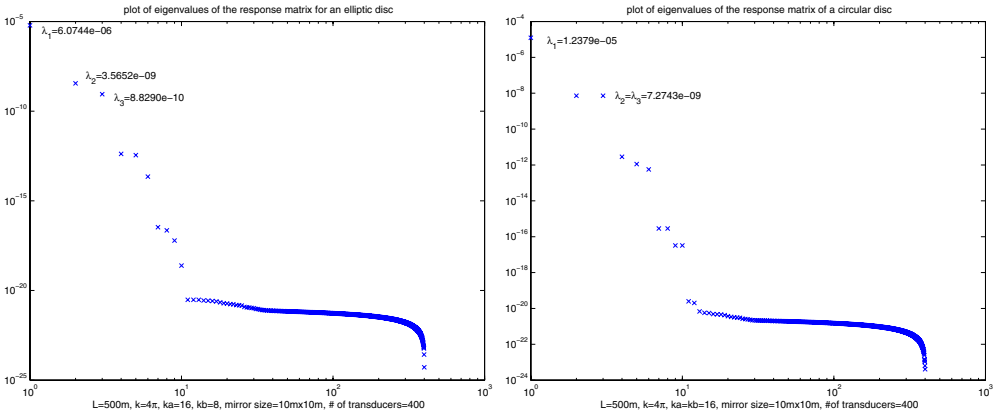
Now we rotate the ellipse by $\frac{\pi}{6}$ so that the two sides of the square array, i.e., our defined y and z axes, are no longer parallel to the two major axes of the ellipse. We have numerically

$$\begin{bmatrix} \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \\ \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \\ \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \end{bmatrix} = \begin{bmatrix} 1 & 9.0735 \times 10^{-17} & 2.3222 \times 10^{-17} \\ 5.4091 \times 10^{-17} & 0.8661 \ (\approx \frac{\sqrt{3}}{2}) & 0.4999 \ (\approx \frac{1}{2}) \\ 3.6460 \times 10^{-17} & 0.4999 & 0.8661 \end{bmatrix},$$

$$|\lambda_1| = 6.0744 \times 10^{-6}, \quad |\lambda_2| = 3.5648 \times 10^{-9}, \quad |\lambda_3| = 8.8302 \times 10^{-10}.$$

The reason that we do not have orthogonality correspondence between $\vec{v}_2, \vec{v}_3$ and $\vec{\tilde{g}}_{1y}, \vec{\tilde{g}}_{1z}$ is because the y and z axes that we choose artificially do not correspond to those two major axes, which are the two intrinsic symmetry lines of the ellipse. The angle ($\frac{\pi}{6}$) between $\vec{v}_2$ and $\vec{\tilde{g}}_{1y}$ is exactly the angle between the y axis and the longer major axis. All the other numerical numbers in the above matrix become clear too. The amplitudes of the three top singular values are not changed since the center of the target is unchanged and so are the integrals on the target for the eigenvalues. The numerical results verify our analysis perfectly. Both the dimension and the symmetry of the target are embedded in the response matrix. The above orthogonality matrix can be used to find the line of symmetry of the object in imaging.

Now we move the elliptic target further away from the array with $L = 2000m$.



(a) (b)

FIG. 5.4. *Eigenvalue plot for a rectangular array:* (a) *for an elliptic target and* (b) *for a circular target.*

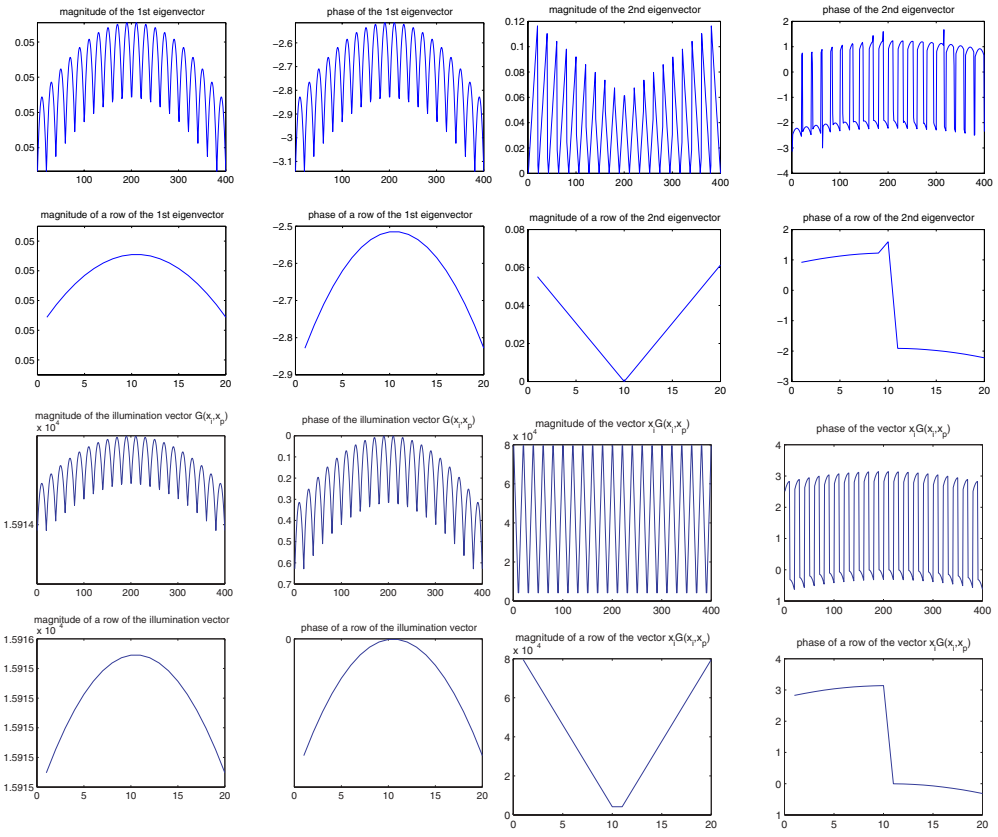


FIG. 5.5. *Amplitude and phase plot of the top eigenvectors and the corresponding illumination vectors.*

We have

$$|\lambda_1| = 1.4469 \times 10^{-6}, \quad |\Omega|\|\vec{\tilde{g}}\|^2 = \pi ab\|\vec{\tilde{g}}\|^2 = 1.6126 \times 10^{-6},$$

$$|\lambda_2| = 5.0757 \times 10^{-11}, \quad \tfrac{k^2}{L^2}\|\vec{\tilde{g}}_{1y}\|^2 \int_\Omega y^2 = \tfrac{\pi a^3 bk^2}{4L^2}\|\vec{\tilde{g}}_{1y}\| = 5.6509 \times 10^{-11},$$

$$|\lambda_3| = 1.2402 \times 10^{-11}, \quad \tfrac{k^2}{L^2}\|\vec{\tilde{g}}_{1z}\|^2 \int_\Omega z^2 = \tfrac{\pi ab^3 k^2}{4L^2}\|\vec{\tilde{g}}_{1z}\| = 1.4127 \times 10^{-11}.$$

Example 2. In this example, we replace the elliptic target by a circular one with radius $a = b = 0.6366m(ka = kb = 8)$ and $L = 500m$. The eigenvalue of the response matrix is plotted in Figure 5.4(b). We have

$$\begin{bmatrix} \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \\ \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \\ \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \end{bmatrix} = \begin{bmatrix} 1 & 2.8620 \times 10^{-17} & 2.7358 \times 10^{-17} \\ 3.9876 \times 10^{-17} & 0.7071(\approx \frac{\sqrt{2}}{2}) & 0.7071 \\ 1.4545 \times 10^{-16} & 0.7071 & 0.7071 \end{bmatrix},$$

$$|\lambda_1| = 1.2379 \times 10^{-5}, \qquad |\Omega|\|\vec{\tilde{g}}\|^2 = \pi ab\|\vec{\tilde{g}}\|^2 = 1.2900 \times 10^{-5},$$

$$|\lambda_2| = |\lambda_3| = 7.2743 \times 10^{-9}, \quad \tfrac{k^2}{L^2}\|\vec{\tilde{g}}_{1y}\|^2 \int_\Omega y^2 = \tfrac{\pi a^3 bk^2}{4L^2}\|\vec{\tilde{g}}_{1y}\| = 7.6038 \times 10^{-9}.$$

In this case, the reason that we do not have orthogonality correspondence between $\vec{v}_2, \vec{v}_3$ and $\vec{\tilde{g}}_{1y}, \vec{\tilde{g}}_{1z}$ is due to grid orientation of our square array. Any two orthogonal radial directions can be the symmetric axes for the circular disc; our square array picks up those two diagonal ones. Or we can interpret it in the following way: since the second and third singular values are equal, the associated singular vectors are not unique and any linear combination of the singular vectors is also a singular vector. MATLAB just chooses a particular combination.

Example 3. In this example, we use the same setup as in Example 1 except that the center of the array is not aligned with the center of the target. The two major axes of the ellipse are parallel to the two sides of the square array. But the center of the ellipse is shifted by $(2m, 2m)$ in the yz plane. The distance is still $L = 500m$. We have numerically

$$\begin{bmatrix} \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_1^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \\ \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_2^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \\ \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}}{\|\vec{\tilde{g}}\|}\right| & \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}_{1y}}{\|\vec{\tilde{g}}_{1y}\|}\right| & \left|\frac{\vec{\bar{v}}_3^T \vec{\tilde{g}}_{1z}}{\|\vec{\tilde{g}}_{1z}\|}\right| \end{bmatrix} = \begin{bmatrix} 1 & 0.5496 & 0.5501 \\ 7.7302 \times 10^{-4} & 0.8354 & 4.9318 \times 10^{-4} \\ 1.9175 \times 10^{-4} & 3.8425 \times 10^{-5} & 0.8351 \end{bmatrix},$$

$$|\lambda_1| = 6.0703 \times 10^{-6}, \quad |\Omega|\|\vec{\tilde{g}}\|^2 = \pi ab\|\vec{\tilde{g}}\|^2 = 6.4496 \times 10^{-6},$$

$$|\lambda_2| = 3.5646 \times 10^{-9}, \quad \tfrac{k^2}{L^2}\|\vec{\tilde{g}}_{1y}\|^2 \int_\Omega y^2 = \tfrac{\pi a^3 bk^2}{4L^2}\|\vec{\tilde{g}}_{1y}\| = 5.4526 \times 10^{-9},$$

$$|\lambda_3| = 8.8251 \times 10^{-10}, \quad \tfrac{k^2}{L^2}\|\vec{\tilde{g}}_{1z}\|^2 \int_\Omega z^2 = \tfrac{\pi ab^3 k^2}{4L^2}\|\vec{\tilde{g}}_{1z}\| = 1.3632 \times 10^{-9}.$$

Again we see that the asymptotic formulas for the first singular vector and singular value are more robust with respect to the center shift.

6. Conclusion. The response matrix of an active array for an extended target is studied. It is shown that the leading singular values and their corresponding singular vectors of the response matrix are related to the location and geometry of the extended target. Asymptotic formulas are derived for the leading singular values and singular vectors. Here we consider only a homogeneous medium and a single target. In the future we will study the effect of random inhomogeneities and multiple targets. We will develop imaging procedures that can detect both locations and sizes of extended targets using an active array.

Acknowledgments. The author would like to thank Chrysoula Tsogka for an interesting discussion that started the work. The author would also like to thank the referees for many helpful suggestions.

REFERENCES

[1] J. Berryman, L. Borcea, G. Papanicolaou, and C. Tsogka, *Statistically stable ultrasonic imaging in random media*, J. Acoust. Soc. Am., 112 (2002), pp. 1509–1522.
[2] P. Blomgren, G. Papanicolaou, and H. Zhao, *Super-resolution in time-reversal acoustics*, J. Acoust. Soc. Am., 111 (2002), pp. 230–248.
[3] D. Chambers and A. Gautesen, *Time reversal for a single spherical scatter*, J. Acoust. Soc. Am., 109 (2001), pp. 2616–2624.
[4] D. H. Chambers, *Analysis of the time-reversal operator for scatterers of finite size*, J. Acoust. Soc. Am., 112 (2002), pp. 411–419.
[5] A. Derode, P. Roux, and M. Fink, *Robust acoustic time reversal with high-order multiple scattering*, Phys. Rev. Lett., 75 (1995), pp. 4206–4209.
[6] Devaney, *Super-resolution processing of multi-static data using time-reversal and MUSIC*, J. Acoust. Soc. Am., to appear.
[7] D. R. Dowling and D. R. Jackson, *Narrow-band performance of phase-conjugate arrays in dynamic random media*, J. Acoust. Soc. Am., 91 (1992), pp. 3257–3277.
[8] M. Fink, *Time reversed acoustics*, Phys. Today, 50 (1997), pp. 34–40.
[9] W. S. Hodgkiss, H. C. Song, W. A. Kuperman, T. Akal, C. Ferla, and D. R. Jackson, *A long-range and variable focus phase-conjugation experiment in shallow water*, J. Acoust. Soc. Am., 105 (1999), pp. 1597–1604.
[10] A. Ishimaru, *Wave Propagation and Scattering in Random Media*, Academic Press, New York, 1978.
[11] W. A. Kuperman, W. Hodgkiss, H. C. Song, T. Akal, C. Ferla, and D. R. Jackson, *Phase conjugation in the ocean*, J. Acoust. Soc. Am., 102 (1997), pp. 1–16.
[12] N. Mordant, C. Prada, and M. Fink, *Highly resolved detection and selective focusing in a waveguide using the D.O.R.T. method*, J. Acoust. Soc. Am., 105 (1999), pp. 2634–2642.
[13] C. Prada, S. Manneville, D. Spoliansky, and M. Fink, *Decomposition of the time reversal operator: Detection and selective focusing on two scatterers*, J. Acoust. Soc. Am., 99 (1996), pp. 2067–2076.
[14] C. Prada, J.-L. Thomas, and M. Fink, *The iterative time reversal process: Analysis of the convergence*, J. Acoust. Soc. Am., 97 (1995), pp. 62–71.
[15] C. Prada, F. Wu, and M. Fink, *The iterative time reversal mirror: A solution to self-focusing in the pulse echo mode*, J. Acoust. Soc. Am., 90 (1991), pp. 1119–1129.
[16] H. C. Song, W. A. Kuperman, W. S. Hodgkiss, T. Akal, and C. Ferla, *Iterative time reversal in the ocean*, J. Acoust. Soc. Am., 105 (1999), pp. 3176–3184.
[17] H. Tortel, G. Micolau, and M. Saillard, *Decomposition of the time reversal operator for electromagnetic scattering*, J. Electromagn. Waves Appl., 13 (1999), pp. 687–719.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 746–761

ASYMPTOTIC BEHAVIOR OF THE NUMBER OF LOST MESSAGES*

VYACHESLAV M. ABRAMOV†

Abstract. The goal of the paper is to study asymptotic behavior of the number of lost messages. Long messages are assumed to be divided into a random number of packets which are transmitted independently of one another. An error in transmission of a packet results in the loss of the entire message. Messages arrive to the $M/GI/1$ finite buffer model and can be lost in two cases as either at least one of its packets is corrupted or the buffer is overflowed. With the parameters of the system typical for models of information transmission in real networks, we obtain theorems on asymptotic behavior of the number of lost messages. We also study how the loss probability changes if redundant packets are added. Our asymptotic analysis approach is based on Tauberian theorems with remainder.

Key words. loss systems, $M/GI/1/n$ queue, busy period, redundancy, loss probability, asymptotic analysis, Tauberian theorems with remainder

AMS subject classifications. 60K25, 60K30, 40E05

DOI. 10.1137/S0036139902405250

1. Introduction.

1.1. Review of the literature and general description of the system.

Long messages in Internet protocols that have to be transmitted are divided into small packets. Upon transmission each packet is transformed by providing additional information related to a given message. Because of the bit errors in transmission of the packet, the message can be lost. The loss probability of a message plays a significant role in the evaluation of network performance and design of network topology.

There are a number of papers in which the loss probability of a message has been studied. Cidon, Khamisy, and Sidi [11] derived recurrence relations for the loss probabilities of packets in a message giving the numerical results for the $M/M/1/n$ buffer model. The complexity of recurrence calculations of that paper are $O(nm^2)$, where m is the size of a message and n is the buffer capacity. Considering the same model, Gurewitz, Sidi, and Cidon [13] obtained another representation for the loss probability by using the ballot theorem (e.g., Takács [17]). In the framework of the same model Altman and Jean-Marie [7] give a comprehensive analysis for the multidimensional generating function of the loss probabilities based on the recurrence relations of the paper of Cidon, Khamisy, and Sidi [11] and analyze the effect of adding redundant packets. Studying a slightly more general model with several sources, Ait-Hellal et al. [6] obtained some asymptotic results and studied the effect of adding redundancy to the loss probability. The aforementioned papers [6], [7], [11], [13] all discuss the problem of complexity of calculations as well as the required memory to store intermediate variables.

In real communication networks the capacity is large. Therefore, asymptotic analysis of the number of lost messages is necessary. The present paper provides asymptotic analysis with sequential application to redundancy of the following model. Assume that each message is divided into a random number of packets each of which

*Received by the editors April 9, 2002; accepted for publication (in revised form) July 21, 2003; published electronically March 11, 2004.
http://www.siam.org/journals/siap/64-3/40525.html
†Department of Mathematics, The Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel, and College of Judea and Samaria, Ariel 44837, Israel (vyachesl@inter.net.il).

is forwarded to the buffer. For the ith message denote its random number of packets by ν_i. We assume that the sequence $\nu_i \geq 1$, $i \geq 1$, consists of independently and identically distributed integer random variables. The interarrival times between messages have an exponential distribution with parameter λ. The buffer can contain only N packets; that is, if immediately before the arrival of message of l packets there are L packets in the buffer, then the message is accepted only if $L + l \leq N$; otherwise the message of l packets is lost. The loss of a message can also occur if at least one packet in a message is corrupted. In this case we assume that if there is enough space, then the message does occupy the buffer, but it is hidden and therefore lost. The probability that at least one packet in a message is corrupted is denoted by p.

In general loss communication networks, a transmission time typically depends on the number of packets in a message. To be realistic we must study a general queueing system with service time depending on batch size. The analysis of such a system is a hard problem. On the other hand, the model with a fixed number of packets in a message, leading to the standard $M/GI/1/n$ queueing system, is not realistic. Therefore, in the following we assume additionally that the random variables ν_i have fixed upper and lower bounds ν^{upper} and ν^{lower}, i.e., $\mathbf{P}\{\nu^{lower} \leq \nu_i \leq \nu^{upper}\} = 1$. This assumption can be considered as a compromise between these two cases. It has a real application in some communication technologies, especially in optical local networks, where a number of small messages following the same direction are combined as one message (bus).[1] Outgoing from the local network, the bus continues on its way being processed by the Internet protocols. When the difference between ν^{upper} and ν^{lower} for the messages is not large, then assuming that a transmission time is independent of the message size seems appropriate.

1.2. Formulation of the model in terms of the queueing theory. In terms of the queueing theory the model can be described as follows. We assume that messages arrive to the finite buffer $M/GI/1$ queue with random number of waiting places ζ. The input rate is equal to λ, and the service time distribution is $B(x)$ with the expectation b. By a queueing system with random number of waiting places we mean the following. We denote

$$\zeta = \inf\left\{ m : \sum_{i=1}^{m} \nu_i \leq N \right\},$$

and according to the assumption $\mathbf{P}\{\nu^{lower} \leq \nu_i \leq \nu^{upper}\} = 1$, there are two fixed values ζ^{upper} and ζ^{lower} depending on N, and $\mathbf{P}\{\zeta^{lower} \leq \zeta \leq \zeta^{upper}\} = 1$.

Let $\zeta_1, \zeta_2, \ldots,$ be a strictly stationary and ergodic sequence of random variables, $\mathbf{P}\{\zeta_i = j\} = \mathbf{P}\{\zeta = j\}$, $\zeta^{lower} \leq j \leq \zeta^{upper}$. If ξ_i is the number of messages in the queue immediately before arrival of the ith message, then the message is lost if $\xi_i > \zeta_i$. Otherwise it joins the queue. We assume that $\xi_1 = 0$.

The existence of the stationary queue-length distribution, i.e.,

$$\mathbf{P}\{\bar{q} = j\} = \lim_{i \to \infty} \mathbf{P}\{\xi_i = j | \xi_1 < \infty\}, \qquad j = 0, 1, \ldots, \zeta^{upper}, \tag{1.1}$$

is shown in the following. The special case when $\mathbf{P}\{\nu_i = l\} = 1$ leads to the standard $M/GI/1/n$ queueing system, where $n = \lfloor N/l \rfloor$ is the integer part of N/l.

[1] For example, one of such technologies was developed in Orika Optical Networks Limited, where the author worked during 2000–2001.

It is also assumed that each message is marked with probability p. We study the asymptotic behavior of the loss probability under assumptions that $\mathbf{E}\zeta$ increases to infinity and p vanishes. The details of these assumptions are clarified in the following consideration. The loss probability is the probability that the message is either marked or lost because of overflowing the queue. We study the cases where the traffic (offered load) $\varrho = \lambda b$ is less than, equal to, and greater than 1.

1.3. Advantages of the approach and methodology. Our approach is based on the asymptotic analysis of the loss queueing systems in the earlier paper of the author (see Abramov [3]). The main method is an application of modern Tauberian theorems with remainder. For the relevant works devoted to asymptotic analysis of the loss and controlled systems with Poisson input, see Abramov [1], [2], Tomkó [18], and other papers. The asymptotic analysis of the $GI/M/1/n$ queueing system was studied in [4], [9], [10]. The advantages of the approach of the present paper are the following.

First, our model is more general than the model from the aforementioned papers: This paper discusses the case of a non-Markovian buffer model where a message contains a random batch of packets, while the aforementioned papers studied a Markovian model with fixed batch size.

Second, the work in [6], [7], [11], [13] discusses a more difficult problem of consecutive losses, remaining in a framework of the standard $M/M/1/n$ queueing system. The present paper flexibly discusses the stationary losses for a nonstandard queueing model with the random number of waiting places. That queueing system belongs to the special class of queueing systems with losses that is exactly defined below.

Third, our asymptotic analysis is much simpler than that of the other papers; our final results and their representation are simple and clear as well.

The traditional approach to asymptotic analysis, based on the final value theorem for z transform, enables us to obtain the main term of asymptotic relation and, in certain cases, a remainder. The modern Tauberian theorems enable us to obtain stronger asymptotic relations using some additional assumptions. These additional assumptions are realistic for the queueing systems considered here, and our asymptotic results are stronger than the earlier asymptotic results obtained for the $M/GI/1/n$ queueing system with the aid of the final value theorem for z transform (see relation (4.15) for its comparison with (4.14)). For some other results related to asymptotic analysis of the $M/GI/1/n$ and $GI/M/1/n$ queueing systems with the aid of the final value theorem, see the bibliography notes and references in Abramov [1].

1.4. What is the main result in this paper? The paper contains a number of theoretical results on the asymptotic behavior of characteristics of the busy period of the system (section 4) and loss probability (section 5). These theoretical results are then used to conclude the effect of adding redundant packets in order to decrease the loss probability.

Although the theoretical results of the paper, related to the cases where the offered load $\varrho < 1$, are standard, the conclusion about adding redundancy is extremely simple and interesting nevertheless. Namely, the stationary loss probability is expressed only via the probability that there is a corrupted packet in the message. This enables us to conclude that adding a number of redundant packets can decrease the loss probability *with the rate of geometric progression* while $\varrho < 1$.

Then the case when ϱ is close to 1 is very important for the performance analysis. For example, it can be a result of adding a number of redundant packets when initially

$\varrho < 1$. That is, to achieve a maximum decrease in the loss probability we allow an increase in the offered load up to the critical value.

Therefore, the results on redundancy, related to the case where $\varrho = 1+\varepsilon$ ($\varepsilon > 0$) is slightly greater than 1, are extremely important. The usefulness of the case $\varrho = 1+\varepsilon$ is that it enables us to obtain more exact conclusions on redundancy based on asymptotic results with remainder. Then, the usefulness of the purely theoretical case $\varrho > 1$ is that it is an intermediate result helping us to study the transient behavior, related to the case $\varrho = 1 + \varepsilon$ for small $\varepsilon > 0$.

1.5. Conclusion on adding redundant packets. The results of the paper enable us to make conclusions on the effect of adding redundant packets as follows. Let ϱ denote the offered load of the system before adding a redundant packet, and let $\breve{\varrho} > \varrho$ be the value of the offered load after adding a redundant packet. While $\breve{\varrho}$ remains not greater than 1, adding redundant packets is profitable. It decreases the loss probability with the rate of a geometric progression. Adding a redundant packet remains profitable if the value $\breve{\varrho} = 1 + \varepsilon$, where ε is a small value of a higher order than p. In some cases adding a redundant packet decreases the loss probability even when the value ε has the same order as p. These cases are studied in section 6.

1.6. The organization of the paper. The paper is organized as follows. There are six sections, with the first an introduction. In section 2 we introduce the class of queueing systems with a random number of waiting places and study the characteristics of the system busy period. The results on the expectations of random variables of the busy period (the number of processed messages, the number of refused messages, etc.) are given by Lemma 2.1. In section 3 we present a number of auxiliary results and the Tauberian theorems with remainder. These results are then used to prove a number of theorems on the asymptotic behavior of the characteristics of the system given on a busy period which in turn are given in section 4. Section 5 presents the results on asymptotic behavior of the loss probabilities under different assumptions. In section 6 we discuss adding redundancy. The central question here is, *How is the loss probability decreased or increased if we add redundant packets into the message*?

2. Characteristics of the system given on a busy period. The aim of this section is to deduce the explicit representations for characteristics of the system during a busy period such as the expected duration of a busy period, expected number of served and lost customers during a busy period, and so on. The queueing system described in section 1.2 is not standard, and the explicit representation for its characteristics cannot be obtained traditionally. Therefore, below we introduce a special class of queueing systems Σ containing the system studied in the paper and described in section 1.2. It will be shown in this section that the above characteristics are the same for all queueing systems of the class Σ. Hence, one can take any queueing system, a representative of class Σ, having a more simple structure than the original system, and study it instead of the original system.

For the sake of convenience, we denote by $\mathcal{S}_1$ the system described in section 1.2. Let $B(x)$ be the probability distribution function of a processing time (in the queueing terminology, a service time), and let λ be the parameter of Poisson input. We also set $\varrho_j = \lambda^j \int_0^\infty x^j \mathrm{d}B(x)$, $j = 1, 2, \ldots$, and $\varrho_1 = \varrho$.

In order to study the characteristics of the system $\mathcal{S}_1$ we introduce a set of systems Σ containing $\mathcal{S}_1$ as an element. The set Σ is a set of $M/GI/1$ queueing systems where λ is the rate of Poisson input, $B(x)$ is the probability distribution function of a service time, and the family of sequences $\{\zeta_i\}$ is more general than in $\mathcal{S}_1$. Each sequence ζ_1,

$\zeta_2, \ldots$ is a family of identically distributed random variables, governing the rejection process and having the same distribution as the random variable ζ. If this sequence is as defined in section 1.2, then we have a description of our system $\mathcal{S}_1$. In order to define the set Σ more exactly, we use the notation for the queueing system $\mathcal{S}_1$ and also introduce the following.

Let ξ_i denote the number of messages in the system $\mathcal{S}_1$ immediately before arrival of the ith message, $\xi_1 = 0$, and let s_i denote the number of service completions between the ith and $i+1$st arrivals. It is clear that

$$\xi_{i+1} = \xi_i - s_i + \mathbf{I}\{\xi_i \le \zeta_i\}, \tag{2.1}$$

where the term $\mathbf{I}\{\xi_i \le \zeta_i\}$ indicates that the ith message is accepted, and obviously s_i is not greater than $\xi_i + \mathbf{I}\{\xi_i \le \zeta_i\}$.

Consider a new queueing system $\mathcal{S}$ as above with the Poisson input rate λ and the probability distribution function of a service $B(x)$, but with the sequence $\widetilde{\zeta}_1, \widetilde{\zeta}_2, \ldots$. Here we assume that the sequence $\{\widetilde{\zeta}_i\}$ is an *arbitrary dependent* sequence of random variables consisting of identically distributed random variables as the random variable ζ. Let $\widetilde{\xi}_i$ denote the number of messages immediately before arrival of the ith message ($\widetilde{\xi}_1 = 0$), and let $\widetilde{s}_i$ denote the number of service completions between the ith and $i+1$st arrivals. Thus, we assume that the initial conditions of both queueing systems $\mathcal{S}_1$ and $\mathcal{S}$ are the same: $\xi_1 = \widetilde{\xi}_1$.

Analogously to (2.1) we have

$$\widetilde{\xi}_{i+1} = \widetilde{\xi}_i - \widetilde{s}_i + \mathbf{I}\{\widetilde{\xi}_i \le \widetilde{\zeta}_i\}. \tag{2.2}$$

Definition. We say that the queueing system $\mathcal{S}$ belongs to the set Σ of queueing systems if $\mathbf{E}\widetilde{\xi}_i = \mathbf{E}\xi_i$, $\mathbf{E}\widetilde{s}_i = \mathbf{E}s_i$, and $\mathbf{P}\{\widetilde{\xi}_i \le \widetilde{\zeta}_i\} = \mathbf{P}\{\xi_i \le \zeta_i\}$ for all $i \ge 1$.

Consider an example of queueing systems belonging to the set Σ, where the sequence $\{\widetilde{\zeta}_i\}$ is strictly stationary but not ergodic. The example is a queueing system with $\widetilde{\zeta}_1 = \widetilde{\zeta}_2 = \ldots$, which we denote by $\mathcal{S}_2$. The example below is artificial rather than realistic, however, its main goal is to help us to show the existence of necessary stationary queue-length probabilities for the queueing system $\mathcal{S}_1$ and to obtain the explicit representations for those probabilities as well.

For $\mathcal{S}_2$ we find by induction for all $i \ge 1$ that

$$\mathbf{E}\widetilde{s}_i = \mathbf{E}s_i, \tag{2.3}$$

$$\mathbf{P}\{\widetilde{\xi}_i \le \widetilde{\zeta}_i\} = \mathbf{P}\{\xi_i \le \zeta_i\}, \tag{2.4}$$

and

$$\mathbf{E}\widetilde{\xi}_i - \mathbf{E}\widetilde{s}_i + \mathbf{P}\{\widetilde{\xi}_i \le \widetilde{\zeta}_i\} = \mathbf{E}\xi_i - \mathbf{E}s_i + \mathbf{P}\{\xi_i \le \zeta_i\}. \tag{2.5}$$

Relations (2.3)–(2.5) show that the queueing system $\mathcal{S}_2 \in \Sigma$. It follows from the definition that if the stationary loss probability exists for at most one of the queueing systems $\mathcal{S} \in \Sigma$, then it exists for all queueing systems of Σ and it is the same. Then, the properties of the queueing system $\mathcal{S}_2$ enable us to conclude similar properties of all queueing systems belonging to the set Σ, including $\mathcal{S}_1$. For example, it is not difficult to show that the expected busy period is the same for all queueing systems of the class Σ. Indeed, let $\widetilde{A}$, $\widetilde{S}$, and $\widetilde{R}$ denote the number of arrived, served, and refused

customers (because of overflowing the buffer) during a busy cycle $\widetilde{C}$, respectively. We have the equations

$$\mathbf{E}\widetilde{A} = \mathbf{E}\widetilde{S} + \mathbf{E}\widetilde{R} = \lambda \mathbf{E}\widetilde{C}, \tag{2.6}$$

$$b\mathbf{E}\widetilde{S} = \mathbf{E}\widetilde{C} - \frac{1}{\lambda}, \tag{2.7}$$

where b is the expected service time. Since the loss probability is the same for all queueing systems $\mathcal{S} \in \Sigma$, then the fraction $\mathbf{E}\widetilde{R}/\mathbf{E}\widetilde{C}$ is the same for all $\mathcal{S} \in \Sigma$ as well. Therefore, it follows from equations (2.6) and (2.7) that the expected duration of a busy period, $\mathbf{E}\widetilde{T} = \mathbf{E}\widetilde{C} - \lambda^{-1}$, is the same for all queueing systems $\mathcal{S} \in \Sigma$.

Recall that for queueing system $\mathcal{S}_2$ we have $\widetilde{\zeta_1}{=}\widetilde{\zeta_2} = \ldots$, i.e., the random variable ζ is modeled once at the initial time moment. Let $\widetilde{T}_\zeta$ denote a busy period of this system. Then, the total expectation formula enables us to write

$$\mathbf{E}\widetilde{T}_\zeta = \sum_{K=\zeta^{lower}}^{\zeta^{upper}} \mathbf{E}T_K \mathbf{P}\{\zeta = K\}, \tag{2.8}$$

where $\mathbf{E}T_K$ is the expected busy period of an $M/GI/1/K$ queueing system with the same sequence of interarrival and service times, and $\mathbf{P}\{\zeta = K\} = \mathbf{P}\{\zeta_j = K\}$. In turn, the expectation $\mathbf{E}T_K$ is determined from the following recurrence relation:

$$\mathbf{E}T_K = \sum_{j=0}^{K} \pi_j \mathbf{E}T_{K-j+1}, \quad \mathbf{E}T_0 = b, \quad \pi_i = \int_0^\infty \mathrm{e}^{-\lambda x} \frac{(\lambda x)^i}{i!} \mathrm{d}B(x) \tag{2.9}$$

(see Tomkó [18], Cooper and Tilt [12], and Abramov [1], [3]), where b is the expectation of a service time.

Now, let T_ζ denote a busy period for the queueing system $\mathcal{S}_1$. According to the above conclusion that $\mathbf{E}T_\zeta = \mathbf{E}\widetilde{T}_\zeta$, and in view of (2.8), we have

$$\mathbf{E}T_\zeta = \sum_{K=\zeta^{lower}}^{\zeta^{upper}} \mathbf{E}T_K \mathbf{P}\{\zeta = K\}, \tag{2.10}$$

where $\mathbf{E}T_K$ are determined from (2.9).

Along with the notation T_ζ for the busy period of the system $\mathcal{S}_1$, we let I_ζ be an idle period and let P_ζ, M_ζ, R_ζ be the characteristics of the system on a busy period: the number of processed messages, the number of marked messages, the number of refused messages, respectively. Here and later we use the following terminology. The term *refused* message is used for the case of overflowing the buffer. Then the term *lost* message is used for the case where a message is either refused or marked. The number of lost messages during a busy period is denoted by L_ζ. Analogously, by loss probability we mean the probability when an arrival message is lost.

LEMMA 2.1. *For the expectations* $\mathbf{E}T_\zeta$, $\mathbf{E}P_\zeta$, $\mathbf{E}M_\zeta$, $\mathbf{E}R_\zeta$ *we have the following representations:*

$$\mathbf{E}P_\zeta = \frac{\lambda}{\varrho} \mathbf{E}T_\zeta, \tag{2.11}$$

$$\mathbf{E}M_\zeta = p\mathbf{E}P_\zeta, \tag{2.12}$$

$$\mathbf{E}R_\zeta = (\varrho - 1)\mathbf{E}P_\zeta + 1. \tag{2.13}$$

Proof. Relations (2.11) and (2.12) follow immediately from Wald's identity. In order to prove (2.13) note that the number of arrivals during a busy cycle equals the number of processed messages during a busy period plus the number of refused messages during a busy period (see relation (2.6)). According to Wald's identity the expected number of arrivals during a busy cycle equals $\lambda(\mathbf{E}T_\zeta + \mathbf{E}I_\zeta)$. Therefore taking into account that $\mathbf{E}I_\zeta = \lambda^{-1}$ from (2.11), we have

$$\mathbf{E}R_\zeta = (\varrho - 1)\mathbf{E}P_\zeta + 1,$$

and the result is proved. □

For the alternative proof of (2.13) see Abramov [3]. (See also the proof in [5].)

3. Auxiliary results. Tauberian theorems with remainder. It is seen from relations (2.10) and (2.9) and Lemma 2.1 that the characteristics of the system during a busy period can be studied in a framework of the recurrence relation

$$Q_k = \sum_{i=0}^{k} r_i Q_{k-i+1}, \tag{3.1}$$

where r_i are nonnegative numbers, $r_0 + r_1 + \cdots = 1$, $r_0 > 0$, and $Q_0 \neq 0$ is an arbitrary real number. Below we recall a number of results on asymptotic behavior of that sequence (3.1).

The known results on representation (3.1) are asymptotic theorems by Takács [17]. Lemma 3.1 below joins two results by Takács: Theorem 5 of [17, p. 22] and relation (35) [17, p. 23]. The results of Takács [17] were then developed by Postnikov [14, sect. 25], [15, sect. 25] (see Lemma 3.2 and Lemma 3.3 below).

Let $r(z) = \sum_{i=0}^{\infty} r_i z^i$, $|z| \leq 1$, $\gamma_m = r^{(m)}(1-0) = \lim_{z\uparrow 1} r^{(m)}(z)$ ($r^{(m)}(z)$ is the mth derivative of $r(z)$). Note that if we denote $Q(z) = \sum_{i=0}^{\infty} Q_i z^i$, then it follows from (3.1) that

$$Q(z) = \frac{Q_0 r(z)}{r(z) - z}.$$

LEMMA 3.1 (Takács [17]). *If $\gamma_1 < 1$, then*

$$\lim_{k\to\infty} Q_k = \frac{Q_0}{1 - \gamma_1}. \tag{3.2}$$

If $\gamma_1 = 1$ and $\gamma_2 < \infty$, then

$$\lim_{k\to\infty} \frac{Q_k}{k} = \frac{2Q_0}{\gamma_2}. \tag{3.3}$$

If $\gamma_1 > 1$, then

$$\lim_{k\to\infty} \Big(Q_k - \frac{Q_0}{\delta^k [1 - r'(\delta)]} \Big) = \frac{Q_0}{1 - \gamma_1}, \tag{3.4}$$

where δ is the least (absolute) root of the equation $z = r(z)$.

LEMMA 3.2 (Postnikov [14], [15]). *Let $\gamma_1 = 1$, $\gamma_3 < \infty$. Then as $k \to \infty$,*

$$Q_k = \frac{2Q_0}{\gamma_2} k + O(\log k). \tag{3.5}$$

LEMMA 3.3 (Postnikov [14], [15]). *Let $\gamma_1 = 1$, $\gamma_2 < \infty$ and $r_0 + r_1 < 1$. Then as $k \to \infty$,*

$$Q_{k+1} - Q_k = \frac{2Q_0}{\gamma_2} + o(1). \tag{3.6}$$

4. Asymptotic results for characteristics of the system during a busy period. This section provides a number of results on asymptotic behavior of characteristics of the system. The first three theorems are related to the case as N increases to infinity, where the cases $\varrho < 1$, $\varrho = 1$, and $\varrho > 1$ are considered. The next two theorems discuss the case when the value ϱ is close to the critical value 1, and as $N \to \infty$, it tends to 1. The last theorem of this section, Theorem 4.6, provides the asymptotic result for the special case when the number of packets in a message is a constant value.

Let us now study the asymptotic behavior of the expectations $\mathbf{E}P_\zeta$, $\mathbf{E}M_\zeta$, and $\mathbf{E}R_\zeta$. We write $\zeta = \zeta(N)$, pointing out the dependence on parameter N. As the buffer size N increases to infinity, both ζ^{lower} and ζ^{upper} tend to infinity, and together with them, $\zeta(N)$ a.s. tends to infinity. Then we have the following.

THEOREM 4.1. *If $\varrho < 1$, then*

$$\lim_{N\to\infty} \mathbf{E}P_{\zeta(N)} = \frac{1}{1-\varrho}. \tag{4.1}$$

If $\varrho = 1$ and $\varrho_2 < \infty$, then

$$\lim_{N\to\infty} \frac{\mathbf{E}P_{\zeta(N)}}{\mathbf{E}\zeta(N)} = \frac{2}{\varrho_2}. \tag{4.2}$$

If $\varrho > 1$, then

$$\lim_{N\to\infty} \Big[\mathbf{E}P_{\zeta(N)} - \frac{1}{\mathbf{E}\varphi^{\zeta(N)}[1 + \lambda\beta'(\lambda - \lambda\varphi)]}\Big] = \frac{1}{1-\varrho}, \tag{4.3}$$

where $\beta(z) = \int_0^\infty e^{-zx} dB(x)$ and φ is the least (absolute) root of functional equation $z - \beta(\lambda - \lambda z) = 0$.

Proof. From (2.9), (2.10), and (2.11) we have

$$\mathbf{E}P_\zeta = \sum_{K=\zeta^{lower}}^{\zeta^{upper}} \mathbf{E}P_K \mathbf{P}\{\zeta = K\},$$

where

$$\mathbf{E}P_K = \sum_{j=0}^{K} \pi_j \mathbf{E}P_{K-j+1}, \quad \mathbf{E}P_0 = 1,$$

$$\pi_j = \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^j}{j!} dB(x).$$

Then applying Lemma 3.1 we have the following. In the case $\varrho < 1$, taking into account that $\zeta(N) \overset{a.s.}{\to} \infty$ as $N \to \infty$, we obtain

$$\lim_{N\to\infty} \mathbf{E}P_{\zeta(N)} = \lim_{N\to\infty} \mathbf{E}P_N = \frac{1}{1-\varrho}.$$

Relation (4.1) is proved.

In the case $\varrho_2 < \infty$ and $\varrho = 1$ we have

$$\lim_{N\to\infty} \frac{\mathbf{E}P_{\zeta(N)}}{N} = \lim_{N\to\infty} \frac{1}{N} \sum_{K=\zeta^{lower}}^{\zeta^{upper}} \mathbf{P}\{\zeta(N) = K\}\mathbf{E}P_K$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{K=\zeta^{lower}}^{\zeta^{upper}} K\mathbf{P}\{\zeta(N) = K\}\frac{2}{\varrho_2} = \frac{2}{\varrho_2} \lim_{N\to\infty} \frac{\mathbf{E}\zeta(N)}{N}.$$

Therefore,

$$\lim_{N\to\infty} \frac{\mathbf{E}P_{\zeta(N)}}{\mathbf{E}\zeta(N)} = \frac{2}{\varrho_2},$$

and relation (4.2) is proved.

In the case $\varrho > 1$ for large N we obtain

$$\mathbf{E}P_{\zeta(N)} = \sum_{K=\zeta^{lower}}^{\zeta^{upper}} \mathbf{P}\{\zeta(N) = K\}\mathbf{E}P_K$$

$$= \sum_{K=\zeta^{lower}}^{\zeta^{upper}} \mathbf{P}\{\zeta(N) = K\}\frac{1}{\varphi^K[1 + \lambda\beta'(\lambda - \lambda\varphi)]} + \frac{1}{1-\varrho} + o(1)$$

$$= \frac{1}{\mathbf{E}\varphi^{\zeta(N)}[1 + \lambda\beta'(\lambda - \lambda\varphi)]} + \frac{1}{1-\varrho} + o(1).$$

Therefore,

$$\lim_{N\to\infty} \Big[\mathbf{E}P_{\zeta(N)} - \frac{1}{\mathbf{E}\varphi^{\zeta(N)}[1 + \lambda\beta'(\lambda - \lambda\varphi)]}\Big] = \frac{1}{1-\varrho},$$

and relation (4.3) is proved. Theorem 4.1 is completely proved. □

THEOREM 4.2. *If $\varrho = 1$ and $\varrho_3 < \infty$, then*

$$\mathbf{E}P_{\zeta(N)} = \frac{2}{\varrho_2}\mathbf{E}\zeta(N) + O(\log N). \tag{4.4}$$

Proof. Applying Lemma 3.2, for large N we have

$$\mathbf{E}P_{\zeta(N)} = \sum_{K=\zeta^{lower}}^{\zeta^{upper}} \mathbf{P}\{\zeta(N) = K\}\mathbf{E}P_K$$

$$= \sum_{K=\zeta^{lower}}^{\zeta^{upper}} K\mathbf{P}\{\zeta(N) = K\}\frac{2}{\varrho_2} + O\{\mathbf{E}[\log\zeta(N)]\}$$

$$= \frac{2}{\varrho_2}\mathbf{E}\zeta(N) + O\{\mathbf{E}[\log \zeta(N)]\}$$

$$= \frac{2}{\varrho_2}\mathbf{E}\zeta(N) + O(\log N).$$

and we obtain relation (4.4). Theorem 4.2 is proved. □

In turn for $\mathbf{E}R_\zeta$ we have the following theorem.

THEOREM 4.3. *If $\varrho < 1$, then*

$$\lim_{N\to\infty} \mathbf{E}R_{\zeta(N)} = 0. \tag{4.5}$$

If $\varrho = 1$, then for all $N \geq 0$

$$\mathbf{E}R_{\zeta(N)} = 1. \tag{4.6}$$

If $\varrho > 1$, then

$$\lim_{N\to\infty} \Big[\mathbf{E}R_{\zeta(N)} - \frac{\varrho - 1}{\mathbf{E}\varphi^{\zeta(N)}[1 + \lambda\beta'(\lambda - \lambda\varphi)]}\Big] = 0. \tag{4.7}$$

Proof. The proof of this theorem is analogous to that of the proof of Theorem 4.1. It follows by application of Lemma 3.1 and relation (2.13) of Lemma 2.1. □

THEOREM 4.4. *Let $\varrho = 1 + \varepsilon$, $\varepsilon > 0$, and $\varepsilon\zeta(N) \to C > 0$ a.s. as $\varepsilon \to 0$ and $N \to \infty$. Assume also that $\varrho_3 = \varrho_3(N)$ is a bounded sequence, and there exists $\tilde{\varrho}_2 = \lim_{N\to\infty} \varrho_2(N)$. Then*

$$\mathbf{E}P_{\zeta(N)} = \frac{e^{2C/\tilde{\varrho}_2} - 1}{\varepsilon} + O(1), \tag{4.8}$$

$$\mathbf{E}R_{\zeta(N)} = e^{2C/\tilde{\varrho}_2} + o(1). \tag{4.9}$$

Proof. It was shown in Subhankulov [16, p. 326], that if $\varrho = 1 + \varepsilon$, $\varepsilon > 0$, $\varepsilon \to 0$, $\varrho_3(N)$ is a bounded sequence, and there exists $\tilde{\varrho}_2 = \lim_{N\to\infty} \varrho_2(N)$, then

$$\varphi = 1 - \frac{2\varepsilon}{\tilde{\varrho}_2} + O(\varepsilon^2). \tag{4.10}$$

Applying (4.10) after some algebra we have

$$1 + \lambda\beta'(\lambda - \lambda\varphi) = \varepsilon + O(\varepsilon^2). \tag{4.11}$$

Then the statements of the theorem follow by applying expansions (4.10) and (4.11) to (4.3) and (4.7). □

THEOREM 4.5. *Let $\varrho = 1 + \varepsilon$, $\varepsilon > 0$, and $\varepsilon\zeta(N) \to 0$ as $\varepsilon \to 0$ and $N \to \infty$. Assume also that $\varrho_3 = \varrho_3(N)$ is a bounded sequence, and there exists $\tilde{\varrho}_2 = \lim_{N\to\infty} \varrho_2(N)$. Then*

$$\mathbf{E}P_{\zeta(N)} = \frac{2}{\tilde{\varrho}_2}\mathbf{E}\zeta(N) + O(1), \tag{4.12}$$

$$\mathbf{E}R_{\zeta(N)} = 1 + o(1). \tag{4.13}$$

Proof. The results follow by expanding (4.8) and (4.9) for small C. □

Special case. If each message contains the same number of packets, say l, then we have the usual $M/GI/1/n$ queueing system, where $n = \lfloor N/l \rfloor$ is the integer part of N/l. For that queueing system all the results in Theorems 4.1–4.5 hold, by replacing $\zeta(N)$ (or $\mathbf{E}\zeta(N)$) by n.

For example, asymptotic relation (4.7) appears as

$$\lim_{n\to\infty} \Big(\mathbf{E}R_n - \frac{\varrho - 1}{\varphi^n[1 + \lambda\beta'(\lambda - \lambda\varphi)]} \Big) = 0. \tag{4.14}$$

Notice that using the final value theorem for z transform, Azlarov and Tahirov [8] obtain the estimation

$$\mathbf{E}R_n = \frac{\varrho - 1}{\varphi^n[1 + \lambda\beta'(\lambda - \lambda\varphi)]} \Big[1 + O\Big(\frac{2\varphi}{1+\varphi} \Big)^n \Big], \tag{4.15}$$

weaker than (4.14).

The theorem below is related to the case of the usual queueing systems only, when the number of packets in a message is fixed. Namely, we have the following.

THEOREM 4.6. *If $\varrho = 1$ and $\varrho_2 < \infty$, then*

$$\mathbf{E}P_{n+1} - \mathbf{E}P_n = \frac{2}{\varrho_2} + o(1), \quad n \to \infty, \tag{4.16}$$

where the index $n+1$ says that P_{n+1} is the number of processed messages during a busy period of the $M/GI/1/n+1$ queueing system.

Proof. The result will follow from Lemma 3.3 if we show that $\beta(\lambda) - \lambda\beta'(\lambda) < 1$. Taking into account that for each $\lambda > 0$,

$$\begin{aligned} \sum_{i=0}^{\infty} \frac{(-\lambda)^i}{i!} \beta^{(i)}(\lambda) &= \sum_{i=0}^{\infty} \int_0^{\infty} \mathrm{e}^{-\lambda x} \frac{(\lambda x)^i}{i!} \mathrm{d}B(x) \\ &= \int_0^{\infty} \sum_{i=0}^{\infty} \mathrm{e}^{-\lambda x} \frac{(\lambda x)^i}{i!} \mathrm{d}B(x) = 1, \end{aligned} \tag{4.17}$$

and all terms

$$\pi_i = \frac{(-\lambda)^i}{i!} \beta^{(i)}(\lambda)$$

are nonnegative, from (4.17) we find that

$$\beta(\lambda) - \lambda\beta'(\lambda) \le 1. \tag{4.18}$$

Thus, the required statement will be proved if we show that for some $\lambda_0 > 0$ the equality

$$\beta(\lambda_0) - \lambda_0\beta'(\lambda_0) = 1 \tag{4.19}$$

is not a case. Indeed, since the function $\beta(\lambda) - \lambda\beta'(\lambda)$ is an analytic function, then according to the maximum absolute value principle for analytic functions, $\beta(\lambda) - \lambda\beta'(\lambda) = 1$ holds for all $\lambda > 0$. Therefore identity (4.19) means that $\pi_i = 0$ for all $i \ge 2$ and for all $\lambda > 0$. Therefore, (4.19) is valid if and only if $\beta(\lambda)$ is a linear function, i.e., $\beta(\lambda) = c_0 + c_1\lambda$, c_0 and c_1 are some constants. However, since $|\beta(\lambda)| \le 1$ we obtain $c_0 = 1$ and $c_1 = 0$, and $\beta(\lambda) \equiv 1$. This is the trivial case where the probability distribution function $B(x)$ is concentrated in point 0. Therefore (4.19) is not a case, and $\beta(\lambda) - \lambda\beta'(\lambda) < 1$. The theorem is proved. □

5. Asymptotic theorems for the loss probabilities. In this section we study the asymptotic behavior of the loss probability by using renewal arguments. The results of this section correspond to those of the previous section. We discuss the behavior of the system for the same cases as $N \to \infty$, as well as when the parameter ρ is close to the critical value 1 and tends to 1 as $N \to \infty$. The theorems of this section are important for our conclusion on adding redundancy, which is given in the next section.

According to renewal arguments the loss probability is determined as

$$\Pi_\zeta = \frac{\mathbf{E}L_\zeta}{\mathbf{E}R_\zeta + \mathbf{E}P_\zeta} = \frac{\mathbf{E}R_\zeta + \mathbf{E}M_\zeta}{\mathbf{E}R_\zeta + \mathbf{E}P_\zeta} = \frac{\mathbf{E}R_\zeta + p\mathbf{E}P_\zeta}{\mathbf{E}R_\zeta + \mathbf{E}P_\zeta}. \tag{5.1}$$

(Recall that L_ζ is the number of lost messages during a busy period.)

THEOREM 5.1. *If* $\varrho < 1$

$$\lim_{N\to\infty} \Pi_{\zeta(N)} = p. \tag{5.2}$$

(Recall that p *is the probability that a message is erroneous because one of its packets is corrupted.)*

Limiting relation 5.2 *is also valid when* $\varrho = 1$ *and* $\varrho_2 < \infty$.

If $\varrho > 1$, *then*

$$\Pi_{\zeta(N)} = \frac{p+\varrho-1}{\varrho}\,\frac{(\varrho-1)+p[1+\lambda\beta'(\lambda-\lambda\varphi)]\mathbf{E}\varphi^{\zeta(N)}}{(\varrho-1)+[1+\lambda\beta'(\lambda-\lambda\varphi)]\mathbf{E}\varphi^{\zeta(N)}} + o(\mathbf{E}\varphi^{\zeta(N)}). \tag{5.3}$$

Proof. The proof follows from Theorems 4.1 and 4.3. □

THEOREM 5.2. *If* $\varrho = 1$ *and* $\varrho_3 < \infty$, *then as* $N \to \infty$

$$\Pi_{\zeta(N)} = p + \frac{(1-p)\varrho_2}{2\mathbf{E}\zeta(N)} + O\Big(\frac{\log N}{N^2}\Big). \tag{5.4}$$

Proof. From (5.1) we have

$$\Pi_{\zeta(N)} = \frac{\mathbf{E}R_{\zeta(N)}}{\mathbf{E}R_{\zeta(N)} + \mathbf{E}P_{\zeta(N)}} + \frac{p\mathbf{E}P_{\zeta(N)}}{\mathbf{E}R_{\zeta(N)} + \mathbf{E}P_{\zeta(N)}} \tag{5.5}$$

$$= \frac{1}{1+\mathbf{E}P_{\zeta(N)}} + \frac{p\mathbf{E}P_{\zeta(N)}}{1+\mathbf{E}P_{\zeta(N)}}.$$

As $N \to \infty$ from Theorem 4.2 we obtain

$$\frac{1}{1+\mathbf{E}P_{\zeta(N)}} = \frac{\varrho_2}{2\mathbf{E}\zeta(N)} + O\Big(\frac{\log N}{N^2}\Big), \tag{5.6}$$

$$\frac{p\mathbf{E}P_{\zeta(N)}}{1+\mathbf{E}P_{\zeta(N)}} = p - \frac{p\varrho_2}{2\mathbf{E}\zeta(N)} + O\Big(\frac{\log N}{N^2}\Big). \tag{5.7}$$

Combining these two asymptotic relations (5.6) and (5.7) we obtain the statement of Theorem 5.2. Theorem 5.2 is proved. □

Note. Under assumptions of Theorem 5.2 assume additionally that $p \to 0$. If $pN \to C > 0$, then

$$\Pi_{\zeta(N)} = \frac{C}{N} + \frac{\varrho_2}{2\mathbf{E}\zeta(N)} + O\Big(\frac{\log N}{N^2}\Big).$$

If $pN \to 0$, then

$$\Pi_{\zeta(N)} = \frac{\varrho_2}{2\mathbf{E}\zeta(N)} + O\Big(p + \frac{\log N}{N^2}\Big).$$

The theorem below also assumes that $p \to 0$. Our result here is the following.

THEOREM 5.3. *Let $\varrho = 1 + \varepsilon$, $\varepsilon > 0$, and $\varepsilon\zeta(N) \to C > 0$ as $\varepsilon \to 0$ and $N \to \infty$, and $p \to 0$. Assume also that $\varrho_3 = \varrho_3(N)$ is a bounded sequence, and there exists $\tilde{\varrho}_2 = \lim_{n\to\infty} \varrho_2(N)$.*

(i) *If $p/\varepsilon \to D \geq 0$, then we have*

$$\Pi_{\zeta(N)} = \left(D + \frac{e^{2C/\widetilde{\varrho}_2}}{e^{2C/\widetilde{\varrho}_2} - 1}\right)\varepsilon + o(\varepsilon). \tag{5.8}$$

(ii) *If $p/\varepsilon \to \infty$, then we have*

$$\Pi_{\zeta(N)} = p + O(\varepsilon). \tag{5.9}$$

Proof. In the case (i) we have

$$p\mathbf{E}P_{\zeta(N)} + \mathbf{E}R_{\zeta(N)} = (D+1)\mathrm{e}^{2C/\widetilde{\varrho}_2} - D + o(1), \tag{5.10}$$

and

$$\mathbf{E}P_{\zeta(N)} + \mathbf{E}R_{\zeta(N)} = \frac{\mathrm{e}^{2C/\widetilde{\varrho}_2} - 1}{\varepsilon} + O(1). \tag{5.11}$$

Therefore from (5.10) and (5.11) we have

$$\Pi_{\zeta(N)} = \left(D + \frac{\mathrm{e}^{2C/\widetilde{\varrho}_2}}{\mathrm{e}^{2C/\widetilde{\varrho}_2} - 1}\right)\varepsilon + o(\varepsilon),$$

and relation (5.8) is proved.

In the case (ii) we have

$$p\mathbf{E}P_{\zeta(N)} + \mathbf{E}R_{\zeta(N)} = \frac{pc}{\varepsilon} + O(1), \tag{5.12}$$

and

$$\mathbf{E}P_{\zeta(N)} + \mathbf{E}R_{\zeta(N)} = \frac{c}{\varepsilon} + O(1), \tag{5.13}$$

where $c = \exp(2C/\widetilde{\varrho}_2)/(\exp(2C/\widetilde{\varrho}_2) - 1)$. Relation (5.9) follows. □

THEOREM 5.4. *Let $\varrho = 1 + \varepsilon$, $\varepsilon > 0$, and $\varepsilon\zeta(N) \to 0$ as $\varepsilon \to 0$ and $N \to \infty$, and $p \to 0$. Assume also that $\varrho_3 = \varrho_3(N)$ is a bounded sequence, and there exists $\tilde{\varrho}_2 = \lim_{n\to\infty} \varrho_2(N)$.*

(i) *If* $p/\varepsilon \to D \geq 0$, *then we have*

$$\Pi_{\zeta(N)} = p + \frac{\widetilde{\varrho}_2}{2\mathbf{E}\zeta(N)} + o\Big(\frac{1}{N}\Big). \tag{5.14}$$

(ii) *If* $p/\varepsilon \to \infty$, *then we have* (5.9).

Proof. The proof of (5.14) follows by expanding (5.8) for small C. The proof in case (ii) trivially follows from (5.12) and (5.13). □

Special case. In the case where each message contains exactly l packets, $n = \lfloor N/l \rfloor$, we obtain the following:

THEOREM 5.5. *If* $\varrho = 1$ *and* $\varrho_2 < \infty$, *then as* $n \to \infty$

$$\Pi_{n+1} - \Pi_n = \frac{\frac{1}{n(n+1)}\ \frac{2}{\varrho_2}(p-1)}{\Big(\frac{2}{\varrho_2} + \frac{1}{n+1}\Big)\Big(\frac{2}{\varrho_2} + \frac{1}{n}\Big)} + o\Big(\frac{1}{n^2}\Big). \tag{5.15}$$

Proof. The proof follows by applying Theorem 4.5 and taking into account the fact that $\mathbf{E}R_n = 1$ for all $n \geq 0$ (see [3] or Lemma 2.1). □

6. Adding redundant packets. We now investigate the effect of adding redundant packets. We assume that adding a redundant packet to the message decreases the probability p that a message is corrupted and increases the offered load and the number of packets in a message. The new parameters of the system after adding a redundant packet are denoted by adding the symbol $\breve{}$ above. For example, $\breve{p}$ is a probability that a message contains a corrupted packet and $\breve{\varrho}$ is the offered load. It follows from Theorem 5.1 that if $\breve{\varrho} \leq 1$ the stationary loss probability coincides with $\breve{p}$. This means that if adding a redundant packet to the message decreases the probability p by γ times, then the same effect is achieved with the loss probability. Thus, adding a number of redundant packets while $\varrho < 1$ can decrease the loss probability geometrically.

In the case where both $\varrho > 1$ and $\breve{\varrho} > 1$, adding a redundant packet to the message changes the stationary loss probability to approximately

$$\frac{\varrho(\breve{p} + \breve{\varrho} - 1)}{\breve{\varrho}(p + \varrho - 1)}.$$

In practice the values p and $\breve{p}$ are small, and even if adding redundant packets can slightly decrease the stationary loss probability, the effect of that action is not considerable.

The case where $\varrho < 1$ and $\breve{\varrho} > 1$ is especially interesting if $\breve{\varrho} = 1 + \delta$, and δ is a small value. For example, if δ is so small that both $\delta\zeta(N)$ and δ/p are also negligible, then a redundant packet decreases the loss probability by approximately the same amount as in the case when both $\varrho < 1$ and $\breve{\varrho} < 1$. However, if δ is of the same order as p or $1/\zeta(N)$, then the special analysis based on the corresponding cases of Theorems 5.3 and 5.4 is necessary. Here we do not provide the details.

Let us consider the cases when both $\varrho > 1$ and $\breve{\varrho} > 1$, where $\varrho = 1 + \varepsilon$ and $\breve{\varrho} = 1 + \breve{\varepsilon}$, and ε and $\breve{\varepsilon}$ are small values as in Theorem 5.3, both satisfying (i). Then the stationary loss probability is changed to approximately

$$\frac{\mathrm{e}^{2C/\widetilde{\varrho}_2} - 1}{\mathrm{e}^{2\breve{C}/\breve{\varrho}_2} - 1}\ \frac{(\mathrm{e}^{2\breve{C}/\breve{\varrho}_2} - 1)\breve{p} + \mathrm{e}^{2\breve{C}/\breve{\varrho}_2}\breve{\varepsilon}}{(\mathrm{e}^{2C/\widetilde{\varrho}_2} - 1)p + \mathrm{e}^{2C/\widetilde{\varrho}_2}\varepsilon} \tag{6.1}$$

times.

For the sake of simplicity let us assume that $\breve{C}/\breve{\varrho}_2 = C/\widetilde{\varrho}_2$. Then (6.1) reduces to

$$\frac{(\mathrm{e}^{2C/\widetilde{\varrho}_2}-1)\breve{p}+\mathrm{e}^{2C/\widetilde{\varrho}_2}\breve{\varepsilon}}{(\mathrm{e}^{2C/\widetilde{\varrho}_2}-1)p+\mathrm{e}^{2C/\widetilde{\varrho}_2}\varepsilon}. \tag{6.2}$$

If we assume that

$$p-\breve{p}=\frac{\mathrm{e}^{2C/\widetilde{\varrho}_2}}{\mathrm{e}^{2C/\widetilde{\varrho}_2}-1}(\breve{\varepsilon}-\varepsilon),$$

then the stationary loss probability remains at approximately the same value, and if

$$p-\breve{p}>\frac{\mathrm{e}^{2C/\widetilde{\varrho}_2}}{\mathrm{e}^{2C/\widetilde{\varrho}_2}-1}(\breve{\varepsilon}-\varepsilon),$$

then the stationary loss probability decreases, otherwise if

$$p-\breve{p}<\frac{\mathrm{e}^{2C/\widetilde{\varrho}_2}}{\mathrm{e}^{2C/\widetilde{\varrho}_2}-1}(\breve{\varepsilon}-\varepsilon),$$

then the stationary loss probability increases.

Acknowledgments. The author thanks Professor Moshe Sidi (Technion) for sending him the files of related papers. The author also thanks the anonymous referees and associate editor for a number of valuable comments.

REFERENCES

[1] V. M. Abramov, *Investigation of a Queueing System with Service Depending on Queue Length*, Donish, Dushanbe, Tadzhikistan, 1991 (in Russian).
[2] V. M. Abramov, *Asymptotic theorems for one queueing system with refusals*, Kibernetika (Ukrainian Academy of Sciences), 2 (1991), pp. 123–124 (in Russian).
[3] V. M. Abramov, *On a property of a refusals stream*, J. Appl. Probab., 34 (1997), pp. 800–805.
[4] V. M. Abramov, *Asymptotic analysis of the $GI/M/1/n$ loss system as n increases to infinity*, Ann. Oper. Res., 112 (2002), pp. 35–41.
[5] V. M. Abramov, *On losses in $M^X/GI/1/n$ queues*, J. Appl. Probab., 38 (2001), pp. 1079–1080.
[6] O. Ait-Hellal, E. Altman, A. Jean-Marie, and I. A. Kurkova, *On loss probabilities in presence of redundant packets and several traffic sources*, Perform. Eval., 36–37 (1999), pp. 485–518.
[7] E. Altman and A. Jean-Marie, *Loss probabilities for messages with redundant packets feeding a finite buffer*, IEEE J. Select. Areas Commun., 16 (1998), pp. 778–787.
[8] T. A. Azlarov and A. Tahirov, *Limit theorems for single-server queueing system with finite number of waiting places*, Proc. USSR Acad. Sci. Engrg. Cybern., 5 (1974), pp. 53–57.
[9] B. D. Choi and B. Kim, *Sharp results on convergence rates for the distribution of $GI/M/1/K$ queues as K tends to infinity*, J. Appl. Probab., 37 (2000), pp. 1010–1019.
[10] B. D. Choi, B. Kim, and I.-S. Wee, *Asymptotic behavior of loss probability in the $GI/M/1/K$ queue as K tends to infinity*, Queueing Syst. Theory Appl., 36 (2000), pp. 437–442.
[11] I. Cidon, A. Khamisy, and M. Sidi, *Analysis of packet loss processes in high-speed networks*, IEEE Trans. Inform. Theory, 39 (1993), pp. 98–108.
[12] R. B. Cooper and B. Tilt, *On the relationship between the distribution of the maximum queue-length in the $M/G/1$ queue and the mean busy period in the $M/G/1/n$ queue*, J. Appl. Probab., 13 (1976), pp. 195–199.
[13] O. Gurewitz, M. Sidi, and I. Cidon, *The ballot theorem strikes again: Packet loss process distribution*, IEEE Trans. Inform. Theory, 46 (2000), pp. 2588–2595.
[14] A. G. Postnikov, *Tauberian Theory and Its Applications*, Trudy Mat. Inst. Steklov, 144 (1979), pp. 1–147 (in Russian).

[15] A. G. Postnikov, *Tauberian Theory and Its Applications*, Proc. Steklov Inst. Math., 1980, pp. 1–138.
[16] M. A. Subhankulov, *Tauberian Theorems with Remainder*, Nauka, Moscow, 1976 (in Russian).
[17] L. Takács, *Combinatorial Methods in the Theory of Stochastic Processes*, John Wiley, New York, 1967.
[18] J. Tomkó, *One limit theorem in the queueing problem as input rate increases infinitely*, Studia Sci. Math. Hungar., 2 (1967), pp. 447–454 (in Russian).

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 762–774

OPTIMAL LOCALIZATION OF EIGENFUNCTIONS IN AN INHOMOGENEOUS MEDIUM*

DAVID C. DOBSON† AND FADIL SANTOSA‡

Abstract. The problem of creating eigenfunctions which are localized arises in the study of photonic bandgap structures. A model problem, that of finding material inhomogeneity in a domain so that one of its Dirichlet eigenfunctions is localized, is considered in this work. The most difficult aspect, that of formulating the problem, is described, and well-posed variational problems are given. A computational approach, based on gradient descent with projection and trajectory continuation, is devised to solve the optimization problem. Numerical examples are provided which demonstrate the capability of the computational method.

Key words. mode localization, optimal design, defect modes

AMS subject classifications. 65K10, 82D25, 49M07

DOI. 10.1137/S0036139903426162

1. Introduction.

We study a problem arising in the design of optical devices that exploit the photonic bandgap phenomenon. This phenomenon occurs in the optical wavelength regime in certain nanostructures with periodic index of refraction. Such materials, called photonic bandgap (PBG) structures by John [8] and Yablonovitch [12], are conceived to be optical analogues of electronic semiconductors. By introducing patterned defects into a photonic bandgap structure, it is possible to control the propagation of light within the structure. Photonic bandgap structures are anticipated to play important roles in future generations of optical devices.

Photonic bandgap phenomenon refers to the existence of a certain frequency band in which waves having frequency in that band cannot propagate in the medium. A bandgap can be created in a medium with periodic structure [7]. If a bandgap exists, it is possible to create a standing wave with frequency in the gap by introducing a so-called "point defect." In its simplest form, a point defect is a localized perturbation to the underlying periodic index of refraction. It is known that such a standing wave will be spatially localized [4]. Optimizing the properties of localized modes is the subject of this work.

To be specific, let us consider the transverse electric-mode for electromagnetic waves in two dimensions. The medium is characterized by the real dielectric property $\epsilon_{\rm p}(x)$, which is unit periodic. It is assumed that the medium $\epsilon_{\rm p}(x)$ has a bandgap. That is, the spectral problem

$$\Delta u + \omega^2 \epsilon_{\rm p}(x) u = 0, \quad x \in \mathbb{R}^2,$$

has a continuous spectrum with a gap.

A defect is modeled by a localized perturbation $\eta(x)$ with compact support; thus the governing equation now takes the form

$$\Delta u + \omega^2 (\epsilon_{\rm p}(x) + \eta(x)) u = 0. \tag{1.1}$$

*Received by the editors April 10, 2003; accepted for publication (in revised form) July 3, 2003; published electronically March 11, 2004.
http://www.siam.org/journals/siap/64-3/42616.html

†Department of Mathematics, University of Utah, Salt Lake City, UT 84112 (dobson@math.utah.edu). The research of this author was supported by National Science Foundation grant DMS-0072439.

‡School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (santosa@math.umn.edu). The research of this author was supported in part by the National Science Foundation.

The stability theorem (see, e.g., [10]) states that the essential spectrum of the perturbed problem is equal to that of the periodic problem. One can interpret the result as saying that the perturbation can only create an additional discrete spectrum.

Defect modes are solution pairs $(\omega, u(x))$ to (1.1) which have the property that ω is in the frequency gap of the periodic medium, and $u(x)$ decays exponentially away from the defect. The existence of these defect modes has been addressed in the work of Figotin and Klein [6] (see also a review of mathematical results on photonic bandgap structures in [10, 5]). It is known, for example, that if the perturbation $\eta(x)$ is sufficiently strong and the frequency bandgap is sufficiently large, then a defect mode can be created.

In this work, we address a more practical question, that of finding $\eta(x)$ so that the defect mode produced has desired attributes. For example, we may wish to create a defect mode that concentrates energy in the smallest spatial region, which is useful in applications for enhancing nonlinear optical effects, or we may wish to create a defect mode with a specific frequency.

We note that one possible way of creating defect modes that are highly localized is to start with an underlying periodic medium that has a large gap. The theoretical results of Figotin and Klein (see, e.g., [6]) state that localization length is minimized by creating a defect mode whose frequency is as far away as possible from the band edge. Maximization of bandgaps has been treated in the work of Cox and Dobson [2, 3]. However, this previous work does not address the issue of how to create defect modes with specified properties.

To simplify the analysis and computation even further, we pose a model problem on a bounded domain Ω. We will look at the Dirichlet eigenvalue problem in the domain and attempt to find material properties that lead to eigenfunctions that are "most localized." This model problem avoids the difficulty of dealing with the original unbounded domain and also lumps the discovery of $\epsilon(x)$ and $\eta(x)$ in a single formulation. It also avoids the issue of explicitly satisfying the conditions needed to create defect modes in a photonic bandgap structure.

We believe that this simple model problem exhibits many of the challenges posed by the original problem and feel that the more complex physics of light can be treated by the approach we propose in this work. Moreover, the eigenfunctions, when highly localized, can be interpreted in terms of unbounded photonic bandgap structures. This is because defect modes in photonic bandgap structures are highly localized and can be well approximated by functions which are zero outside of a bounded domain.

The paper is organized as follows. We first provide a statement of the optimal design problem. In section 3 we investigate conditions under which mathematically sound formulations of the design problem can be posed. We find two well-posed formulations, each of which has some practical deficiencies. In section 4, we present a numerical method for solving the optimal design problem which explicitly assumes a finite-dimensional implementation. The method does not directly solve either of the two optimization problems formulated in section 3 but could be adapted to do so and is quite efficient at finding good designs. We show numerical examples of designs created by our approach. We find the results quite startling in the sense that the process often produces what appears to be a periodic structure with a defect. The paper ends with a discussion section that outlines some open theoretical and computational issues.

2. Problem description. We assume that the dielectric coefficient $\epsilon(x)$ of the medium satisfies $0 < \epsilon_0 \leq \epsilon(x) \leq \epsilon_1 < \infty$. We are interested in modes $u \in H_0^1(\Omega)$

satisfying the Dirichlet eigenvalue problem

$$-\Delta u = \lambda \epsilon\, u \quad \text{in } \Omega, \tag{2.1a}$$

$$u = 0 \quad \text{on } \partial\Omega \tag{2.1b}$$

for some eigenvalue λ. Here Ω is a simply connected bounded domain in $\mathbb{R}^2$ with Lipschitz continuous boundary. The quantity ϵu^2 is proportional to energy density in the medium. Normalizing the eigenfunctions to have unit energy in the domain, we specify

$$\int_\Omega \epsilon u^2 = 1. \tag{2.2}$$

Associated with problem (2.1) is the following variational problem: find $u \in H_0^1(\Omega)$ such that

$$a(u,v) = \lambda b(\epsilon; u, v) \quad \text{for all } v \in H_0^1(\Omega), \tag{2.3}$$

where

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v, \quad b(\epsilon; u, v) = \int_\Omega \epsilon\, uv.$$

The symmetric bilinear form $a(u,v)$ defines an associated L^2–self-adjoint operator A through the formula $a(u,v) = \langle Au, v\rangle$. It is well known that A has a compact self-adjoint inverse $A^{-1} : L^2(\Omega) \to L^2(\Omega)$, and the problem (2.3) is equivalent to

$$v = \lambda A^{-1/2} B(\epsilon) A^{-1/2} v,$$

where $B(\epsilon)$ is defined by $b(\epsilon; u, v) = \langle B(\epsilon)u, v\rangle$.

The admissible class of dielectric coefficient is defined

$$\mathcal{A} = \{\epsilon \in L^\infty(\Omega) : 0 < \epsilon_1 \le \epsilon(x) \le \epsilon_2, \ \text{ a.e.}\}.$$

Since the operator $A^{-1/2} B A^{-1/2}$ is compact and self-adjoint, the spectral theorem implies that for each given $\epsilon \in \mathcal{A}$, problem (2.3) admits an infinite sequence of non-negative real eigenvalues

$$0 < \lambda_0(\epsilon) \le \lambda_1(\epsilon) \le \lambda_2(\epsilon) \le \cdots, \tag{2.4}$$

listed according to multiplicity, and associated eigenfunctions $u_j(\epsilon)$, $j = 1, 2, \ldots, \infty$, (assumed to be scaled so that $\int \epsilon u_j^2 = 1$). For each fixed ϵ, the set of eigenfunctions $\{u_j(\epsilon)\}$ forms an orthogonal basis for $L^2(\Omega)$, although the sequence $u_j(\epsilon)$ is not uniquely defined, due to sign ambiguities and the possibility of multiple eigenvalues.

Our goal is to find material parameters $\epsilon \in \mathcal{A}$ such that a particular corresponding eigenfunction $u(\epsilon)$ is "most localized" at a particular point. We measure the degree of localization by the moment

$$J(\epsilon, u) = \int_\Omega w \epsilon u^2,$$

where w is some fixed prescribed weight function. In the numerical experiments to follow, we will assume $0 \in \Omega$ and take $w(x) = |x|^2$.

3. Optimization problems. The purpose of this section is to investigate conditions under which a well-defined optimization problem can be formulated. In the minimization method to be described in the following section, we will specifically make use of the finite-dimensionality of the numerical implementation in order to justify the algorithm. Before doing so, it is important to understand the difficulties associated with the underlying problem before discretization.

Without some constraints, the problem of minimizing $J(\epsilon, u(\epsilon))$ over $\mathcal{A}$ is not well-posed. By considering eigenfunctions corresponding to higher and higher frequencies, a sequence ϵ_n could be constructed which drives $J(\epsilon_n)$ to zero, but for which ϵ_n does not converge in $\mathcal{A}$. For example, one could choose ϵ_n to be a sequence of truncated photonic bandgap structures with point defects, with spatial frequency tending toward infinity. To ensure existence of a solution, we must further constrain the problem. We will consider two ways of imposing constraints, leading to a "global problem" and a "local problem."

3.1. Global problem. For each fixed $\epsilon \in \mathcal{A}$, the sequence of eigenvalues $\lambda_k(\epsilon)$, $k = 0, 1, 2, \ldots$, as in (2.4), is well defined. Choose some $N < \infty$. Define

$$E_N(\epsilon) = \{u \in H_0^1(\Omega) : \ u \text{ is an eigenfunction associated with some } \lambda_j(\epsilon),\ j \le N, \text{ satisfying } \int \epsilon u^2 = 1\}.$$

Note that $E_N(\epsilon)$ is finite-dimensional for each ϵ, but its dimension may change as ϵ is varied due to multiple eigenvalues. The global problem we consider is

$$\inf_{\epsilon \in \mathcal{A}} \min_{u \in E_N(\epsilon)} J(\epsilon, u) = \int_\Omega w\epsilon u^2. \tag{3.1}$$

PROPOSITION 3.1. *Problem* (3.1) *admits a solution* $\epsilon \in \mathcal{A}$.

Proof. By Proposition 4.3.i in Cox and McLaughlin [1], each eigenvalue $\lambda_j(\epsilon)$ is weak$*$ L^∞ continuous over $\mathcal{A}$. Also $\mathcal{A}$ is weak$*$ compact, so

$$\sup_{\epsilon \in \mathcal{A}} \lambda_N(\epsilon) = C < \infty.$$

The variational problem (2.3) then immediately gives a uniform upper bound on $\|u\|_{H^1(\Omega)}$ for all $u \in E_N(\epsilon)$, independent of $\epsilon \in \mathcal{A}$. Considering then a minimizing sequence ϵ_n with some subsequence (still denoted ϵ_n) converging weak$*$ to some $\epsilon \in \mathcal{A}$, any corresponding minimizing eigenfunctions $u_n \in E_N(\epsilon_n)$ have a subsequence (again indexed by n) such that u_n converges weakly in $H_0^1(\Omega)$ (hence strongly in $L^2(\Omega)$) to some u in $H_0^1(\Omega)$. Since $u_n^2 \to u^2$ in L^1, it follows that $J(\epsilon_n, u_n) \to J(\epsilon, u)$. Finally, the fact that $u \in E_N(u)$ follows easily from (2.3). ∎

The global problem (3.1) is interesting in that it seeks an absolute minimum over all admissible designs with a prescribed upper bound on frequency. This is very close to the problem that we would like to solve computationally. It does, however, have some drawbacks. First, nothing in the problem formulation prevents solutions from occurring at a multiple eigenvalue, in which one eigenfunction is localized while others are not. Second, the problem is computationally awkward, since the objective function is not everywhere differentiable.

Next we propose another problem which removes each of these drawbacks but introduces a new objection in that it is "local."

3.2. Local problem. Choose some $\epsilon_0 \in \mathcal{A}$ which yields an associated eigenvalue $\lambda_k(\epsilon_0)$ such that

$$\min_j |\lambda_k(\epsilon_0) - \lambda_j(\epsilon_0)| \geq 2\delta > 0.$$

In other words, the kth eigenvalue is distinct and separated from the other eigenvalues by 2δ. Then define a new admissible set

$$\mathcal{A}_\delta = \{\epsilon \in \mathcal{A} : \min_j |\lambda_k(\epsilon) - \lambda_j(\epsilon)| \geq \delta\},$$

where it is assumed that the eigenvalues are ordered according to multiplicity as in (2.4). Thus $\mathcal{A}_\delta$ contains a set of material parameters for which the kth eigenvalue is always distinct and bounded away from all other eigenvalues. Note that $\mathcal{A}_\delta$ is not empty, since it at least contains ϵ_0. Because of the weak$*$ continuity of $\lambda(\epsilon)$, it follows that $\mathcal{A}_\delta$ is weak$*$ compact. Unfortunately, however, it is not necessarily convex. Nevertheless, we can formulate the "local problem"

$$\inf_{\epsilon \in \mathcal{A}_\delta} J(\epsilon) = \int_\Omega w\epsilon u_k^2, \tag{3.2}$$

where $\pm u_k$ is a basis eigenfunction which spans the one-dimensional eigenspace associated with $\lambda_k(\epsilon)$, again normalized so that $\int \epsilon u_k^2 = 1$.

PROPOSITION 3.2. *Problem* (3.2) *admits a solution* $\epsilon \in \mathcal{A}_\delta$.

Proof. The proof follows essentially the same direct method argument as Proposition 3.1. The key point is that the weak$*$ continuity of $\lambda_k(\epsilon)$ establishes both the compactness of $\mathcal{A}_\delta$ and the fact that $\sup_{\epsilon \in \mathcal{A}_\delta} \lambda_k(\epsilon) < \infty$. The latter fact yields the uniform H_0^1 upper bound on the eigenfunctions, and the argument follows that of Proposition 3.1. □

The advantage of this problem formulation is that it excludes multiple eigenvalues and the associated changes in dimension of eigenspaces as ϵ moves through the design space. The disadvantage is that due to the nonconvexity of the admissible set $\mathcal{A}_\delta$, any derivative-based minimization method would be forced to search for solutions locally near the known design ϵ_0.

4. A minimization method. In this section, we introduce a computational method which can be adapted for solving either problem (3.1) or problem (3.2). However, as pointed out in the previous section, each problem has some practical deficiencies. The method developed here is a somewhat more general algorithm which starts with an eigenfunction of an initial design medium ϵ and iteratively decreases the objective function associated with this particular eigenfunction as it updates the medium. The algorithm handles the situation when the eigenvalue associated with the eigenfunction becomes a repeated eigenvalue at some design and can be viewed as a steepest descent approach with trajectory continuation. The continuation is needed to track the same eigenfunction as the algorithm explores the admissible designs through the iterations.

From now on, we tacitly assume that the eigenproblem (2.1) has been discretized, for example, by finite differences, into a corresponding matrix eigenproblem

$$Au = \lambda S(\epsilon)u, \tag{4.1}$$

where u and ϵ now represent finite-dimensional vectors, and $S(\epsilon)$ is a diagonal matrix which multiplies the entries of u pointwise by the elements in ϵ. Most of the following still carries through when A and S are infinite-dimensional operators. Assume that the discretization makes A and S symmetric and positive definite.

4.1. Derivative calculation. In preparing to compute derivatives of the objective function, it will be helpful to reformulate the eigenproblem in such a way that ϵ does not appear explicitly in the normalization constraint (2.2). Let b be the vector with entries $b_j = 1/\sqrt{\epsilon_j}$, and let B be the diagonal matrix with b on the main diagonal. Thus $S(\epsilon)^{-1/2} = B$. It will simplify the following calculations to consider the vector b as our design variable. Set $v = B^{-1}u$. The eigenproblem (4.1) with normalization of the eigenvector can then be equivalently stated as

$$\begin{aligned} BABv - \lambda v &= 0, \\ \langle v, v\rangle &= 1. \end{aligned}$$

Note that for any eigenpair (λ, v) satisfying this problem, we have $\lambda = \langle v, BABv\rangle$. Given a design vector b, and any associated eigenvector $v(b)$, we can then consider the objective function

$$J(b) = \frac{1}{2}\langle v(b), Wv(b)\rangle, \tag{4.2}$$

where $v(b)$ solves

$$BABv - \langle v, BABv\rangle v = 0, \tag{4.3a}$$

$$\langle v, v\rangle = 1. \tag{4.3b}$$

Here, W is a symmetric matrix which represents multiplication by the weight function w on the discretized vector v.

Let δb be a small perturbation in b. The linearized response $DJ(b)(\delta b)$ in the objective $J(b)$ is

$$DJ(b)(\delta b) = \langle \delta v, Wv\rangle,$$

where δv is the linearized response in v to δb. Differentiating (4.3), we find that δv satisfies

$$\begin{aligned} &BAB\delta v - \langle v, BABv\rangle \delta v - 2\langle BABv, \delta v\rangle v \\ &\quad = -(\delta B)ABv - BA(\delta B)v + \langle v, (\delta B)ABv\rangle v + \langle v, BA(\delta B)v\rangle v, \end{aligned} \tag{4.4}$$

where $\delta B = DB(b)(\delta b)$ is simply the diagonal matrix with δb on the diagonal.

Now define an adjoint vector q as the solution to the equation

$$BABq - \langle v, BABv\rangle q - 2\langle q, v\rangle BABv = Wv. \tag{4.5}$$

Assuming that the unit eigenvector v is associated with a simple eigenvalue, the operator $BAB - \langle v, BABv\rangle I$ has a one-dimensional null space, spanned by v. Also $BABv = \lambda v$, so the third term is the rank-one projection $-2\lambda vv^T$. The sum of the three operators acting on q thus has a trivial null space, and the adjoint equation has a unique solution q.

By a straightforward calculation using (4.4) and (4.5) we then have

$$\begin{aligned} DJ(b)(\delta b) &= \langle \delta v, Wv\rangle \\ &= -\langle \delta b, \operatorname{diag}(q)ABv\rangle - \langle \delta b, \operatorname{diag}(v)ABq\rangle + 2\langle q, v\rangle\langle \delta b, \operatorname{diag}(v)ABv\rangle. \end{aligned}$$

Setting

$$g = \operatorname{diag}(q)ABv - \operatorname{diag}(v)ABq + 2\langle q, v\rangle \operatorname{diag}(v)ABv, \tag{4.6}$$

we have $DJ(b)(\delta b) = \langle \delta b, g\rangle$; hence we identify the vector g with the *gradient* of J.

4.2. Gradient descent. As mentioned above, the gradient g (4.6) is not necessarily well defined unless the eigenvector v used in the objective is associated with a simple eigenvalue.

The (negative) gradient at a given point b gives a direction to move in the design space which will result in a *particular* eigenvector becoming more localized. It is important to note that the same calculation above applies to *any* eigenvector $v(b)$, provided it is associated with a simple eigenvalue. This actually creates an algorithmic problem: suppose we implement a simple gradient descent algorithm in which steps are iteratively taken in the direction of $-g$. One would expect that in moving through the design space, eigenvalues will cross, and it may quickly become unclear which eigenvector we are supposed to be localizing. If we jump back and forth between different eigenvectors during the course of the optimization, there should be no expectation of convergence. In fact, this behavior was observed in preliminary numerical experiments.

A solution to the problem of tracking eigenvectors is provided by the following theorem.

THEOREM 4.1 (Kato, [9]). *Let $T(\tau)$ be a symmetric $N \times N$ matrix whose entries are analytic functions of τ. Then there exist N holomorphic vector-valued functions $\{v_j(\tau)\}_{j=1}^N$ which are orthonormal eigenvectors for T.*

Notice that the matrix $T(\tau) = (B - \tau \operatorname{diag}(g))A(B - \tau \operatorname{diag}(g))$, which results from a step in the direction of $-g$, satisfies the hypothesis of Theorem 4.1. Thus as we take a step along the direction $-g$, whatever eigenvector we are optimizing can be continued as an analytic function of the step length, *even if the associated eigenvalue crosses with other eigenvalues*. This observation leads to the following algorithm.

Basic Algorithm

1. Choose an initial design b_0 and a distinct eigenvalue $\lambda_k(b_0)$, with associated eigenvector v_0. Set $n = 0$. Choose a step parameter $\tau > 0$.
2. Compute the gradient g of $J(b_n)$, associated with the distinct eigenvector v_n.
3. Let $u = \operatorname{argmin}\{\|\pm u - v_n\| : u$ is an eigenvector of $\operatorname{diag}(b_n - \tau g) A \operatorname{diag}(b_n - \tau g)$, with $\langle u, u\rangle = 1\}$.
4. If $J(u) < J(v_n)$, then
   $v_{n+1} = u$,
   $b_{n+1} = P(b_n - \tau g)$,
   else $\tau = \tau/2$.
5. Set $n = n+1$ and check for convergence. If no convergence and τ is not too small, continue with step 2.

The idea in step 3 is to select, out of all eigenvectors at the next iterate, the one closest to the current eigenvector (modulo sign). Since each eigenvector varies analytically with respect to the step length, and the new eigenvectors are mutually orthogonal, for small enough step length we are guaranteed that this will select the analytic continuation of the current eigenvector.

The operator P in step 4 simply projects the step back into the admissible set, i.e., P is defined by $P(b) = \max\{\min\{b, b_1\}, b_0\}$, where $b_0 = 1/\sqrt{\epsilon_1}$ is the lower bound for b, and $b_1 = 1/\sqrt{\epsilon_0}$ is the upper bound.

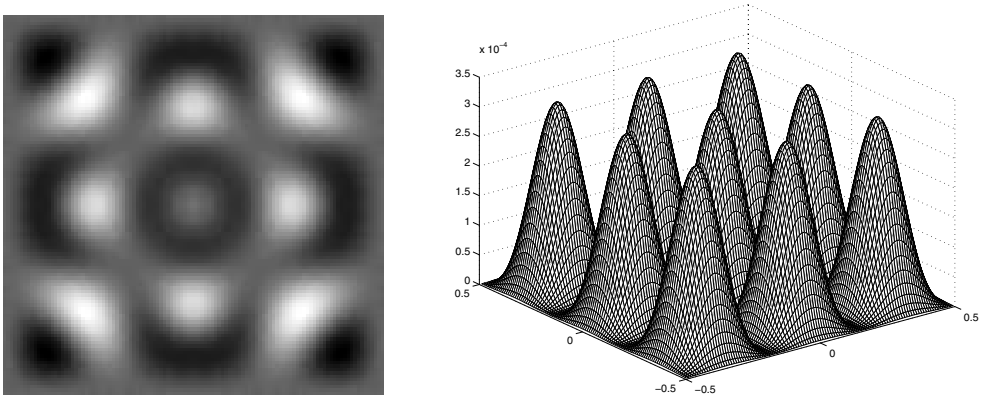
This basic algorithm solves neither the global problem nor the local problem described in section 3 but instead tracks along a single well-defined eigenvector, which

thanks to Theorem 4.1, can be followed all over the design space. The algorithm could be modified to find approximate local minima for either of the problems in section 3 but the basic implementation described here seems to be more versatile and of greater practical use.

5. Numerical results. To implement the basic algorithm, problem (2.1) was discretized by a simple 5-point finite difference scheme on a uniform square grid. Each step in the basic algorithm is then a relatively straightforward linear algebra operation. Our implementation uses Matlab with sparse matrix data structures wherever possible. For efficiency, eigenvalues and eigenvectors are found iteratively, using data from the previous optimization step as starting points. The adjoint equation (4.5) is also solved iteratively, using a biconjugate gradient algorithm. In the following we illustrate the results of three numerical experiments.

In the first experiment, we start with a homogeneous initial design $\epsilon(x) \equiv 1$. After discretizing problem (2.1) on a 112×112 grid, the first several eigenvalues and eigenvectors were computed. The homogeneous medium admits numerous multiple eigenvalues. A distinct eigenvalue λ_{11} was chosen whose associated eigenvector v_0 has the energy distribution pictured in Figure 5.1(b). The initial value of the objective was $J(v_0) \approx 6.44 \times 10^{-1}$. The initial gradient g is shown in Figure 5.1(a). We set the upper and lower bounds on $\epsilon(x)$ at $\epsilon_1 = 8$ and $\epsilon_0 = 1$. After approximately 2000 iterations of the algorithm, the design shown in Figure 5.2(a) was obtained, with energy density as pictured in Figure 5.2(b). The algorithm produced a localized eigenfunction despite the poorly localized initial guess. This was the case in all of our experiments: the algorithm does not seem to require a "good" initial guess. The final value of the objective was $J(v_f) \approx 9.71 \times 10^{-2}$.

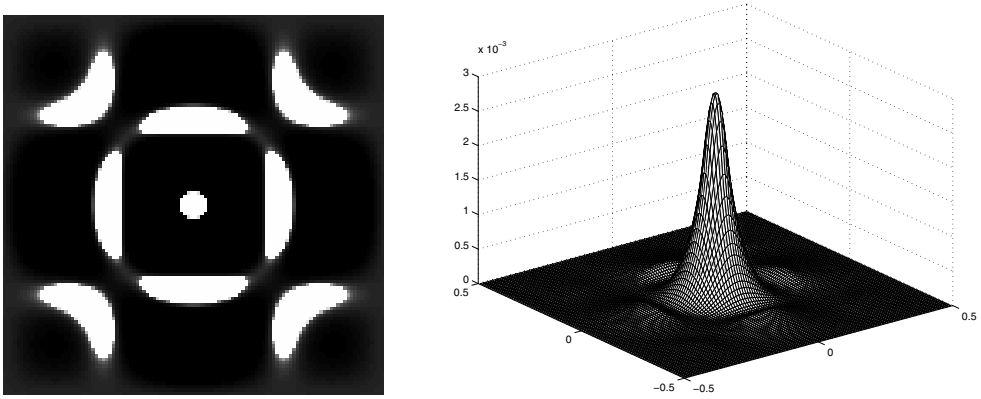
For the second experiment, we begin again with a homogeneous background on a 112×112 grid, but now with a single point material defect centered at the origin, as shown in Figure 5.3(a). The inclusion separates some of the eigenvalues and provides more variety for the choice of initial eigenvectors. An eigenvector v_0 associated with a distinct eigenvalue λ_{22}, whose frequency is somewhat higher than in the previous



(a) Gradient at first step.

(b) Energy density of initial eigenvector v_0.

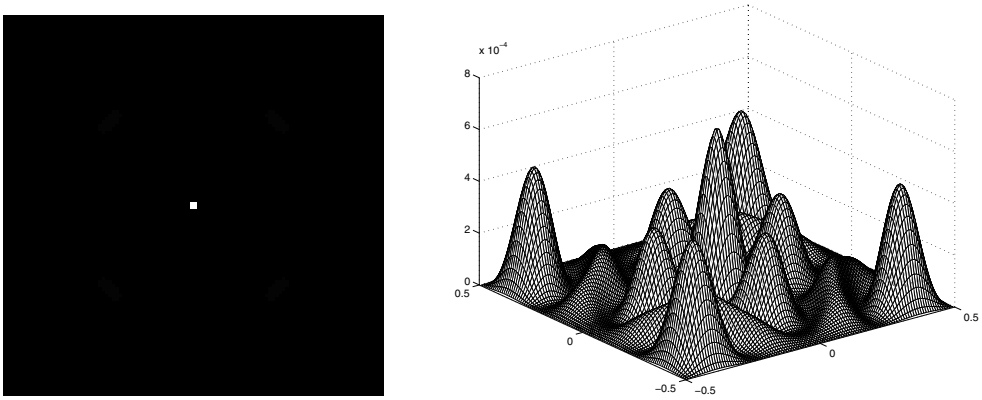
FIG. 5.1. *Initial values for the first experiment. Initial design $\epsilon(x)$ was constant.*



(a) Final profile of $\epsilon(x)$.

(b) Energy density of eigenvector.

FIG. 5.2. *Final values for the first experiment.*



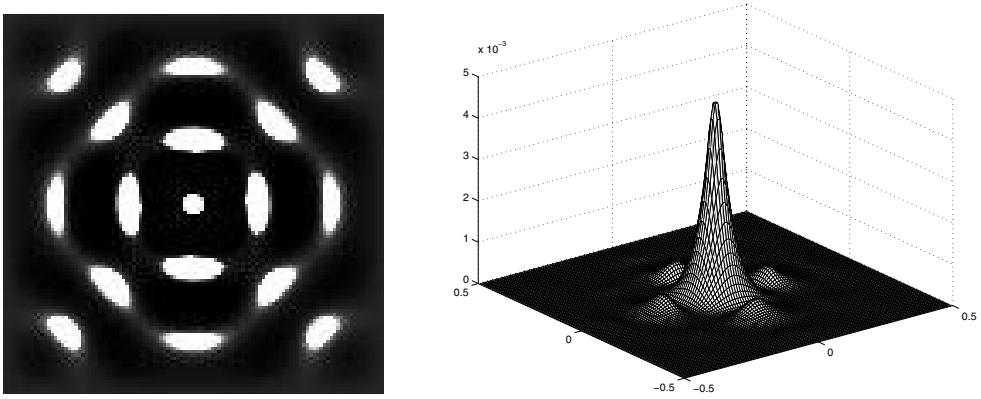
(a) Initial $\epsilon(x)$ is constant except for a small inclusion centered at the origin.

(b) Energy density of initial eigenvector.

FIG. 5.3. *Initial values for the second experiment.*

example, was chosen. The energy density of the initial eigenvector is pictured in Figure 5.3(b). The initial value of the objective was $J(v_0) \approx 6.59 \times 10^{-1}$. The material constraints were set at $\epsilon_1 = 8$ and $\epsilon_0 = 1$. After approximately 3000 iterations, the design shown in Figure 5.4 was obtained. Perhaps the most interesting feature of this example is the nearly periodic structure of the design away from the defect, resembling a photonic bandgap structure. The final value of the objective was $J(v_f) \approx 8.91 \times 10^{-2}$. This example achieves higher energy density at the origin than the previous example, due to the higher frequency of the mode.

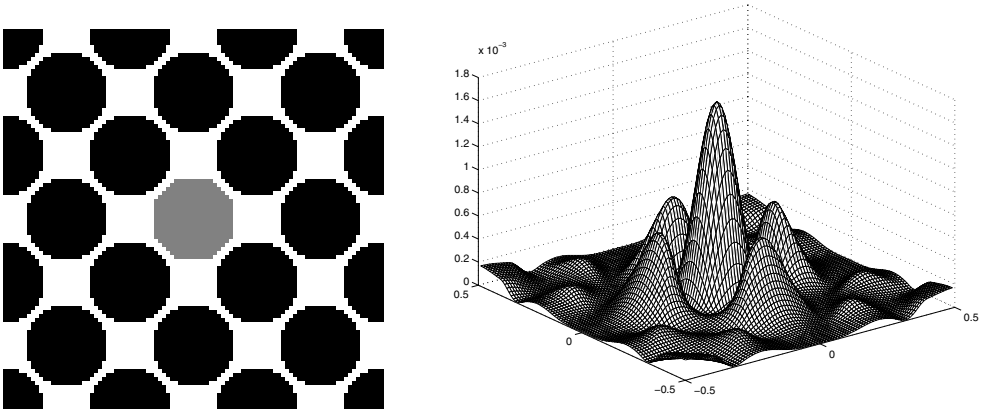
For the final numerical experiment, we take as an initial guess a hexagonal



(a) Profile of $\epsilon(x)$.

(b) Energy density of eigenvector.

FIG. 5.4. *Final values for the second experiment.*

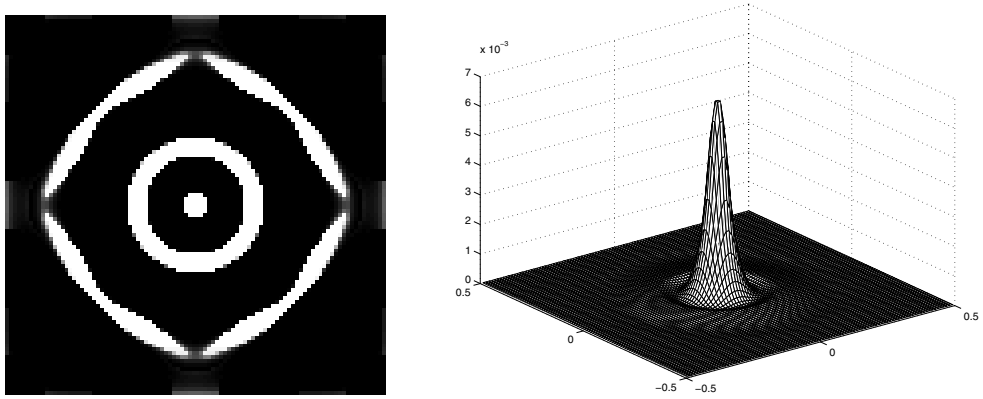


(a) Initial $\epsilon(x)$ is a hexagonal photonic bandgap structure with a defect at the origin.

(b) Energy density of the defect mode eigenvector, showing some localization.

FIG. 5.5. *Initial values for the third experiment.*

photonic bandgap structure with localized point defect as shown in Figure 5.5(a). The white areas in the image represent dielectric coefficient $\epsilon_1 = 9$; the black areas (holes) are $\epsilon_0 = 1$. The gray defect hole is filled with a material with $\epsilon = 5$. This structure, with periodic boundary conditions, exhibits a complete photonic bandgap for E-parallel wave propagation in a narrow frequency range and admits a localized mode with most of the energy concentrated within the defect. Such a localized mode, calculated by a "supercell" method with periodic boundary conditions, is pictured in Figure 5.5(b). The structure, discretized on a 96×96 grid, was taken as the initial design b_0. We then changed to Dirichlet boundary conditions, recalculated the



(a) Optimized profile of $\epsilon(x)$. (b) Energy density of optimized eigenvector.

FIG. 5.6. *Final values for the third experiment.*

modes, and identified that one (and only one) localized eigenvector v_0 which had energy density similar to Figure 5.5(b). This localized mode v_0 was taken as the eigenfunction to optimize. The initial value of the objective was $J(v_0) \approx 3.97 \times 10^{-1}$. After nearly 10000 iterations, the structure b_f in Figure 5.6(a) was obtained, with eigenfunction energy density v_f shown in Figure 5.6(b). The final value of the objective was $J(v_f) \approx 4.57 \times 10^{-2}$. The irregularities in the outer ring of the design may be due to the influence of the Dirichlet boundary conditions; i.e., it is possible that on a larger domain the optimal solution would be concentric annuli.

The large number of iterations exhibited by the algorithm in each of the previous examples is partly due to obvious deficiencies in the algorithm and partly due to inherent difficulties with the problem. The main weaknesses of the algorithm are poor step size control and inefficient parameterization of the design space. The numerical experiments presented here suggest that optimal designs are "bang-bang," i.e., optimal $\epsilon(x)$ designs take on only the values ϵ_1 or ϵ_0 (analysis to support this assertion will be presented in another paper). For this reason, an approach based on a level-set parameterization of the designs (see [11]) would probably be more efficient.

The large number of iterations required by the algorithm is also a symptom of the inherent ill-conditioning of the problem. Near an optimal design, modal energy decays very rapidly away from the origin. It follows from (4.6) that the magnitude of the gradient also decays rapidly with increasing $|x|$. The objective J is thus increasingly insensitive to changes in the design away from the origin. As the algorithm iterates, rapid local "convergence" near the center of the picture is normally observed (producing large decreases in J), followed by extremely slow changes further away from the origin, and much smaller decreases in J.

In computing the examples above, eigenvalue crossings often occurred as the iteration proceeded. The algorithm had no problems tracking the correct eigenvector through the crossings. In the second example, a nearby eigenvalue asymptotically approached the eigenvalue being optimized but never crossed. When the iteration stopped, the distance between the two was less than 10^{-4}, so that a small perturbation of the design could result in the eigenspace associated with the optimized eigenvector

becoming multidimensional.

6. Discussion. We studied a simplified version of a problem arising in the design of photonic bandgap devices. The problem solved involves finding a distribution of the material properties in an inhomogeneous medium such that one of its Dirichlet eigenfunctions is highly localized. We describe two versions of well-posed optimization problems associated with the design problem. A numerical method, which essentially is a descent algorithm with trajectory following, is devised to solve the problem numerically. We demonstrate the behavior of the algorithm in several examples.

We find that the results are remarkable not only in that we find a highly localized eigenfunction, but that the resulting medium resembles a photonic bandgap structure consisting of a periodic background and a defect. It is important, however, to emphasize that the designs produced by this algorithm are *not* themselves necessarily photonic bandgap structures. Each structure obviously supports modes other than the highly localized state which is optimized. If the defect design is inserted in an infinitely periodic photonic bandgap structure, the other modes may propagate, or they may remain localized, depending on the bandgap frequencies of the surrounding structure.

This demonstration project points to several research directions. First, as described in the previous section, it would be desirable to develop a faster algorithm for solving the optimization procedure. Second, the approach described in this work should be applied to a more realistic model governed by the vector Maxwell equations.

Finally, as pointed out in the numerical examples, the basic computational approach does not control the separation between the frequency of the localized mode and that of neighboring modes. As a practical consequence, particularly when losses are included, the frequency response of the structure to broadband sources may be somewhat spread out, without the narrow peak associated with high quality resonators. The *quality factor* "Q" of a resonator (or resonance) is usually defined to be inversely proportional to the width of the peak. For some engineering applications, it would be useful to modify this approach to produce designs for defect resonances with a specified quality factor "Q." This could be achieved by controlling the distance between the optimized eigenvalue and its neighbors across a family of eigenproblems associated with various propagation directions through the structure.

REFERENCES

[1] S. Cox and J. McLaughlin, *Extremal eigenvalue problems for composite membranes* I, Appl. Math Optim., 22 (1990), pp. 153–167.

[2] S. Cox and D. Dobson, *Maximizing band gaps in two-dimensional photonic crystals*, SIAM J. Appl. Math., 59 (1999), pp. 2108–2120.

[3] S. Cox and D. Dobson, *Band structure optimization of two-dimensional photonic crystals in H-polarization*, J. Comput. Phys., 158 (2000), pp. 214–224.

[4] A. Figotin and A. Klein, *Localization of light in lossless inhomogeneous dielectrics*, J. Opt. Soc. Amer. A, 15 (1998), p. 1423.

[5] A. Figotin and A. Klein, *Localized classical waves created by defects*, J. Statist. Phys., 86 (1997), pp. 165–177.

[6] A. Figotin and A. Klein, *Midgap defect modes in dielectric and acoustic media*, SIAM J. Appl. Math., 58 (1998), pp. 1748–1773.

[7] J. Joannopoulus and S. Johnson, *Photonic Crystals: The Road from Theory to Practice*, Kluwer, Norwell, MA, 2002.

[8] S. John, *Strong localization of photons in certain disordered dielectric superlattices*, Phys. Rev. Lett., 58 (1987), p. 2486.

[9] T. Kato, *Perturbation Theory for Linear Operators*, corrected printing of the 2nd ed., Springer-Verlag, Berlin, 1980.

[10] P. Kuchment, *The mathematics of photonic crystals*, in Mathematical Modeling in Optical Science, Frontiers Appl. Math. 22, G. Bao, L. Cowsar, and W. Masters, eds., SIAM, Philadelphia, 2001, pp. 207–272.
[11] F. Santosa, *A level-set approach for inverse problems involving obstacles*, ESAIM Contrôle Optim. Calc. Var., 1 (1995–1996), pp. 17–33.
[12] E. Yablonovitch, *Inhibited spontaneous emission in solid-state physics and electronics*, Phys. Rev. Lett., 58 (1987), pp. 2059–2062.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 775–799

A DIFFUSE INTERFACE MODEL FOR ALLOYS WITH MULTIPLE COMPONENTS AND PHASES*

HARALD GARCKE†, BRITTA NESTLER‡, AND BJÖRN STINNER†

Abstract. A nonisothermal phase field model for alloys with multiple phases and components is derived. The model allows for arbitrary phase diagrams. We relate the model to classical sharp interface models by formally matched asymptotic expansions. In addition we discuss several examples and relate our model to the ones already existing.

Key words. phase field models, sharp interface models, phase transitions, partial differential equations, alloy systems, matched asymptotic expansions

AMS subject classifications. 35K55, 82C26, 34E05, 35B25, 82C24

DOI. 10.1137/S0036139902413143

1. Introduction.

The phase field method is a powerful methodology to describe phase transition phenomena. The method has been used to describe solidification processes [7, 34] as well as microstructure evolution in solids [15] and liquid-liquid interfaces [28]. There are phase field models for pure substances [7, 34] and binary alloys [9, 21] for eutectic, peritectic, and monotectic systems [45, 31, 32, 33, 39]. Furthermore, the evolution of grain boundaries also can be modelled by phase field models or order parameter models [12, 16]. For recent reviews of phase field methods we refer to [13, 5, 14].

Traditionally the evolution of interfaces, such as the liquid-solid interface, has been modelled as a moving boundary problem. This means that pure phases are separated by a sharp interface. In the phases, partial differential equations, e.g., describing mass and heat diffusion, are solved. These equations are coupled by boundary conditions on the interface, such as the Stefan condition demanding energy balance and the Gibbs–Thomson equation. Across the sharp interface certain quantities (e.g., the heat flux, the concentration or the energy) may suffer jump discontinuities.

In phase field models the individual phases are distinguished by one or more so-called phase fields. In different phases the phase fields attain different values and interfaces are now modelled by a diffuse interface; i.e., the phase fields and all other quantities do not jump across an interface, but they change smoothly on a very thin transition layer (the diffuse interface). For example, for a solid-liquid phase transition we choose a phase field taking the value one in the solid and zero in the liquid; across an interface, the phase field varies smoothly from one to zero.

The use of diffuse interface models to describe interfacial phenomena dates back to van der Waals [42], Landau and Ginzburg [26], and Cahn and Hilliard [10]. In the theory of solidification this idea was introduced by Langer [27] and Caginalp [7]. Caginalp and Fife [8] used asymptotic expansions to relate the phase field models

*Received by the editors August 16, 2002; accepted for publication (in revised form) August 20, 2003; published electronically March 11, 2004. This work was supported by the DFG (Deutsche Forschungsgemeinschaft) through the Schwerpunktprogramm 1095 "Analysis, Modeling and Simulation of Multiscale Problems."

http://www.siam.org/journals/siap/64-3/41314.html

†NWF I - Mathematik, Universität Regensburg, 93040 Regensburg, Germany (harald.garcke@mathematik.uni-regensburg.de, bjoern.stinner@mathematik.uni-regensburg.de).

‡Fachbereich Informatik, FH Karlsruhe, Moltkestraße 30, 76133 Karlsruhe, Germany (britta.nestler@fh-karlsruhe.de).

proposed by Langer to classical free boundary problems in the sharp interface limit. This relation has also been rigorously established for some cases (see, for example, [38, 41] and the references therein).

Since the original phase field model is not derived from thermodynamical principles, a number of so-called thermodynamically consistent phase field models were proposed in the 1990s (see Penrose and Fife [34], Alt and Pawlow [2], Wang et al. [44]). All of these models guarantee a positive entropy production.

The classical asymptotics leads to restrictions on parameters which often makes it difficult to perform practical computations of realistic solidification processes. This is particularly true in the regime of small undercooling. In recent years Karma and Rappel [23, 24] (see also [25, 1, 30]) used the so-called thin interface asymptotics to realize numerical simulations in this regime. There, the Gibbs–Thomson equation is approximated to a higher order and the temperature profile in the interfacial region is recovered with a higher accuracy when compared to the classical asymptotics. Further numerical simulations (see [35, 36, 37]) confirm the superiority of this approach in the case of small undercooling.

So far, generalizing this approach to more general situations (see the discussion in [25]) and, in particular, extending the approach to phase field systems handling multiple phases are still an open problem. Therefore, as a first step, we apply classical sharp interface asymptotics to handle general systems with multiple phases and components. The task of making this approach more efficient by the use of thin interface asymptotics is left to further research.

The aim of this paper is to derive a phase field model that

- is thermodynamically consistent,
- allows for an arbitrary number of phases and components,
- is defined solely via the bulk free energies of the individual phases, the surface energy densities (surface entropy densities, respectively) of the interfaces, and diffusion and mobility coefficients, and
- yields classical moving boundary problems in the sharp interface limit.

The third requirement enables us to define the full set of phase field evolution equations by quantities which (in principal) can be measured. Since the bulk free energies determine the phase diagrams (see, e.g., Chalmers [11], Haasen [22]) our model can be used to model phase transitions for arbitrary phase diagrams. We note that in a multi-phase field model computing the surface free energy densities (or surface entropy densities) is difficult. Here one can make use of the studies by Garcke, Nestler, and Stoth [18], in which free energies for phase field methods with good calibration properties have been developed. This means that for given surface free energies (also called surface tensions) one can calibrate the parameters in the free energies of the phase field model in such a way that the sharp interface limit is defined via the given surface tensions. In particular the sharp interface problem is defined with the help of the surface free energies.

In the following section we introduce the phase field model in its full generality and state the corresponding sharp interface model. In section 3 we give examples and relate the model we propose to models already existing in the literature. Furthermore, we discuss a variety of different applications for the new model. Due to its general formulation, the model has the capability to describe phase transformation processes in nonisothermal multicomponent alloys as well as in grain structure evolution. Different phases and different crystal orientations can be distinguished at the same time by an appropriate choice of the phase field variables. This allows us to treat

effects occurring on different length scales such as eutectic grains and interdendritic structures.

Finally, we show in section 4 via formally matched asymptotic expansions that the phase field model yields the sharp interface model in the limit when the interfacial thickness tends to zero.

2. The models. We consider a domain $\Omega \subset \mathbf{R}^d$, $d \in \{1, 2, 3\}$, and we assume that the system has N components with M different phases possible.

2.1. The phase field model. The phase field model is based on an entropy functional of the form

$$S(e, c, \phi) = \int_\Omega \left(s(e, c, \phi) - \left(\varepsilon a(\phi, \nabla\phi) + \frac{1}{\varepsilon} w(\phi) \right) \right) dx. \tag{1}$$

We assume that the bulk entropy density s depends on the internal energy density e, the concentrations of the N components c_i, $1 \le i \le N$, and the phase field variable $\phi = (\phi_\alpha)_{\alpha=1}^M$. The variable ϕ_α denotes the local fraction of phase α, and we require that the concentrations of the components and the phase field variables fulfill the constraints

$$\sum_{i=1}^N c_i = 1, \quad \sum_{\alpha=1}^M \phi_\alpha = 1. \tag{2}$$

It will be convenient to use the free energy as a thermodynamical potential. We therefore postulate the Gibbs relation

$$df = -s dT + \sum_i \mu_i dc_i + \sum_\alpha r_\alpha d\phi_\alpha \tag{3}$$

(see Alt and Pawlow [3], who show that the Gibbs relation is a consequence of the entropy principle). Here, T is the temperature, $\mu_i = f_{,c_i}$ are the chemical potentials, and $r_\alpha = f_{,\phi_\alpha}$ are potentials due to the appearance of different phases.

We set

$$e = f + sT, \tag{4}$$

and hence

$$de = Tds + \sum_i \mu_i dc_i + \sum_\alpha r_\alpha d\phi_\alpha, \tag{5a}$$

$$ds = \frac{1}{T} de - \sum_i \frac{\mu_i}{T} dc_i - \sum_\alpha \frac{r_\alpha}{T} d\phi_\alpha. \tag{5b}$$

If we interpret s as a function of (e, c, ϕ), then we have

$$s_{,e} = \frac{1}{T}, \quad s_{,c_i} = \frac{-\mu_i}{T}, \quad s_{,\phi_\alpha} = \frac{-r_\alpha}{T}.$$

Later it will be convenient to switch among the variables (T, c, ϕ), (e, c, ϕ), (T, μ, ϕ), and $(-\frac{1}{T}, \frac{1}{T}\mu, \phi)$, and we therefore assume for the rest of this paper that

- $c \mapsto f(T, c, \phi)$ is strictly convex,
- $T \mapsto f(T, c, \phi)$ is strictly concave.

This will make the above exchanges of variables possible.

We note that given the free energy densities of the pure phases, we obtain the total free energy as a suitable interpolation of the free energies f_α, i.e., such that $f(T, c, e_\alpha) = f_\alpha(T, c)$, with e_α being the αth coordinate vector.

So far we have neglected interfacial effects. The thermodynamics of the interface gives additional contributions to entropy and free energy. Let us first consider how interfacial contributions are accounted for in a sharp interface model. Let $\Gamma_{\alpha\beta}$ denote an interface between phases α and β and let $\nu_{\alpha\beta}$ denote the unit normal at $\Gamma_{\alpha\beta}$ pointing into the β-phase. Then in sharp interface models an interfacial term

$$-\sum_{\substack{\alpha<\beta\\ \alpha,\beta=1}}^{M} \int_{\Gamma_{\alpha\beta}} \gamma_{\alpha\beta}(\nu_{\alpha\beta}) d\mathcal{H}^{d-1} \tag{6}$$

with a positive function $\gamma_{\alpha\beta}$ on S^{d-1} is added to the entropy (see [29], [43]). The notation $d\mathcal{H}^{d-1}$ indicates integration with respect to the $(d-1)$-dimensional surface measure.

In diffuse interface models the surface entropy functional (6) is replaced by a Ginzburg–Landau type functional of the form

$$-\int_\Omega \left(\varepsilon a(\phi, \nabla\phi) + \frac{1}{\varepsilon} w(\phi) \right) dx. \tag{7}$$

Here, a is the gradient energy density which is assumed to be homogeneous of degree two in the second variable; i.e.,

$$a(\phi, \eta X) = \eta^2 a(\phi, X) \quad \forall (\phi, X) \in \mathbf{R}^M \times \mathbf{R}^{d\times M} \text{ and } \forall \eta \in \mathbf{R}^+,$$

and w is a nonconvex function with exactly M global minima at the points $e_\beta = (\delta_{\alpha,\beta})_{\alpha=1}^M$, $1 \le \beta \le M$, with $w(e_\alpha) = 0$. It has been shown under appropriate assumptions on a that the functional (7) converges to the perimeter functional (6) when ε converges to zero. We refer to [18], [19] and section 3 for appropriate choice of a and w. We assume in this paper that a and w and, hence, the interfacial contributions to the entropy, do not depend on (T, c), but these dependences can be included, leading to a much more complicated model.

Our goal is to derive balance equations,

$$\partial_t e = -\nabla \cdot J_0 \qquad \text{(energy balance)}, \tag{8a}$$

$$\partial_t c_i = -\nabla \cdot J_i \qquad \text{(mass balances, } i = 1, \dots, N), \tag{8b}$$

that are coupled to

$$\partial_t \phi_\alpha = \textit{right-hand side (RHS)} \tag{8c}$$

in such a way that the second law of thermodynamics is fulfilled in an appropriate local version. Here, J_0 is the energy flux and $J_1, \dots, J_N$ are the fluxes of the components $c_1, \dots, c_N$. In order to derive appropriate expressions for the fluxes $J_0, \dots, J_N$, we use the generalized thermodynamic potentials (compare (5b)) $\frac{\delta S}{\delta e} = \frac{1}{T}$ and $\frac{\delta S}{\delta c_i} = \left(\frac{-\mu_i}{T}\right)$, which will drive the evolution. Now we appeal to nonequilibrium thermodynamics and postulate that the fluxes are linear functions of the thermodynamic driving forces

$\nabla\frac{\delta S}{\delta e}, \nabla\frac{\delta S}{\delta c_1}, \ldots, \nabla\frac{\delta S}{\delta c_N}$ to obtain

$$J_0 = L_{00}(T,c,\phi)\nabla\frac{\delta S}{\delta e} + \sum_{j=1}^{N} L_{0j}(T,c,\phi)\nabla\frac{\delta S}{\delta c_j}$$

$$= L_{00}(T,c,\phi)\nabla\frac{1}{T} + \sum_{j=1}^{N} L_{0j}(T,c,\phi)\nabla\frac{-\mu_j}{T}, \tag{9a}$$

$$J_i = L_{i0}(T,c,\phi)\nabla\frac{\delta S}{\delta e} + \sum_{j=1}^{N} L_{ij}(T,c,\phi)\nabla\frac{\delta S}{\delta c_j}$$

$$= L_{i0}(T,c,\phi)\nabla\frac{1}{T} + \sum_{j=1}^{N} L_{ij}(T,c,\phi)\nabla\frac{-\mu_j}{T} \tag{9b}$$

with mobility coefficients

$$(L_{ij})_{i,j=0,\ldots,N}.$$

To fulfill the constraint $\sum_{i=1}^{N} c_i = 1$ during the evolution, we assume

$$\sum_{i=1}^{N} L_{ij} = 0, \qquad j = 0,\ldots,N, \tag{10}$$

which implies $\sum_{i=1}^{N} J_i = 0$, and, hence, $\partial_t(\sum_{i=1}^{N} c_i) = \nabla\cdot(\sum_{i=1}^{N} J_i) = 0$. We further assume that L is symmetric (Onsager relations). In addition, L is assumed to be positive semidefinite; i.e.,

$$\sum_{i,j=0}^{N} L_{ij}\xi_i\xi_j \geq 0 \qquad \forall \xi = (\xi_0,\ldots,\xi_N) \in \mathbf{R}^{N+1}. \tag{11}$$

This condition will later ensure that an entropy inequality is satisfied. We note that we include cross effects between mass and energy diffusion in the model. One can neglect them by setting $L_{i0} = 0$ and $L_{0j} = 0$ for all $i,j \in \{1,\ldots,N\}$.

For the nonconserved phase field variables $\phi_1,\ldots,\phi_M$, we assume that the evolution is such that the system locally tends to maximize entropy conserving concentration and energy at the same time. Therefore we postulate

$$\begin{aligned} \omega\varepsilon\partial_t\phi_\alpha &= \frac{\delta S}{\delta\phi_\alpha} - \lambda \\ &= \varepsilon\big(\nabla\cdot a_{,X_\alpha}(\phi,\nabla\phi) - a_{,\phi_\alpha}(\phi,\nabla\phi)\big) - \frac{1}{\varepsilon}w_{,\phi_\alpha}(\phi) - \frac{f_{,\phi_\alpha}}{T} - \lambda, \end{aligned} \tag{12}$$

where we denote with $a_{,X_\alpha}$ the derivative with respect to the variables corresponding to $\nabla\phi_\alpha$. ω is (in this paper) a constant kinetic coefficient and λ is an appropriate Lagrange multiplier such that the constraint $\sum_{\alpha=1}^{M}\phi_\alpha = 1$ is satisfied; i.e.,

$$\lambda = \frac{1}{M}\sum_{\alpha}\left[\varepsilon\left(\nabla\cdot a_{,X_\alpha} - a_{,\phi_\alpha}\right) - \frac{1}{\varepsilon}w_{,\phi_\alpha} - \frac{f_{,\phi_\alpha}}{T}\right]. \tag{13}$$

Relevant for the dynamics are the variational derivatives of S that take the constraints (2) into account. We can therefore reformulate (9b) and (12) in terms of

the projection of $(\frac{\delta S}{\delta e}, \frac{\delta S}{\delta c_j}, \frac{\delta S}{\delta \phi_\alpha})$ onto the tangent space of the linear subspace whose elements satisfy the constraints. Defining

$$\Sigma^K = \{d \in \mathbf{R}^K : \quad \sum_{k=1}^{K} d_k = 1\},$$

and its tangent space

$$T\Sigma^K = \{d \in \mathbf{R}^K : \quad \sum_{k=1}^{K} d_k = 0\},$$

the constraints (2) read as $c \in \Sigma^N$ and $\phi \in \Sigma^M$. In the following, P^K will denote the projection onto $T\Sigma^K$. Then the relevant quantities for the definition of the fluxes are

$$\left(P^N\Big(-\frac{1}{T}\mu\Big)\right)_i = -\frac{1}{T}\left(\mu_i - \frac{1}{N}\sum_j \mu_j\right) = -\frac{1}{T}\frac{1}{N}\sum_j(\mu_i - \mu_j),$$

whereas there are no changes to $\frac{\delta S}{\delta e}$. We note that the quantities

$$\overline{\mu}_i = \frac{1}{N}\sum_j(\mu_i - \mu_j)$$

can be interpreted as generalized chemical potential differences. For two components we obtain $\overline{\mu}_1 = (\mu_1 - \mu_2)/2$, i.e., the usual chemical potential difference multiplied by the factor $1/2$.

With the above notation we can rewrite the fluxes as

$$J_0 = L_{00}(T,c,\phi)\nabla\frac{1}{T} + \sum_{j=1}^{N} L_{0j}(T,c,\phi)\nabla\frac{-\bar{\mu}_j}{T},$$

$$J_i = L_{i0}(T,c,\phi)\nabla\frac{1}{T} + \sum_{j=1}^{N} L_{ij}(T,c,\phi)\nabla\frac{-\bar{\mu}_j}{T}.$$

Similarly we can rewrite (12) as

$$\omega\varepsilon\partial_t\phi = P^M\left[\varepsilon\big(\nabla\cdot a_{,X}(\phi,\nabla\phi) - a_{,\phi}(\phi,\nabla\phi)\big) - \frac{1}{\varepsilon}w_{,\phi}(\phi) - \frac{f_{,\phi}}{T}\right].$$

Altogether the total entropy density is given by

$$\text{bulk entropy} + \text{surface entropy} = s(e,c,\phi) - \left(\varepsilon a(\phi,\nabla\phi) + \frac{1}{\varepsilon}w(\phi)\right),$$

and a straightforward computation shows (setting $\mu_0 = -1$)

$$\begin{aligned}
\partial_t(\text{entropy}) &= \partial_t\left(s(e,c,\phi) - \varepsilon a(\phi,\nabla\phi) - \frac{1}{\varepsilon}w(\phi)\right)\\
&= \sum_{i,j=0}^{N}\nabla\frac{-\mu_i}{T}\cdot L_{ij}\nabla\frac{-\mu_j}{T} - \nabla\cdot\left(\sum_{i,j=0}^{N}\frac{-\mu_i}{T}L_{ij}\nabla\frac{-\mu_j}{T}\right)\\
&\quad + \omega\varepsilon\sum_\alpha\left(\partial_t\phi_\alpha\right)^2 - \varepsilon\sum_\alpha\nabla\cdot\left(a_{,X_\alpha}\partial_t\phi_\alpha\right)\\
&\geq -\nabla\cdot\left(\sum_{i=0}^{N}\frac{-\mu_i}{T}J_i - \varepsilon\sum_{\alpha=1}^{M}a_{,X_\alpha}\partial_t\phi_\alpha\right).
\end{aligned}$$

The above inequality shows that the local entropy production is positive where the entropy flux J_s is given by

$$J_s = \sum_{i=0}^{N} \left(\frac{-\mu_i}{T} J_i \right) - \varepsilon \sum_{\alpha=1}^{M} a_{,p_\alpha} \partial_t \phi_\alpha. \tag{14}$$

The first term represents the entropy flux due to mass and energy diffusion, and the second one is due to moving phase boundaries (compare [2]). We refer to Alt and Pawlow [3], who show that for conserved phase fields (they call them order parameters) either the energy flux or the entropy flux has to depend on $\partial_t \phi$ in order to describe phase transitions.

2.2. The sharp interface model. In section 4 we use the method of asymptotic expansions to relate the phase field model of the previous subsection to the sharp interface model which we state in the following. We obtain that when the domain Ω is separated in phase regions $\Omega_1,\ldots,\Omega_M$ occupied by the pure phases $1,\ldots,M$ such that in every phase Ω_α, $\alpha = 1,\ldots,M$, the following evolution equations hold:

$$\partial_t e^\alpha = -\nabla \cdot \left(L^\alpha_{00}(T^\alpha, c^\alpha)\, \nabla \frac{1}{T^\alpha} - \sum_{j=1}^{N} L^\alpha_{0j}(T^\alpha, c^\alpha)\, \nabla \frac{\mu^\alpha_j}{T^\alpha} \right) \quad \text{(energy balance)}, \tag{15}$$

$$\partial_t c^\alpha_i = -\nabla \cdot \left(L^\alpha_{i0}(T^\alpha, c^\alpha)\, \nabla \frac{1}{T^\alpha} - \sum_{j=1}^{N} L^\alpha_{ij}(T^\alpha, c^\alpha)\, \nabla \frac{\mu^\alpha_j}{T^\alpha} \right) \forall i \ \text{(mass balances)}. \tag{16}$$

These equations can be formulated in the variables (T, μ) (in which case the internal energy e^α and the concentrations c^α are given as $e^\alpha = e^\alpha(T^\alpha, \mu^\alpha)$ and $c^\alpha = c^\alpha(T^\alpha, \mu^\alpha)$) or, more commonly, in the variables (T, c) (in which case the internal energy e^α and the chemical potentials μ^α are given as $e^\alpha = e^\alpha(T^\alpha, c^\alpha)$ and $\mu^\alpha = c^\alpha(T^\alpha, c^\alpha)$).

On a (smooth) boundary $\Gamma_{\alpha\beta}$ between two phases α and β we have (assuming an isotropic surface energy)

$$T^\alpha = T^\beta =: T \qquad \text{(continuity of temperature)}, \tag{17}$$

$$\bar\mu^\alpha_i = \bar\mu^\beta_i =: \bar\mu_i \quad \forall i \qquad \text{(continuity of chemical potentials)}, \tag{18}$$

$$[e]^\beta_\alpha\, v = [J_0]^\beta_\alpha \cdot \nu \qquad \text{(energy balance)}, \tag{19}$$

$$[c_i]^\beta_\alpha\, v = [J_i]^\beta_\alpha \cdot \nu \quad \forall i \qquad \text{(mass balances)}, \tag{20}$$

$$m_{\alpha\beta}\, v = \gamma_{\alpha\beta}\kappa + \frac{[f]^\beta_\alpha - \sum_i \bar\mu_i\, [c_i]^\beta_\alpha}{T} \quad \text{(Gibbs–Thomson relation)}. \tag{21}$$

Here, $\nu = \nu_{\alpha\beta}$ is the unit normal pointing into β, v is the speed of Γ in this direction, and κ is the mean curvature. The quantities

$$\bar\mu^\alpha_i = \mu^\alpha_i - \frac{1}{N} \sum_{j=1}^{N} \mu^\alpha_j = \frac{1}{N} \sum_{j=1}^{N} (\mu^\alpha_i - \mu^\alpha_j), \tag{22}$$

where $\mu^\alpha_i = f^\alpha_{,c_i}(T, c)$ are the generalized chemical potential differences in phase α, and $[\cdot]^\beta_\alpha$ denotes the jump of the quantity in the brackets across the interface. The quantity

$\gamma_{\alpha\beta}$ is the surface entropy density and the relation between the surface entropy and the entropy density in the phase field model is given by

$$\gamma_{\alpha\beta} = \inf_{p} \left\{ 2 \int_{-1}^{1} \sqrt{w(p)} \sqrt{a(p, p' \otimes \nu)} \right\}, \tag{23}$$

where the infimum is taken over all Lipschitz continuous functions p connecting the minima of w corresponding to the phases adjacent to the interface, i.e., $p(-1) = e_\alpha$ and $p(1) = e_\beta$. The kinetic coefficient $m_{\alpha\beta}$ can also be expressed in terms of the minimizer p (see [17]).

In general, a and w might depend on temperature and on the concentrations leading to a temperature- and concentration-dependent surface entropy in the sharp interface limit. In this case, the surface terms would also enter the internal energy.

For a thin interface analysis of a partially linearized model for pure substances we refer to [30]. Performing a thin interface analysis for our model would require studying higher order corrections of fields like s, f, T, and c in the interface region. We do not pursue this issue further at this stage.

We note that the Gibbs–Thomson equation can be derived by locally maximizing entropy, conserving concentration and energy at the same time. For a stationary flat interface the equations (17), (18), and (21) yield the classical equilibrium for phase boundaries. The equilibrium condition at a flat boundary at rest separating phases α and β is

$$\bar{\mu}_i^\alpha = \bar{\mu}_i^\beta \qquad \text{for all} \quad i = 1, \dots, N.$$

In addition the temperature has to be the same and (see (21))

$$[f]_\alpha^\beta - \sum_i \bar{\mu}_i \, [c_i]_\alpha^\beta = 0.$$

For M phases to be in equilibrium we therefore have $(N+1)(M-1)$ conditions. For each phase we can choose $N-1$ components and the temperature. All together there are

$$MN - (N+1)(M-1) = N - M + 1$$

degrees of freedom. This is the *Gibbs phase rule*. We note that for two component systems the equilibrium conditions between two phases lead to the well-known common tangent construction.

Finally, at triple junctions where three phases α, β, and δ meet, a force balance of the form

$$\gamma_{\alpha\beta}\tau_{\alpha\beta} + \gamma_{\beta\delta}\tau_{\beta\delta} + \gamma_{\delta\alpha}\tau_{\delta\alpha} = 0 \tag{24}$$

has to hold (compare [19]). Here, $\tau_{\alpha\beta}, \tau_{\beta\delta}$, and $\tau_{\delta\alpha}$ are the tangents to the interfaces $\Gamma_{\alpha\beta}, \Gamma_{\beta\delta}$, and $\Gamma_{\delta\alpha}$. All are assumed to either point in the direction of the triple junction or point away from the triple junction at the same time. It can be easily seen that this force balance is equivalent to certain angle conditions at the triple junction.

In the appendix we will demonstrate that the entropy does not decrease for solutions of the above problem. In particular, for a closed system we obtain, using

appropriate transport theorems and assuming $m \geq 0$ and $L = (L_{ij})_{i,j=1,\dots,N}$ is positive semidefinite, the following:

$$\frac{d}{dt}\Big(\int_\Omega s(e,c)dx - \int_\Gamma \gamma d\mathcal{H}^{d-1}\Big) = \int_\Omega \Big(\nabla\frac{1}{T}\cdot J_0 + \sum_i \nabla\frac{-\bar{\mu}_i}{T}\cdot J_i\Big)dx, \\ + \int_\Gamma mv^2 d\mathcal{H}^{d-1} \quad \geq 0,$$

where the integral over Γ is an integral over all possible interfaces.

3. Examples. In this section we will first demonstrate that the phase field method is able to model systems with a very general class of phase diagrams. In the way it is formulated, the model can describe systems with concave entropies $s_\alpha(e,c)$ in the pure phases. This corresponds to free energies $f_\alpha(T,c)$ which are convex in c and concave in T. In the case where $f(T,c)$ is not convex in the variable c, the free energy needs to contain gradients of the concentrations (as in the Cahn–Hilliard model).

We will first give a rather general example, which already covers most examples in practice, and then discuss relations to existing models and possible partial linearizations of the system.

3.1. Possible choices of the free energy. Choosing the phase field ϕ such that $\phi = e_M$ corresponds to the liquid phase, we define bulk free energies for the individual phases by

$$f_\alpha(T,c) = \sum_{i=1}^N \left(c_i L_i^\alpha \frac{T - T_i^\alpha}{T_i^\alpha} + \frac{R}{v_m} T c_i \ln(c_i)\right) - c_v T(\ln(T) - 1)$$

with $L_i^M = 0$, and L_i^α, $i = 1,\dots,N$, $\alpha = 1,\dots,M-1$, being the latent heat per unit volume of the phase transition from phase α to the liquid phase of the pure component i. Furthermore, T_i^α, $i = 1,\dots,N$, $\alpha = 1,\dots,M-1$, is the melting temperature of the ith component in the phase α, and c_v is the specific heat, which is assumed to be independent of c and ϕ; the molar volume v_m is supposed to be a constant, and R is the gas constant. Then we define the total free energy density as follows:

$$f(T,c,\phi) := \sum_{\alpha=1}^M \sum_{i=1}^N \left(c_i L_i^\alpha \frac{T - T_i^\alpha}{T_i^\alpha} h(\phi_\alpha)\right) \\ + \sum_{i=1}^N \left(\frac{R}{v_m} T c_i \ln(c_i)\right) - c_v T(\ln(T) - 1), \tag{25}$$

where h is a monotone function on $[0,1]$ that satisfies $h(0) = 0$ and $h(1) = 1$. Examples are $h(\phi) = \phi$ and $h(\phi) = \phi^2(3 - 2\phi)$. The last one has the property $h'(0) = h'(1) = 0$ which is suitable for phase field models as we will see below. With this choice of h the function f is an interpolation of the individual free energy densities f_α.

We can calculate

$$s = -f_{,T} = -\sum_{\alpha=1}^M \sum_{i=1}^N \left(c_i \frac{L_i^\alpha}{T_i^\alpha} h(\phi_\alpha)\right) - \sum_{i=1}^N \left(\frac{R}{v_m} c_i \ln(c_i)\right) + c_v \ln(T), \tag{26}$$

so that

$$e = f + Ts = -\sum_{\alpha=1}^{M}\sum_{i=1}^{N}(c_i L_i^\alpha h(\phi_\alpha)) + c_v T. \tag{27}$$

We note that if $L_i^\alpha = L^\alpha$ for all components i, then e does not depend on c. The chemical potentials are given as

$$\mu_i(T,c,\phi) = f_{,c_i}(T,c,\phi) = \sum_{\alpha=1}^{M}\left(L_i^\alpha \frac{T-T_i^\alpha}{T_i^\alpha}h(\phi_\alpha)\right) + \frac{R}{v_m}T(\ln(c_i)+1). \tag{28}$$

Expressions for the quantities above in the pure phases are obtained by setting $\phi_\alpha = e_\alpha$. For example, we have

$$\mu_i^\alpha = \partial_{c_i} f_\alpha = \partial_{c_i} f(T,c,e_\alpha) = L_i^\alpha \frac{T-T_i^\alpha}{T_i^\alpha} + \frac{R}{v_m}T(\ln(c_i)+1)$$

for the chemical potential of the ith component in the phase α.

Now we give some examples for the terms modelling interfacial contributions to the free energy. The simplest form of the gradient energy is

$$a(\phi,\nabla\phi) = |\nabla\phi|^2 = \sum_{\alpha=1}^{M}|\nabla\phi_\alpha|^2.$$

However, it has been shown [17, 19, 39] that gradient energies of the form

$$a(\phi,\nabla\phi) = \sum_{\substack{\alpha,\beta=1\\ \alpha<\beta}} A_{\alpha\beta}(\phi_\alpha\nabla\phi_\beta - \phi_\beta\nabla\phi_\alpha),$$

where $A_{\alpha\beta}$ are convex functions that are homogeneous of degree two, are more convenient with respect to the calibration of parameters in the phase field model to the surface terms in the sharp interface model. A choice that leads to isotropic surface terms is

$$a(\phi,\nabla\phi) = \sum_{\alpha<\beta}\frac{\tilde\gamma_{\alpha\beta}}{\tilde m_{\alpha\beta}}|\phi_\alpha\nabla\phi_\beta - \phi_\beta\nabla\phi_\alpha|^2$$

with constants $\tilde\gamma_{\alpha\beta}$ and $\tilde m_{\alpha\beta}$ that can be related to $\gamma_{\alpha\beta}$ and $m_{\alpha\beta}$ in (21) (cf. [17]). For the bulk potential one may take the standard multiwell potential

$$w_{st}(\phi) = 9\sum_{\alpha<\beta}\tilde m_{\alpha\beta}\tilde\gamma_{\alpha\beta}\phi_\alpha^2\phi_\beta^2$$

or a higher order variant

$$\tilde w_{st}(\phi) = w_{st}(\phi) + \sum_{\alpha<\beta<\delta}\gamma_{\alpha\beta\delta}\phi_\alpha^2\phi_\beta^2\phi_\delta^2.$$

For practical computations the multiobstacle potential yields good calibration properties. It is defined by

$$w_{ob}(\phi) = \frac{16}{\pi^2}\sum_{\alpha<\beta}\tilde m_{\alpha\beta}\tilde\gamma_{\alpha\beta}\phi_\alpha\phi_\beta$$

with a higher order variant

$$\tilde{w}_{ob}(\phi) = w_{ob}(\phi) + \sum_{\alpha<\beta<\delta} \gamma_{\alpha\beta\delta}\phi_\alpha\phi_\beta\phi_\delta,$$

where w_{ob} and $\tilde{w}_{ob}$ are defined to be infinity whenever ϕ is not on the Gibbs simplex $G = \{d \in \Sigma^M : d_\alpha \geq 0\}$. We refer to [18] and [19] for a further discussion of the properties of the surface terms.

3.2. Possible choices of the mobility matrix. Here we give an example only for the part of the mobility matrix $(L_{ij})_{i,j=0,\dots,N}$ that defines mass diffusion resulting from chemical potential differences; i.e., we do not specify $L_{i0} = L_{0i}$ for $0 \leq i \leq N$. An example for those terms, which in particular define cross effects between mass and energy diffusion, will be given in section 3.4.

If $l_i(c_i, T, \phi)$ are the nonnegative bare mobilities of the pure components, we can argue as in [4] to obtain

$$L_{ij}(T, c, \phi) = l_i(T, c_i, \phi)\left(\delta_{ij} - \Big(\sum_{q=1}^{N} l_q(T, c_q, \phi)\Big)^{-1} l_j(T, c_j, \phi)\right), \quad 1 \leq i, j \leq N.$$

To give a simple example, we assume that all bare mobilities are the same constant (e.g., $l_i(T, c_i, \phi) = 1$). Hence

$$(L_{ij})_{i,j=1}^N = id - \frac{1}{N}\mathbf{1} \otimes \mathbf{1},$$

where $\mathbf{1} = (1, \dots, 1)$ and $\otimes$ is the tensor product. Often it is more reasonable to assume that the bare mobilities l_i are linear in c_i, and in the simplest case ($l_i(T, c_i, \phi) = c_i$) we obtain

$$(L_{ij})_{i,j=1}^N = (c_i(\delta_{ij} - c_j))_{i,j=1}^N.$$

Choosing a free energy of the form (25) and taking (28) into account, we get the following equations for the concentrations:

$$\begin{aligned}\partial_t c_i = -\nabla \cdot \Bigg[& L_{i0}\nabla\frac{1}{T} + \sum_{j=1}^{N} c_i(\delta_{ij} - c_j)\nabla\Big(- \sum_{\alpha=1}^{M}\left(L_j^\alpha\left(\frac{1}{T_j^\alpha} - \frac{1}{T}\right)h(\phi_\alpha)\right) \\ & - \frac{R}{v_m}(\ln(c_j) + 1)\Big)\Bigg] \\ = \nabla \cdot & \left[L_{i0}\nabla\frac{1}{T} + \sum_{\alpha=1}^{M}\sum_{j=1}^{N} L_{ij}\nabla\Big(L_j^\alpha\left(\frac{1}{T_j^\alpha} - \frac{1}{T}\right)h(\phi_\alpha)\Big)\right] + \frac{R}{v_m}\Delta c_i.\end{aligned}$$

3.3. Relation to the Penrose–Fife model. In this subsection we will demonstrate that our model includes the model of Penrose and Fife [34] as a special case. In this case there is only one component, and we can neglect the variable c. There are two phases, so we will write the equations in terms of the solid fraction $\psi = \phi_1$. Then by (2), $\phi_2 = 1 - \psi$.

The first phase, the solid one, is characterized by $\phi = 1$; hence $\psi = 1$. We assume its free energy density to be

$$f^s = L\frac{T - T_m}{T_m} - c_v T(\ln(T) - 1),$$

where T_m is the melting temperature and L the latent heat of the solid-liquid phase transition. The second phase, the liquid one, is characterized by $\phi = e_2$; therefore $\psi = 0$, and we take the free energy density to be

$$f^l = -c_v T(\ln(T) - 1).$$

We have

$$f(T, \psi) = L\frac{T - T_m}{T_m} h(\psi) - c_v T(\ln(T) - 1);$$

hence

$$s(T, \psi) = -\frac{L}{T_m} h(\psi) + c_v \ln(T)$$

so that $e(T, \psi) = -Lh(\psi) + c_v T$. The evolution equation for the energy density yields

$$c_v \partial_t T - Lh'(\psi)\partial_t \psi = -\nabla \cdot \left(L_{00} \nabla \frac{1}{T} \right).$$

Now we choose $L_{00} = c_v K_2 T^2$, $\lambda(\psi) = Lh'(\psi)/c_v$, and

$$a(\phi, \nabla\phi) = \frac{c}{2}|\nabla\phi|^2 = \frac{c}{2}(|\nabla\phi_1|^2 + |\nabla\phi_2|^2),$$

where $c = \kappa_1 c_v/(2\varepsilon)$ for some constant κ_1. Setting $\omega = 1$, $K_1 = c_v/(2\varepsilon)$ and

$$s_0(\psi) = -\frac{1}{\varepsilon c_v} w(\psi, 1 - \psi) - \frac{L}{c_v T_m} h(\psi).$$

We arrive at the system

$$\begin{aligned} \partial_t \psi &= K_1 \left(\frac{\lambda(\psi)}{T} + s_0'(\psi) + \kappa_1 \Delta\psi \right), \\ \partial_t T - \lambda(\psi)\partial_t \psi &= K_2 \Delta T \end{aligned}$$

which is the model of Penrose and Fife [34, Chapter 6].

3.4. A linearized model. In this subsection we are going to partially linearize our model. This is done in such a way that the evolution equations in the pure phases are linear and they indeed reduce to standard linear diffusion equations. We restrict ourselves to binary systems but a generalization to higher order systems is straightforward.

We denote by $c = c_1$ the concentration of the first component; therefore $c_2 = 1 - c$. Using that L is symmetric and the algebraic constraints (10), we obtain

$$L_{01} = L_{10} = -L_{02} = -L_{20} \qquad \text{and} \qquad L_{11} = L_{22} = -L_{12} = -L_{21}.$$

Furthermore, we introduce the chemical potential difference

$$\mu = f_{,c} = f_{,c_1} - f_{,c_2} = \mu_1 - \mu_2.$$

Then the conservation laws for energy and concentration read (up to a factor 2 in the last term of the right-hand sides)

$$\partial_t e = -\nabla \cdot L_{00} \nabla \frac{1}{T} - \nabla \cdot L_{10} \nabla \frac{-f_{,c}}{T}, \tag{29}$$

$$\partial_t c = -\nabla \cdot L_{10} \nabla \frac{1}{T} - \nabla \cdot L_{11} \nabla \frac{-f_{,c}}{T}. \tag{30}$$

Choosing

$$L_{11} = D\frac{T}{f_{,cc}}, \quad L_{10} = L_{01} = e_{,c} D \frac{T}{f_{,cc}}, \quad \text{and} \quad L_{00} = e_{,c}^2 D \frac{T}{f_{,cc}} + KT^2,$$

the system (29)–(30) reduces to

$$\partial_t e = \nabla \cdot \left(K \nabla T + e_{,c} D \nabla c + e_{,c} D \frac{f_{,c\phi}}{f_{,cc}} \nabla \phi \right), \tag{31}$$

$$\partial_t c = \nabla \cdot \left(D \nabla c + D \frac{f_{,c\phi}}{f_{,cc}} \nabla \phi \right). \tag{32}$$

Here K and D are coefficients that may depend on ϕ. Equations (31) and (32) then have to be coupled to the phase field system (12).

We assume as in (27) that the internal energy density is affine linear in the variables (T, c). Then the system (31)–(32) reduces in regions where ϕ is constant, i.e., in the pure phases, to (here K and D are constants)

$$c_v \partial_t T = \nabla \cdot K \nabla T = K \Delta T, \quad \partial_t c = \nabla \cdot D \nabla c = D \Delta c.$$

Here c_v is the specific heat. These are classical linear diffusion equations for temperature (Fourier's law) and concentration (Fick's law).

3.5. Relation to the Caginalp model. If we further linearize the system it can be seen that our model leads to a generalization of the original phase field model [7] to the case of alloy solidification. We consider a three-phase system for a binary alloy. We choose the free energy density

$$f(T, c, \phi) = \Big(\kappa \frac{c}{2} - \sum_{\alpha=1}^{3} L_1^\alpha \phi_\alpha \Big) cT - c_v T(\ln(T) - 1) - \sum_{\alpha=1}^{3} L_2^\alpha \phi_\alpha,$$

where L_2^α are latent heat coefficients and L_1^α and κ, respectively, are coefficients entering the chemical potentials. Then we get

$$\begin{aligned} s &= -f_{,T} = -\Big(\kappa \frac{c}{2} - \sum_{\alpha=1}^{3} L_1^\alpha \phi_\alpha \Big) c + c_v \ln(T), \\ e &= f + Ts = c_v T - \sum_\alpha L_2^\alpha \phi_\alpha, \\ \frac{\mu}{T} &= \frac{f_{,c}}{T} = \kappa c - \sum_\alpha L_1^\alpha \phi_\alpha, \\ \frac{r_\alpha}{T} &= \frac{f_{,\phi_\alpha}}{T} = -L_1^\alpha c - \frac{L_2^\alpha}{T}. \end{aligned}$$

Choosing the mobility matrix as in the previous subsection we obtain

$$\partial_t e = \partial_t \Big(c_v T - \sum_\alpha L_2^\alpha \phi_\alpha \Big) = \nabla \cdot (K \nabla T),$$
$$\partial_t c = \nabla \cdot D \nabla \Big(\kappa c - \sum_\alpha L_1^\alpha \phi_\alpha \Big).$$

For the gradient energy we take the isotropic function $a(\phi, \nabla\phi) = \frac{1}{2}\sum_\alpha |\nabla\phi_\alpha|^2$. Then the equations for the phase field variables are

$$\omega\varepsilon\partial_t\phi_\alpha = \varepsilon\Delta\phi_\alpha - \frac{1}{\varepsilon} w_{,\phi_\alpha}(\phi) + L_1^\alpha c + \frac{L_2^\alpha}{T} - \lambda,$$

where λ is the Lagrange multiplier (13). Now we linearize the term $\frac{1}{T}$ in the above equation around a temperature T_m to obtain

$$\omega\varepsilon\partial_t\phi_\alpha = \varepsilon\Delta\phi_\alpha - \frac{1}{\varepsilon} w_{,\phi_\alpha}(\phi) + L_1^\alpha c + L_2^\alpha \left(\frac{1}{T_m} - \frac{1}{T_m^2}(T - T_m) \right) - \lambda.$$

The equations for (T, c) are linear and all terms in the equation for ϕ are linear except for the term $w_{,\phi_\alpha}$. A complete linearization cannot be expected because systems with moving interfaces can never be linear, as can be easily seen for the sharp interface model.

Finally, we note that this simplification of the model leads to a linearized phase diagram; in particular, the magnitude of the jump of the concentration in the sharp interface model is constant for each of the phase boundaries.

3.6. Fields of application. In this paragraph, we comment on the generality of the presented phase field model, on the new features, and on the various different applications to solidification processes, microstructure formation, and polycrystalline grain growth. With the phase field model set up for an arbitrary number of alloy components and phases in a nonisothermal system, the set of governing equations is able to describe the coupled heat and mass diffusion processes as well as the phase transformations in multicomponent systems. Due to the flexibility to choose parameters in the gradient and in the potential free energy, the model consists of enough degrees of freedom to prescribe the physics of each phase boundary and interface separately by defining values for appropriate surface energies $\tilde{\gamma}_{\alpha\beta}$ and for the mobilities $\tilde{m}_{\alpha\beta}$. The model allows for both kinetic and surface energy anisotropies. Different types of anisotropy such as smooth and crystalline expressions corresponding to Wulff shapes with a different number of vertices can be realized in three dimensions. Considering the application point of view, the effect of the type and strength of anisotropy on the growth structure can be investigated. Examples of experimentally observed anisotropic characteristics in eutectic systems are tilted or spiral phase formations and the growth of neighboring eutectic grains.

The phase field variables ϕ_α can represent different phases and different grains of orientational variants at the same time. Therefore, phenomena such as eutectic grain formation involving different length scales (grains on the larger scale and a eutectic structure on a smaller scale) and interpretations of the nonconserved order parameters can be described using the new model. A main focus of application in future development is the two- and three-dimensional numerical simulation of solidification in multicomponent alloy systems with arbitrary phase diagrams. By choosing the

specific thermodynamical quantities—the latent heats of fusion L_i^α and the melting temperatures T_i^α—and by inserting these data as input parameters for the numerical simulations, different types of phase transformations, such as peritectics, eutectics, and monotectics, are modelled. In particular, the stability of ternary eutectic lamellae with phase arrays of different period length and phase permutations will be investigated by phase field simulations in a forthcoming paper. The results of computed structures are compared with a generalization of the classical Jackson–Hunt theory for ternary eutectics. The occurrence of a ternary phase impurity leads to the formation of eutectic colonies. The resulting complex structure is of multiscale type and can also be modelled with the new approach.

4. Relating the models by asymptotic expansions. By matched asymptotic expansions we want to establish the relation between the phase field model and the sharp interface model that were described in section 2. We are going to generalize methods developed by Caginalp and Fife [8], Bronsard, Garcke, and Stoth [6], Garcke and Novick-Cohen [20], and Garcke, Nestler, and Stoth [17]. We restrict ourselves to two space dimensions, i.e., $d = 2$, but generalizations are possible.

Since the quantities $(T, \bar{\mu})$ are continuous across a phase boundary it will be convenient to use them in the asymptotic expansions. More precisely we will use the variables ϕ and $u = (\frac{-1}{T}, \frac{\bar{\mu}_1}{T}, \dots, \frac{\bar{\mu}_N}{T})$. Since $f(T, \cdot, \phi)$ is strictly convex and $f(\cdot, c, \phi)$ is strictly concave, we obtain that the mappings

$$(T, c, \phi) \mapsto (u, \phi) \quad \text{and} \quad (e, c, \phi) \mapsto (u, \phi)$$

are both invertible and an exchange of variables between these quantities is possible.

We will use the variables (u, ϕ) in the asymptotics but the equations can always be reinterpreted with respect to the variables (T, c, ϕ) or (e, c, ϕ). We write the conservation laws as

$$\partial_t c_i(u, \phi) = \nabla \cdot \sum_{j=0}^{N} L_{ij}(u, \phi) \nabla u_j, \quad 0 \le i \le N,$$

where we have set $c_0 = e$.

The phase field equations are

$$\omega \varepsilon \partial_t \phi = P^M \Big[\varepsilon \big(\nabla \cdot a_{,X}(\phi, \nabla\phi) - a_{,\phi}(\phi, \nabla\phi) \big) - \frac{1}{\varepsilon} w_{,\phi}(\phi) + u_0 f_{,\phi}(T(u, \phi), c(u, \phi), \phi) \Big].$$

We assume that the matrix $L = (L_{ij})_{i,j=0}^N$ is strictly positive definite for all arguments on the space

$$H^N := \Big\{ d = (d_i)_{i=0}^N \in \mathbf{R}^{N+1} : \sum_{i=1}^{N} d_i = 0 \Big\} = \mathbf{R} \times T\Sigma^N.$$

In addition, we will frequently make use of the fact that a is homogeneous of degree two in the variable X. In particular, we have (cf. [17])

$$a_{,X}(\phi, \eta X) : X = 2\eta a(\phi, X), \tag{33}$$

$$a_{,\phi}(\phi, \eta X) : X = \eta^2 a_{,\phi}(\phi, X), \tag{34}$$

$$a(\phi, 0) = 0, \tag{35}$$

$$a_{,X}(\phi, 0) = 0. \tag{36}$$

4.1. Outer expansion. We expect, based on experiences from numerical simulations, that several phases arise which are separated by diffuse interfaces whose thickness is of order ε. We will see that these phases correspond to the M minima of the potential w. In such a phase, away from an interface to another phase, we consider an outer expansion in the bulk region. For a function b in (t,x) we present the ansatz

$$b_{out}(t,x) = \sum_{K=0}^{\infty} \varepsilon^K b_{out}^K(t,x). \tag{37}$$

In this way we expand the variables u_j and ϕ_α, $0 \le j \le N$, $1 \le \alpha \le M$. For the constraints $\phi \in \Sigma^M$ and $u \in H^N$ to be satisfied we assume

$$\begin{aligned} &\phi_{out}^0 \in \Sigma^M, \quad \phi_{out}^K \in T\Sigma^M, \quad K \ge 1, \\ &u_{out}^K \in H^N, \quad K \ge 0. \end{aligned}$$

First we consider the equation for the phase field variables. We expand $P^M w_{,\phi}(\phi)$ as

$$P^M w_{,\phi}(\phi) = P^M w_{,\phi}(\phi_{out}^0) + \varepsilon (P^M w_{,\phi})_{,\phi}(\phi_{out}^0) \cdot \phi_{out}^1 + O(\varepsilon^2).$$

To leading order $O(\varepsilon^{-1})$ the equation (12) becomes

$$0 = P^M w_{,\phi}(\phi_{out}^0) = w_{,\phi}(\phi_{out}^0) - \frac{1}{M} \left(\sum_{\alpha=1}^{M} w_{,\phi_\alpha}(\phi_{out}^0) \right) \mathbf{1}. \tag{38}$$

As we are searching for stable solutions for this equation, ϕ_{out}^0 is one of the base vectors $\{e_\beta\}_{1 \le \beta \le M}$. We can conclude that to leading order the whole domain Ω is partitioned into phases which are characterized by the M possible values of ϕ_{out}^0.

The $O(1)$-equations for the conserved variables are $(0 \le i \le N)$

$$\partial_t c_i(u_{out}^0, \phi_{out}^0) = \nabla \cdot \sum_{j=0}^{N} L_{ij}(u_{out}^0, \phi_{out}^0) \nabla u_{j,out}^0. \tag{39}$$

Boundary conditions for these equations will be obtained by matching with the inner expansion. One should note that we have expanded the coefficients L_{ij} in $(u_{out}^0, \phi_{out}^0)$ in the same way as $P^M w_{,\phi}$ in ϕ_{out}^0. In phase α, i.e., at points where $\phi_{out}^0 = e_\alpha$, we write $L_{ij}^\alpha(u) = L_{ij}(u, e_\alpha)$. Then the $O(1)$-equations become

$$\partial_t c_i(u_{out}^0, e_\alpha) = \nabla \cdot \sum_{j=0}^{N} L_{ij}^\alpha(u_{out}^0) \nabla u_{j,out}^0.$$

Since $c_0 = e$, $u_0 = -\frac{1}{T}$, and $u_j = \frac{\mu_j}{T}$ we obtain (15) and (16). We note that an upper index in (15) and (16) refers to the phase, whereas an upper index in this section refers to the order in the expansion.

4.2. Inner expansion. Now we consider an interfacial region where two phases meet. Without loss of generality we assume that $\phi_{out}^0 = e_1$ in one of the outer regions, denoted by Ω_1, and $\phi_{out}^0 = e_2$ in the other one, denoted by Ω_2. We assume that these two regions are separated by a family $\{\Gamma_t\}_t$ of evolving smooth curves. Let ψ be a smooth function such that $s \mapsto \psi(t,s)$ is an arc-length parametrization of Γ_t. The

unit tangential vector $\tau(t,x)$ on Γ_t in $x=\psi(t,s)$ is given by $\tau(t,x)=\partial_s\psi(t,s)$, and the unit normal $\nu(t,x)$ on Γ_t in $x=\psi(t,s)$ is such that (ν,τ) is positively oriented. We choose the orientation in the parametrization ψ such that ν points into Ω_1.

Since the parametrization is smooth, it is possible to introduce new space coordinates $(z(t,x),s(t,x))$ in a strip S around Γ_t in the following way. We define $r(t,x)=d(x,\Gamma_t)$ to be the signed distance between a point x and Γ_t; i.e., r is positive in Ω_1 and negative in Ω_2. Then the variable z is defined by $z(t,x)=\frac{1}{\varepsilon}r(t,x)$. Let P_t be the projection of S onto Γ_t. Then by the smoothness of Γ_t one can use the strip S narrow enough such that there is exactly one $s(t,x)$ for every $x\in S$ such that $P_t(x)=\psi_t(s)$. The following holds:

$$\begin{aligned}\nabla_x z(t,x)&=\frac{1}{\varepsilon}\nu(t,P_t(x)),\\ \nabla_x s(t,x)&=\tau(t,P_t(x))+O(\varepsilon).\end{aligned}$$

In the new variables (t,z,s) we present for some real function b in (t,x) the ansatz

$$b_{in}(t,x)=\sum_{K=0}^{\infty}\varepsilon^K b_{in}^K(t,z(t,x),s(t,x)). \tag{40}$$

Introducing the notation $\nu(P_t(x))=\nu(t,s(t,x))$ and, similarly, $\tau(P_t(x))=\tau(t,s(t,x))$, we obtain

$$\nabla_x b_{in}(t,z(t,x),s(t,x))=\frac{1}{\varepsilon}[\partial_z b_{in}(t,z,s)]\nu(t,s)+[\partial_s b_{in}(t,z,s)]\tau(t,s)+O(\varepsilon),$$

and for some vector field $\vec{b}$ we have

$$\nabla_x\cdot\vec{b}(t,z(t,x),s(t,x))=\frac{1}{\varepsilon}(\partial_z\vec{b}(t,z,s))\cdot\nu(t,s)+(\partial_s\vec{b}(t,z,s))\cdot\tau(t,s)+O(\varepsilon).$$

Moreover, it follows that

$$\begin{aligned}\partial_t z(t,x)&=\partial_t\frac{1}{\varepsilon}d(x,\Gamma_t)=-\frac{1}{\varepsilon}v(P_t(x)),\\ \partial_t s(t,x)&=-v_\tau(P_t(x))+O(\varepsilon),\end{aligned}$$

where v is the normal velocity and v_τ the tangential velocity. We note that v_τ depends on the parametrization, whereas v is an intrinsic quantity. This leads to

$$\frac{d}{dt}b_{in}^K(t,z(t,x),s(t,x))=\partial_t b_{in}^K(t,z,s)-\frac{1}{\varepsilon}v\partial_z b_{in}^K(t,z,s)-v_\tau\partial_s b_{in}^K(t,z,s)+O(\varepsilon).$$

Now we expand ϕ and u in the variables (t,z,s) and we assume

$$\begin{aligned}&\phi_{in}^0\in\Sigma^M,\quad \phi_{in}^K\in T\Sigma^M,\quad K\geq 1,\\ &u_{in}^K\in H^N,\quad K\geq 1,\end{aligned}$$

to ensure that the constraints on ϕ and u are satisfied. Taking a Taylor expansion of L_{ij} around (u_{in}^0,ϕ_{in}^0) and writing $L_{ij}^{0,in}=L_{ij}(u_{in}^0,\phi_{in}^0)$, we obtain from the conservation laws for mass and energy to lowest order, i.e., $O(\varepsilon^{-2})$,

$$0=\frac{d}{dz}\left(\sum_{j=0}^{N}L_{ij}^{0,in}\partial_z u_{j,in}^0\right),\quad 0\leq i\leq N, \tag{41}$$

where we used that $\partial_z \nu = 0$. Integrating yields

$$L\partial_z u^0_{in} = k \tag{42}$$

for some vector $k \in \mathbf{R}^{N+1}$. Later, the matching with the outer solution will give $k = 0$.

We have

$$\partial_z \nu = 0, \quad \partial_z \tau = 0, \quad \partial_s \tau = \kappa\nu, \quad \partial_s \nu = -\kappa\tau,$$

where κ is the curvature of Γ_t. Concerning the sign of the curvature we note that for a circle of radius r whose normal is outward oriented (with our orientation the tangent is then running counterclockwise) the curvature is $-1/r$.

Hence the $O(\varepsilon^{-1})$-equations of the conserved quantities are

$$\begin{aligned} -v\partial_z c_i(u^0_{in}, \phi^0_{in}) = &-\kappa \left(\sum_{j=0}^{N} L^{0,in}_{ij} \partial_z u^0_{j,in} \right) + \frac{d}{dz} \left(\sum_{j=0}^{N} L^{0,in}_{ij} \partial_z u^1_{j,in} \right) \\ &+ \frac{d}{dz} \left(\sum_{j=0}^{N} ((L_{ij})^{0,in}_{,u} \cdot u^1_{j,in} + (L_{ij})^{0,in}_{,\phi} \cdot \phi^1_{in}) \partial_z u^0_{j,in} \right). \end{aligned} \tag{43}$$

These equations will further simplify when an expression for u^0_{in} has been derived.

Now we consider the equations for the phase field variables. As done in [17] we expand the a-terms in $(\phi^0_{in}, \partial_z \phi^0_{in} \otimes \nu)$, the w-term in ϕ^0_{in}, and the f-term in (u^0_{in}, ϕ^0_{in}). To leading order $O(\varepsilon^{-1})$ we then obtain the equation

$$0 = \frac{d}{dz} \left(P^M a_{,X}(\phi^0_{in}, \partial_z \phi^0_{in} \otimes \nu) \right) \nu - P^M a_{,\phi}(\phi^0_{in}, \partial_z \phi^0_{in} \otimes \nu) - P^M w_{,\phi}(\phi^0_{in}). \tag{44}$$

Multiplying this equation with $\partial_z \phi^0_{in} \in T\Sigma^M$ gives

$$0 = \frac{d}{dz} \left(a_{,X}(\phi^0_{in}, \partial_z \phi^0_{in} \otimes \nu) : (\partial_z \phi^0_{in} \otimes \nu) - a(\phi^0_{in}, \partial_z \phi^0_{in} \otimes \nu) - w(\phi^0_{in}) \right). \tag{45}$$

The equation of order $O(1)$ is

$$\begin{aligned} -\omega v \partial_z \phi^0_{in} = &\frac{d}{dz} \left[(P^M a_{,X})_{,\phi} \cdot \phi^1_{in} \right. \\ &\left. + (P^M a_{,X})_{,X} : (\partial_s \phi^0_{in} \otimes \tau + \partial_z \phi^1_{in} \otimes \nu) \right] \nu + \frac{d}{ds} (P^M a_{,X}) \tau \\ &- (P^M a_{,\phi})_{,\phi} \cdot \phi^1_{in} - (P^M a_{,\phi})_{,X} : (\partial_s \phi^0_{in} \otimes \tau + \partial_z \phi^1_{in} \otimes \nu) \\ &- (P^M w_{,\phi})_{,\phi} \cdot \phi^1_{in} + P^M u^0_{0,in} f_{,\phi}(T(u^0_{in}, \phi^0_{in}), c(u^0_{in}, \phi^0_{in}), \phi^0_{in}), \end{aligned} \tag{46}$$

where w and all its derivatives are evaluated in ϕ^0_{in} and a and its derivatives in $(\phi^0_{in}, \partial_z \phi^0_{in} \otimes \nu)$.

4.3. Matching and resulting jump conditions. For some quantity $b(t, x)$ we gave by (37) and (40) expansions in bulk regions, respectively, in a strip around an interface between such regions. Now we want to match these expansions in an overlap domain. We will need the matching conditions of order zero and one. For the outer expansions in Ω_1 and Ω_2 we will use the subscripts b_{out1} and b_{out2}.

We observe that near Γ_t we can express the functions $b^K_{out}(t,x)$ in the variables (t,z,s). By expanding in a Taylor series at the point $(0,s(t,x))$ which corresponds to the boundary point $\psi_t(s(t,x)) \in \Gamma_t$ (remember that $z(t,x) = \frac{1}{\varepsilon} r(t,x)$ and $\partial_r = \nu \cdot \nabla_x$), we obtain

$$\begin{aligned} b^K_{out}(t,x) &= b^K_{out}(t, r(t,x), s(t,x)) \\ &= b^K_{out}(t,0,s(t,x)) + r\partial_r(b^K_{out})(t,0,s(t,x)) + O(r^2) \\ &= b^K_{out}(t,0,s(t,x)) + \varepsilon z(\nabla_x b^K_{out}(t,0,s(t,x)) \cdot \nu(t,0,s(t,x))) + O(\varepsilon^2), \end{aligned}$$

where $b^K_{out}(t,0,s)$ and $\nabla_x b^K_{out}(t,0,s)$ mean the evaluation in $(t, P_t(x))$. We get

$$b_{out}(t,x) = b^0_{out}(t,0,s) + \varepsilon \left(z(\nabla_x b^0_{out}(t,0,s) \cdot \nu(t,s)) + b^1_{out}(t,0,s) \right) + O(\varepsilon^2).$$

Now we consider an intermediate variable $z_\varepsilon = \eta(\varepsilon) z$ for some $z > 0$, where $\eta(\varepsilon)$ is some function in ε in the overlap domain of validity of the two expansions (which we suppose to exist); i.e., $\eta = o(1)$ and $\varepsilon = o(\eta)$. Because of $z = r/\varepsilon$ we have $z_\varepsilon \to \pm\infty$ as $\varepsilon \to 0$.

We substitute the variable z in our expansions by this intermediate variable z_ε and consider their difference; the expansions of u match if, in the limit as $\varepsilon \to 0$, the terms of every order ε^K vanish. For the $O(1)$-terms this means

$$\begin{aligned} 0 &\overset{!}{=} \lim_{\varepsilon \searrow 0} \left(b^0_{out1}(t,0,s) - b^0_{in}(t,z_\varepsilon,s) \right) = \lim_{z_\varepsilon \to \infty} \left(b^0_{out1}(t,0,s) - b^0_{in}(t,z_\varepsilon,s) \right), \\ 0 &\overset{!}{=} \lim_{\varepsilon \nearrow 0} \left(b^0_{out2}(t,0,s) - b^0_{in}(t,z_\varepsilon,s) \right) = \lim_{z_\varepsilon \to -\infty} \left(b^0_{out2}(t,0,s) - b^0_{in}(t,z_\varepsilon,s) \right), \end{aligned}$$

while for the $O(\varepsilon^1)$-terms the matching condition is

$$\begin{aligned} 0 &\overset{!}{=} \lim_{z_\varepsilon \to \infty} \left(z_\varepsilon \nabla_x b^0_{out1}(t,0,s) \cdot \nu(t,s) + b^1_{out1}(t,0,s) - b^1_{in}(t,z_\varepsilon,s) \right), \\ 0 &\overset{!}{=} \lim_{z_\varepsilon \to -\infty} \left(z_\varepsilon \nabla_x b^0_{out2}(t,0,s) \cdot \nu(t,s) + b^1_{out2}(t,0,s) - b^1_{in}(t,z_\varepsilon,s) \right). \end{aligned}$$

First we apply the matching conditions on the functions $u^0_{j,in}$, $0 \le j \le N$, solving the differential equations (42). The assumption on L yields

$$\partial_z u^0_{in} = L^{-1} k.$$

By the matching conditions of order zero, u^0_{in} must be bounded if $|z| \to \infty$. Then the assumption on L necessarily gives $k = 0$ so that u^0_{in} is constant.

Since u^0_{in} is constant, we obtain that $u^0_{out1}(t,0,s) = u^0_{out2}(t,0,s)$ and hence u, and therefore the temperature and the chemical potential differences are in the sharp interface limit continuous across an interface.

Now, due to $\partial_z u^0_{j,in} = 0$, the $O(\varepsilon^{-1})$-equations (44) for the conserved variables simplify to

$$-v \partial_z c_i(u^0_{in}, \phi^0_{in}) = \frac{d}{dz} \left(\sum_{j=0}^{N} L_{ij}(u^0_{in}, \phi^0_{in}) \partial_z u^1_{j,in} \right).$$

Integrating with respect to z from $-\infty$ to ∞ (or, more correctly, integrating from $-R$ to R and then considering the limit as $R \to \infty$) and using that $v(t,s)$ is independent

of z, we obtain

$$v\left[c_i(u^0_{in},\phi^0_{in})\right]^{z\nearrow\infty}_{z\searrow-\infty} = -\left[\sum_{j=0}^{N} L_{ij}(u^0_{in},\phi^0_{in})\partial_z u^1_{j,in}\right]^{z\nearrow\infty}_{z\searrow-\infty}.$$

As has been shown in [8, 6] the matching conditions of order one for the $b^1_{j,in}$ yield

$$\partial_z b^1_{j,in} \to \nabla_x b^0_{j,out1}\cdot\nu \qquad \text{for} \qquad z\to\infty \tag{47}$$

and

$$\partial_z b^1_{j,in} \to \nabla_x b^0_{j,out2}\cdot\nu \qquad \text{for} \qquad z\to-\infty, \tag{48}$$

where the right-hand sides are evaluated in $(t,x)=(t,\psi_t(s))$ or, in the other coordinates, in $(t,r,s)=(t,0,s(t,x))$. In fact, these are the boundary values of $\nabla_x u^0_{j,out\beta}\cdot\nu$, $\beta\in\{1,2\}$, on Γ_t. After matching for the phase field variables ϕ we obtain

$$\begin{aligned}
v[c_i]^1_2 &= v\big(c_i(u^0_{out1},\phi^0_{out1}) - c_i(u^0_{out2},\phi^0_{out2})\big)(t,x)\\
&= v\left[c_i(u^0_{in},\phi^0_{in})\right]^{z\nearrow\infty}_{z\searrow-\infty}\\
&= -\Big(\sum_{j=0}^{N} L^{0,out1}_{ij}\nabla_x u^0_{j,out1} - L^{0,out2}_{ij}\nabla_x u^0_{j,out2}\Big)(t,x)\cdot\nu(t,x)\\
&= \big(J_i(u^0_{out1},\phi^0_{out1}) - J_i(u^0_{out2},\phi^0_{out2})\big)(t,x)\cdot\nu(t,x)\\
&= [J_i]^1_2\cdot\nu.
\end{aligned}$$

We will refer to this fact as the jump condition for the inner energy density $e=c_0$ and the concentrations c_i, $1\le i\le N$.

4.4. Matching and the Gibbs–Thomson relation. In the bulk regions we have $\phi^0_{out\beta}=e_\beta$, $\beta\in\{1,2\}$. Hence for each s, we have to solve equation (44) of second order in z with respect to the boundary conditions e_1 for $z\to\infty$ and e_2 for $z\to-\infty$.

By integrating (45) and using (35), (36) and $w(e_1)=w(e_2)=0$ we obtain

$$0 = a_{,X}(\phi^0_{in},\partial_z\phi^0_{in}\otimes\nu) : (\partial_z\phi^0_{in}\otimes\nu) - a(\phi^0_{in},\partial_z\phi^0_{in}\otimes\nu) - w(\phi^0_{in}).$$

Using (33) we deduce

$$a(\phi^0_{in},\partial_z\phi^0_{in}\otimes\nu) = w(\phi^0_{in}), \tag{49}$$

which is known as equipartition of energy. We set

$$\begin{aligned}
&C^{0,1}_{\alpha\beta}([-1,1],\Sigma^M) =\\
&\left\{p:[-1,1]\to\Sigma^M \mid p \text{ Lipschitz continuous, } p(-1)=e_\alpha \text{ and } p(1)=e_\beta\right\},
\end{aligned} \tag{50}$$

and define the surface entropy for some $e\in\mathbf{R}^n$ to be

$$\gamma_{\alpha\beta}(e) = \inf\left\{2\int_{-1}^{1}\sqrt{w(p)}\sqrt{a(p,p'\otimes e)}(y)dy \mid p\in C^{0,1}_{\alpha\beta}\right\}. \tag{51}$$

As shown in [40, 17], if a minimizer exists for $e = \nu(t,s)$, then a reparametrization of the minimizer fulfills (44) and, in addition,

$$\gamma_{2,1}(\nu) = \int_{-\infty}^{\infty} \left(a(\phi_{in}^0, \partial_z \phi_{in}^0 \otimes \nu) + w(\phi_{in}^0) \right) dz. \tag{52}$$

Now we want to deduce the Gibbs–Thomson law. We multiply the equation (44) for ϕ_{in}^0 by $\partial_z \phi_{in}^1 \in T\Sigma^M$ and the equation (47) for ϕ_{in}^1 by $\partial_z \phi_{in}^0 \in T\Sigma^M$. Observe that we can drop the projections P^M. Then we sum up the two equations and integrate from $-\infty$ to ∞ with respect to z. Some straightforward calculations together with the matching conditions for the boundary values yield the following solvability condition for equation (47):

$$\begin{aligned} -\omega v \int_{-\infty}^{\infty} (\partial_z \phi_{in}^0(z,s))^2 = \frac{d}{ds} \left(\int_{-\infty}^{\infty} a_{,X}(\phi_{in}^0(z,s), \partial_z \phi_{in}^0(z,s) \otimes \nu(s)) \cdot \partial_z \phi_{in}^0(z,s) dz \right) \tau(s) \\ + \int_{-\infty}^{\infty} u_{0,in}^0 f_{,\phi}(T(u_{in}^0, \phi_{in}^0) c(u_{in}^0, \phi_{in}^0), \phi) \cdot \partial_z \phi_{in}^0 dz. \end{aligned} \tag{53}$$

Using that $u_{0,in}^0$ and $\bar{u}_{in}^0 = (u_{1,in}^0, \dots, u_{N,in}^0)$ are independent of z, the last term on the RHS of (53) yields

$$\begin{aligned} &\int_{-\infty}^{\infty} u_{0,in}^0 f_{,\phi}(T_{in}^0, c_{in}^0, \phi_{in}^0) \cdot \partial_z \phi_{in}^0 dz \\ &= \int_{-\infty}^{\infty} \left(\frac{d}{dz} \left(u_{0,in}^0 f(T_{in}^0, c_{in}^0, \phi_{in}^0) \right) - u_{0,in}^0 f_{,c}(T_{in}^0, c_{in}^0, \phi_{in}^0) \cdot \partial_z c_{in}^0 \right) dz \\ &= \int_{-\infty}^{\infty} \left(\frac{d}{dz} \left(u_{0,in}^0 f(T_{in}^0, c_{in}^0, \phi_{in}^0) \right) + \bar{u}_{in}^0 \cdot \partial_z c_{in}^0 \right) dz \\ &= \left[u_{0,in}^0 f(T_{in}^0, c_{in}^0, \phi_{in}^0) + \bar{u}_{in}^0 \cdot c_{in}^0 \right]_{z \searrow -\infty}^{z \nearrow \infty} \\ &=: \left[u_0^0 \left(f(T^0, c^0, \phi^0) - f_{,c}(T^0, c^0, \phi^0) \cdot c^0 \right) \right]_2^1. \end{aligned}$$

Here we use the abbreviation $T_{in}^0 = T(u_{in}^0, \phi_{in}^0)$, $c_{in}^0 = c(u_{in}^0, \phi_{in}^0)$, $T^0 = T(u^0, \phi^0)$, and $c^0 = c(u^0, \phi^0)$. Finally, as $[c^0] \in T\Sigma^N$ we obtain

$$\int_{-\infty}^{\infty} u_{0,in}^0 f_{,\phi}(T_{in}^0, c_{in}^0, \phi_{in}^0) \cdot \partial_z \phi_{in}^0 dz = -\left(\frac{[f^0]_2^1 - \mu^0 \cdot [c^0]_2^1}{T^0} \right)(t,x).$$

Calculating the total derivative of $\gamma_{2,1}$, which becomes with (52)

$$D\gamma_{2,1}(\nu) = \int_{-\infty}^{\infty} a_{,X} \cdot \partial_z \phi_{in}^0 dz,$$

and setting

$$m(\nu) = \omega \int_{-\infty}^{\infty} (\partial_z \phi_{in}^0)^2 dz,$$

we reduce the solvability condition to (writing $\nabla_s \cdot g = (\partial_s g) \cdot \tau$ for the surface divergence of some vector field g on Γ_t)

$$m(\nu) v = -\nabla_s \cdot D\gamma_{2,1}(\nu) + \frac{[f^0]_2^1 - \mu^0 \cdot [c^0]_2^1}{T^0}.$$

Considering ν and γ as functions in an angle $\theta \in [0, 2\pi)$, i.e., setting $\nu(\theta) = (\cos(\theta), \sin(\theta))$ and $\hat{\gamma}(\theta) = \gamma(\nu(\theta))$, one can derive (see [17])

$$\nabla_s \cdot D\gamma_{2,1}(\nu) = -(\hat{\gamma}_{2,1}(\theta) + \hat{\gamma}''_{2,1}(\theta))\kappa$$

with the curvature $\kappa = -\nabla_s \cdot \nu$ which may be inserted into the solvability condition to yield

$$m(\nu)v = (\hat{\gamma}_{2,1}(\theta) + \hat{\gamma}''_{2,1}(\theta))\kappa + \frac{[f^0]_2^1 - \mu^0 \cdot [c^0]_2^1}{T^0}.$$

Finally, the force balance at triple junctions (24) can be derived as in [17]. Therefore, all equations defining the sharp interface model have been derived by asymptotic expansions.

5. Appendix. In this appendix we will show that for the sharp interface model described in section 2 the entropy does not decrease in time. We consider a situation where a bounded domain Ω is partitioned into M phases $\Omega_1(t), \dots, \Omega_M(t)$ which are separated by smooth boundaries $\Gamma_{\alpha\beta}(t) = \overline{\Omega}_\alpha \cap \overline{\Omega}_\beta \cap \Omega$. For simplicity we restrict ourselves to two space dimensions, but the calculations can also be done in higher dimensions.

Given some domain $R(t) \subset \Omega$ with smooth boundary $\partial R(t)$ and a smooth evolving curve $\Gamma(t) \subset \Omega$ with normal velocity v, we will make use of the following transport identities:

$$\frac{d}{dt}\left(\int_{\Gamma(t)} \gamma \, d\mathcal{H}^1\right)\Big|_{t=t_0} = -\int_{\Gamma(t_0)} \gamma\kappa v \, d\mathcal{H}^1 + \sum_{endpoints} \dot{p} \cdot \tau \quad \text{and}$$

$$\frac{d}{dt}\left(\int_{R(t)} u \, dx\right)\Big|_{t=t_0} = \int_{R(t_0)} \partial_t u \, dx + \int_{\partial R(t_0)} uv \, d\mathcal{H}^1(x)$$

for some smooth function $u = u(t, x)$ and some constant γ; κ is the curvature of the interface Γ, and ν is the unit normal. By $\dot{p}$ we denote the velocity of the endpoints of Γ and by τ the exterior tangent vector to $\Gamma(t)$ at the endpoints.

Let the evolution in each phase be given by

$$\partial_t e^q = -\nabla \cdot J_0^q, \quad \partial_t c_i^q = -\nabla \cdot J_i^q, \quad 1 \le i \le N, \quad 1 \le q \le M,$$

with the fluxes given in (15) and (16). We assume that the functions are smooth in their domain Ω_q and that the fluxes vanish at the external boundary of Ω. Observe that $-\partial_t c = \nabla \cdot J \in T\Sigma^N$. Then

$$\begin{aligned}
\frac{d}{dt}\left(\int_{\Omega(t)} s(e,c) \, dx\right)\Big|_{t=t_0} &= \sum_\alpha \int_{\Omega_\alpha(t_0)} \partial_t s(e,c) \, dx - \sum_{\alpha<\beta} \int_{\Gamma_{\alpha\beta}(t_0)} [s]_\alpha^\beta v \, d\mathcal{H}^1 \\
&= \sum_\alpha \int_{\Omega_\alpha(t_0)} \Big(s_{,e}\partial_t e + \sum_i s_{,c_i}\partial_t c_i\Big) \, dx - \sum_{\alpha<\beta} \int_{\Gamma_{\alpha\beta}(t_0)} [s]_\alpha^\beta v \, d\mathcal{H}^1 \\
&= -\sum_\alpha \int_{\Omega_\alpha(t_0)} \left(\frac{1}{T}\nabla \cdot J_0 + \sum_i \frac{-\bar{\mu}_i}{T}\nabla \cdot J_i\right) dx \\
&\quad - \sum_{\alpha<\beta} \int_{\Gamma_{\alpha\beta}(t_0)} [s]_\alpha^\beta v \, d\mathcal{H}^1
\end{aligned}$$

$$= \sum_\alpha \int_{\Omega_\alpha(t_0)} \nabla \frac{1}{T} \cdot J_0 + \sum_i \nabla \frac{-\bar{\mu}_i}{T} \cdot J_i \, dx$$
$$+ \sum_{\alpha<\beta} \int_{\Gamma_{\alpha\beta}(t_0)} \left(\left[\frac{1}{T} J_0 + \sum_i \frac{-\bar{\mu}_i}{T} J_i \right]_\alpha^\beta \cdot \nu - [s]_\alpha^\beta v \right) d\mathcal{H}^1.$$

The fact that L is positive semidefinite leads to

$$\nabla \frac{1}{T} \cdot J_0 + \sum_i \nabla \frac{-\bar{\mu}_i}{T} \cdot J_i \geq 0.$$

In addition, we make use of the continuity conditions (17), (18) and the jump conditions (19), (20) to obtain

$$\frac{d}{dt} \left(\int_{\Omega(t)} s(e,c)\, dx \right) \Big|_{t=t_0} \geq \sum_{\alpha<\beta} \int_{\Gamma_{\alpha\beta}(t_0)} \left(\frac{1}{T} [e]_\alpha^\beta v + \sum_i \frac{-\bar{\mu}_i}{T} [c_i]_\alpha^\beta v - \frac{[Ts]_\alpha^\beta}{T} v \right) d\mathcal{H}^1$$
$$= \sum_{\alpha<\beta} \int_{\Gamma_{\alpha\beta}(t_0)} \frac{[f]_\alpha^\beta - \sum_i \mu_i [c_i]_\alpha^\beta}{T} v \, d\mathcal{H}^1.$$

Furthermore, we have

$$\frac{d}{dt} \left(- \int_{\Gamma_{\alpha\beta}(t)} \gamma_{\alpha\beta}\, d\mathcal{H}^1 \right) \Big|_{t=t_0} = \int_{\Gamma_{\alpha\beta}(t_0)} \gamma_{\alpha\beta} \kappa v \, d\mathcal{H}^1 - \sum_{endpoints} \dot{p} \cdot \tau_{\alpha\beta} \gamma_{\alpha\beta}$$

so that we get

$$\frac{d}{dt} S \Big|_{t=t_0} = \frac{d}{dt} \left(\int_{\Omega(t)} s(e,c)\, dx - \sum_{\alpha<\beta} \int_{\Gamma_{\alpha\beta}(t)} \gamma_{\alpha\beta}\, d\mathcal{H}^1 \right) \Big|_{t=t_0}$$
$$\geq \sum_{\alpha<\beta} \int_{\Gamma_{\alpha\beta}(t_0)} \left(\frac{[f]_\alpha^\beta - \sum_i \mu_i [c_i]_\alpha^\beta}{T} + \gamma_{\alpha\beta} \kappa \right) v \, d\mathcal{H}^1$$
$$= \sum_{\alpha<\beta} \int_{\Gamma_{\alpha\beta}(t_0)} m(\nu) v^2 \, d\mathcal{H}^1 \geq 0.$$

In the last equality we used the Gibbs–Thomson relation (21), the fact that the mobility coefficient m is supposed to be positive, the force balance at triple junctions (24), and the fact that in a closed system the interfaces intersect the exterior boundary by a 90° angle condition (compare [6] and the references therein).

REFERENCES

[1] R. F. Almgren, *Second-order phase field asymptotics for unequal conductivities*, SIAM J. Appl. Math., 59 (1999), pp. 2086–2107.

[2] H. W. Alt and I. Pawlow, *A mathematical model of dynamics of non-isothermal phase separation*, Phys. D, 59 (1992), pp. 389–416.

[3] H. W. Alt and I. Pawlow, *On the entropy principle of phase transition models with a conserved order parameter*, Adv. Math. Sci. Appl., 6 (1996), pp. 291–376.

[4] J. W. Barrett, J. F. Blowey, and H. Garcke, *On fully practical finite element approximations of degenerate Cahn-Hilliard systems*, M2AN Math. Model. Numer. Anal., 35 (2002), pp. 713–748.

[5] W. J. Boettinger, J. A. Warren, C. Beckermann, and A. Karma, *Phase-field simulations of solidification*, Ann. Rev. Mater. Res., 32 (2002), p. 163–194.
[6] L. Bronsard, H. Garcke, and B. Stoth, *A multi-phase Mullins–Sekerka system: Matched asymptotic expansions and an implicit time discretization for the geometric evolution problem*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 481–506.
[7] G. Caginalp, *Stefan and Hele Shaw type models as asymptotic limits of the phase field equations*, Phys. Rev. A, 39 (1989), pp. 5887–5896.
[8] G. Caginalp and P. C. Fife, *Dynamics of layered interfaces arising from phase boundaries*, SIAM J. Appl. Math., 48 (1988), pp. 506–518.
[9] G. Caginalp and W. Xie, *An analysis of phase-field alloys and transition layers*, Arch. Ration. Mech. Anal., 142 (1998), pp. 293–329.
[10] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system* I. *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.
[11] B. Chalmers, *Principles of solidification*, Krieger, Melbourne, FL, 1977.
[12] L. Q. Chen and W. Yang, *Computer simulation of the domain dynamics of a quenched system with a large number of non-conserved order parameters: The grain–growth kinetics*, Phys. Rev. B, 50 (1994), pp. 15752–15756.
[13] L. Q. Chen, *Phase-field models for microstructural evolution*, Ann. Rev. Mater. Res., 32 (2002), p. 113–140.
[14] D. N. Fan and L. Q. Chen, *Diffuse-interface description of grain boundary motion*, Philosophical Magazine Letters, 75 (1997), pp. 187–196.
[15] E. Fried and M. E. Gurtin, *Dynamic solid-solid transitions with phase characterized by an order parameter*, Phys. D, 72 (1994), pp. 287–308.
[16] H. Garcke and B. Nestler, *A mathematical model for grain growth in thin metallic films*, Math. Models Methods Appl. Sci., 10 (2000), pp. 895–921.
[17] H. Garcke, B. Nestler, and B. Stoth, *On anisotropic order parameter models for multi–phase systems and their sharp interface limits*, Phys. D, 115 (1998), pp. 87–108.
[18] H. Garcke, B. Nestler, and B. Stoth, *A multiphase field concept: Numerical simulations of moving phase boundaries and multiple junctions*, SIAM J. Appl. Math., 60 (1999), pp. 295–315.
[19] H. Garcke, B. Nestler, and B. Stoth, *Anisotropy in multi-phase systems: A phase field approach*, Interfaces Free Bound., 1 (1999), pp. 175–198.
[20] H. Garcke and A. Novick-Cohen, *A singular limit for a system of degenerate Cahn–Hilliard equations*, Adv. Differential Equations, 5 (2000), pp. 401–434.
[21] W. L. George and J. A. Warren, *A parallel 3D dendritic growth simulator using the phase-field method*, J. Comput. Phys., 177 (2002), pp. 264–283.
[22] P. Haasen, *Physikalische Metallkunde*, 3rd ed., Springer, Berlin, 1994.
[23] A. Karma and W.-J. Rappel, *Phase-field method for computationally efficient modeling of solidification with arbitrary interface kinetics*, Phys. Rev. E, 53 (1996), pp. 3017–3020.
[24] A. Karma and W.-J. Rappel, *Quantitative phase-field modeling of dendritic growth in two and three dimensions*, Phys. Rev. E, 57 (1998), pp. 4323–4349.
[25] A. Karma, *Phase-field formulation for quantitative modeling of alloy solidification*, Phys. Rev. Lett., 87 (2001), p. 115701-1–115701-4.
[26] L. D. Landau and V. I. Ginzburg, *K teorii sverkhrovodimosti*, Zh. Eksp. Teor. Fiz, 20 (1950), pp. 1064–1082. English translation: *On the theory of superconductivity*, in Collected Papers of L.D. Landau, D. ter Haar, ed., Pergamon, Oxford, UK, 1965, pp. 626–633.
[27] J. S. Langer, *Models of pattern formation in first–order phase transitions*, in Directions in Condensed Matter Physics, World Scientific, Singapore, 1986, pp. 165–186.
[28] J. Lowengrub and L. Truskinovsky, *Quasi-incompressible Cahn-Hilliard fluids and topological transitions*, Proc. Roy. Soc. London Ser. A, 454 (1998), pp. 2617–2654.
[29] S. Luckhaus, *Solidification of Alloys and the Gibbs-Thomson Law*, Preprint 335 SFB 256, Universität Bonn, Bonn, Germany, 1994.
[30] G. B. McFadden, A. A. Wheeler, and D. M. Anderson, *Thin interface asymptotics for an energy/entropy approach to phase-field models with unequal conductivities*, Phys. D, 144 (2000), pp. 154–168.
[31] B. Nestler and A. A. Wheeler, *A multi-phase-field model of eutectic and peritectic alloys: Numerical simulation of growth structures*, Phys. D, 138 (2000), pp. 114–133.
[32] B. Nestler, A. A. Wheeler, and H. Garcke, *Modelling of microstructure formation and interface dynamics*, Comp. Mater. Sci., 26 (2003), pp. 111–119.
[33] B. Nestler, A. A. Wheeler, L. Ratke, and C. Stöcker, *Phase-field model for solidification of a monotectic alloy with convection*, Phys. D, 141 (2000), pp. 133–154.

[34] O. Penrose and P. C. Fife, *Thermodynamically consistent models of phase field type for the kinetics of phase transition*, Phys. D, 43 (1990), pp. 44–62.
[35] N. Provatas, N. Goldenfeld, and J. Dantzig, *Adaptive mesh refinement computation of solidification microstructures using dynamic data structures*, J. Comput. Phys., 148 (1999), pp. 265–290.
[36] M. Plapp and A. Karma, *Multiscale random-walk algorithm for simulating interfacial pattern formation*, Phys. Rev. Lett., 84 (2000), pp. 1740–1743.
[37] A. Karma, Y. H. Lee, and M. Plapp, *Three-dimensional dendrite-tip morphology at low undercooling*, Phys. Rev. E, 61 (2000), pp. 3996–4006.
[38] H. M. Soner, *Convergence of the phase-field equations to the Mullins-Sekerka problem with kinetic undercooling*, Arch. Ration. Mech. Anal., 131 (1995), pp. 139–197.
[39] I. Steinbach, F. Pezolla, B. Nestler, M. Seesselberg, R. Prieler, G. J. Schmitz, and J. L. L. Rezende, *A phase field concept for multi phase systems*, Phys. D, 94 (1996), pp. 35–147.
[40] P. Sternberg, *Vector–valued local minimizers of nonconvex variational problems*, Rocky Mountain J. Math., 21 (1991), pp. 799–807.
[41] B. Stoth, *A sharp interface limit of the phase field equations, one-dimensional and axisymmetric*, European J. Appl. Math., 7 (1996), pp. 603–633.
[42] J. D. van der Waals, *The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density*, in Konink. Nederl. Akad. Wetensch. Afd. Natuurk. Eerste Reeks, 1 (1893), in Dutch. English translation (with commentary): J. S. Rowlinson, J. Statist. Phys., 20 (1979), pp. 197–244.
[43] A. Visintin, *Models of Phase Transition*, Birkhäuser, Boston, 1996.
[44] S.-L. Wang, R. F. Sekerka, A. A. Wheeler, B. T. Murray, S. R. Coriell, R. J. Braun, and G. B. McFadden, *Thermodynamically-consistent phase-field models for solidification*, Phys. D, 69 (1993), pp. 189–200.
[45] A. A. Wheeler, G. B. McFadden, and W. J. Boettinger, *Phase-field model for solidification of a eutectic alloy*, Proc. Roy. Soc. London Ser. A, 452 (1996), pp. 495–525.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 800–818

TRAVELING SOLITONS IN THE DAMPED-DRIVEN NONLINEAR SCHRÖDINGER EQUATION*

I. V. BARASHENKOV† AND E. V. ZEMLYANAYA‡

Abstract. The well-known effect of the linear damping on the moving nonlinear Schrödinger soliton (even when there is a supply of energy via the spatially homogeneous driving) is to quench its momentum to zero. Surprisingly, the zero momentum does not necessarily mean zero velocity. We show that two or more parametrically driven damped solitons can form a complex traveling with zero momentum at a nonzero constant speed.

All traveling complexes we have found so far have turned out to be unstable. Thus, the parametric driving is capable of sustaining the uniform motion of damped solitons, but some additional agent is required to stabilize it.

Key words. traveling waves, nonlinear Schrödinger equation, parametric driving, dissipative solitons, bifurcations

AMS subject classifications. 35Q51, 35Q55, 74J35

DOI. 10.1137/S0036139903424837

1. Introduction.

The amplitude of a nearly harmonic wave propagating in a nonlinear dispersive medium satisfies a nonlinear Schrödinger equation. Confining ourselves to the generic, cubic nonlinearity of the "focusing" type, the resulting nonlinear Schrödinger equation is of the form

$$i\Psi_t + \Psi_{xx} + 2|\Psi|^2\Psi = -i\gamma\Psi, \quad \gamma > 0. \tag{1.1}$$

The $-i\gamma\Psi$ term in the right-hand side accounts for dissipative losses (which were assumed to be small in the derivation of (1.1)). In the underlying physical system the dissipation is compensated for by pumping the energy into the system, in one way or another. The pumping is modeled by adding a driving term to the right-hand side of (1.1).

Like a simple pendulum, the distributed system can be driven externally or parametrically. The typical form of the corresponding amplitude equation is

$$i\Psi_t + \Psi_{xx} + 2|\Psi|^2\Psi = he^{i\Omega t} - i\gamma\Psi, \tag{1.2}$$

and

$$i\Psi_t + \Psi_{xx} + 2|\Psi|^2\Psi = h\overline{\Psi}e^{2i\Omega t} - i\gamma\Psi, \tag{1.3}$$

respectively. (The overline in the right-hand side of (1.3) indicates complex conjugation.) Both the externally and parametrically driven nonlinear Schrödinger equations

*Received by the editors March 17, 2003; accepted for publication (in revised form) July 29, 2003; published electronically March 11, 2004.

http://www.siam.org/journals/siap/64-3/42483.html

†Department of Applied Mathematics, University of Cape Town, Rondebosch 7701, South Africa (igor@cenerentola.mth.uct.ac.za, igor@maths.uct.ac.za). The research of this author was supported by the NRF of South Africa under grant 2053723, by the Johnson Bequest Fund, and by the URC of the University of Cape Town.

‡Department of Applied Mathematics, University of Cape Town, Rondebosch 7701, South Africa. Permanent address: Laboratory for Information Technologies, Joint Institute for Nuclear Research, Dubna 141980, Russia (elena@jinr.ru). The research of this author was supported by the Russian Foundation for Fundamental Research under grant 0301-00657, by the Visiting Lecturer's Fund of UCT, and by NRF travel grant 2060193.

arise in a great variety of physical contexts. In particular, the parametric equation (1.3) describes the nonlinear Faraday resonance in a vertically oscillating water tank [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] and the effect of phase-sensitive amplifiers on solitons in optical fibers [11, 12, 13, 14]. The same equation controls the magnetization waves in an easy-plane ferromagnet placed in a combination of a static and microwave field [15] and the amplitude of synchronized oscillations in vertically vibrated pendula lattices [16, 17, 18, 19, 20].

Both equations (1.2) and (1.3) exhibit soliton solutions [21, 22, 23, 24, 1, 2, 15], stable and unstable [15, 22], which can also form (stable and unstable) multisoliton complexes [25, 26, 27, 28, 29]. All localized solutions that have been found so far were nonpropagating. In fact, it is widely accepted that the nonlinear Schrödinger solitons simply *cannot* travel in the presence of dissipation. This perception is based mainly on the rate equation

$$\dot{P} = -2\gamma P, \tag{1.4}$$

which is straightforward from (1.2) and (1.3). Here P is the total field momentum,

$$P = \frac{i}{2}\int_{-\infty}^{\infty} (\overline{\Psi}_x \Psi - \Psi_x \overline{\Psi}) dx. \tag{1.5}$$

In the undamped case ($\gamma = 0$) the momentum is conserved; however, if $\gamma > 0$, P decays to zero and this seems to suggest that a solitary wave, initially moving with a nonzero velocity, will have to slow down and eventually stop [30].

Another indication that only quiescent solitons are possible in the damped-driven Schrödinger equation comes ostensibly from the singular [2, 31] and inverse scattering-based perturbation theory [21, 32, 33]. Here we should mention, however, that these techniques are well developed only in the one-soliton sector and in the case of several well-separated solitons. They either make use of the smallness of the perturbation in the right-hand sides of (1.2) and (1.3) [2, 21, 32] or utilize an explicit form of the perturbed soliton (to study its stability and bifurcation) [33]. In any case, the resulting finite-dimensional system of equations for the parameters of the soliton and radiation leads to the conclusion that the soliton's velocity has to decay to zero as $t \to \infty$.

Meanwhile, the moving solitary waves could play a significant role in a variety of physical situations modeled by the damped-driven nonlinear Schrödinger equations. Stable traveling waves could compete with nonpropagating localized attractors; unstable ones might arise as long-lived transients and intermediate states in spatiotemporal chaotic regimes. Both types of moving solitary waves could mediate energy dissipation in damped-driven systems. One more reason for not rejecting the unstable solutions outright is their possible persistence within the (directly or parametrically driven) Ginzburg–Landau equations of which the Schrödinger equations (1.2) and (1.3) are special cases [35, 36, 37, 38, 39, 40, 41, 42, 43]. The diffusion and nonlinear damping (the terms $ic_1\Psi_{xx}$ and $-ic_2|\Psi|^{2n}\Psi$, to be added to the right-hand sides of (1.2) and (1.3)) are known to have a stabilizing effect on the Ginzburg–Landau pulses; hence the unstable Schrödinger solitons may gain stability as they are continued to nonzero positive c_1 and c_2.

The purpose of this paper is to show that the damped-driven nonlinear Schrödinger equations do support solitary waves traveling with nonzero velocities. For the demonstration of this fact we confine our study to the *parametrically* driven Schrödinger only. The *externally* driven equation is left as an object of future research.

Two complementary strategies will be pursued to achieve our goal. First, in section 3, we consider the *motionless damped* solitons ($V = 0$, $\gamma \neq 0$) and derive the condition under which they can be continued to nonzero velocity. Having identified values of γ for which this condition is satisfied, we perform the numerical continuation obtaining a branch of solitary waves with nonzero V and γ. Our second approach is presented in section 4; the idea is to continue *undamped traveling* waves ($\gamma = 0$, $V \neq 0$) to nonzero dampings. We show that this is possible only if the traveling wave has zero momentum. For complexes with $P = 0$, we then carry out the numerical continuation in γ. Finally, in section 5 we discuss the consistency of results obtained within these two complementary approaches.

We examined, numerically, the stability of all solutions obtained within both approaches. The general framework of the stability analysis is outlined in section 2. Results of this analysis are presented along with results of the numerical continuation. Section 6 summarizes the conclusions of our study.

2. Mathematical preliminaries. For purposes of this paper we transform equation (1.3) to an autonomous form. First, we normalize the driving frequency Ω to unity; after that, the substitution $\Psi(x,t) = e^{it}\psi(x,t)$ takes (1.3) to

$$i\psi_t + \psi_{xx} + 2|\psi|^2\psi - \psi = h\overline{\psi} - i\gamma\psi. \tag{2.1}$$

This is the representation of the parametrically driven damped nonlinear Schrödinger equation that we are going to work with in this paper. We confine ourselves to uniformly traveling solutions of the form

$$\psi(x,t) = \psi(x - Vt) \equiv \psi(\xi), \tag{2.2}$$

where $\psi(\xi) \to 0$ as $|\xi| \to \infty$. These satisfy an ordinary differential equation,

$$-iV\psi_\xi + \psi_{\xi\xi} + 2|\psi|^2\psi - \psi = h\overline{\psi} - i\gamma\psi. \tag{2.3}$$

The analytical part of this paper deals mainly with identifying those of the previously found solutions of (2.3) with $V = 0$ or $\gamma = 0$ which can be continued in V and γ, respectively. The actual continuation will be carried out numerically. Our numerical method employs a predictor-corrector continuation algorithm with a fourth-order accurate Newtonian solver. Typically, the infinite line was approximated by an interval $(-100, 100)$. The discretization step size was typically 0.005. The numerical residual was set to be 10^{-10}; that is, the grid solution would be deemed accurate if the difference between the left- and right-hand sides in (2.3) were smaller than 10^{-10}.

Along with the continuation of solutions in V and γ, we will be analyzing their stability to small perturbations. To this end, we linearize (2.1) in the comoving frame of reference. Letting $\psi(x,t) = u(\xi) + iv(\xi) + \delta\psi(\xi,t)$, where u and v are the real and imaginary parts of the solution that we are linearizing about, and assuming that the linear perturbation depends on time exponentially,

$$\delta\psi(\xi,t) = e^{\lambda t}\left[\delta u(\xi) + i\delta v(\xi)\right],$$

we arrive at an eigenvalue problem

$$\mathcal{H}_0\,\vec{y} = (\lambda + \gamma)J\,\vec{y}, \tag{2.4}$$

where the operator $\mathcal{H}_0$ is defined by

$$\mathcal{H}_0 = \begin{pmatrix} -\partial_\xi^2 + 1 + h - 6u^2 - 2v^2 & -V\partial_\xi - 4uv \\ V\partial_\xi - 4uv & -\partial_\xi^2 + 1 - h - 6v^2 - 2u^2 \end{pmatrix}, \tag{2.5}$$

the skew-symmetric matrix J is

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

and the column vector $\vec{y}(\xi) = (\delta u, \delta v)^T$. The eigenvalue problem (2.4) was solved by expanding δu and δv over a Fourier basis, typically with 500 modes, on the interval $(-50, 50)$.

The last point that we need to touch upon in this preliminary section is the integrals of motion of (1.3), or, more precisely, the quantities which are conserved in the absence of dissipation. When $\gamma = 0$, (1.3) conserves the momentum (given by (1.5) where one only needs to replace $\Psi \to \psi$) and energy,

$$E = \int_{-\infty}^{\infty} (|\psi_x|^2 + |\psi|^2 - |\psi|^4 + h \ \mathrm{Re}\, \psi^2)\, dx. \tag{2.6}$$

In the damped case, the momentum decays according to the rate equation (1.4) while the energy satisfies

$$\dot{E} = 2\gamma \left(\int_{-\infty}^{\infty} |\psi|^4 dx - E \right). \tag{2.7}$$

3. Continuation of damped solitons to nonzero velocities.

3.1. Continuability criterion. Our first strategy is to attempt to continue stationary solutions with nonzero γ to nonzero V. Two basic soliton solutions, denoted ψ_+ and ψ_-, are available explicitly:

$$\begin{aligned} \psi_\pm(x) &= e^{-i\theta_\pm} A_\pm \,\mathrm{sech}\,(A_\pm x)\,, \\ A_\pm &= \sqrt{1 \pm \sqrt{h^2 - \gamma^2}}, \\ \theta_+ = \frac{1}{2} \arcsin \frac{\gamma}{h}, &\quad \theta_- = \frac{\pi}{2} - \theta_+. \end{aligned} \tag{3.1}$$

The two solitons can form a variety of stationary complexes. These are denoted, symbolically, by $\psi_{(++)}$, $\psi_{(--)}$, $\psi_{(+-+)}$, $\psi_{(-+-)}$, and so on [29]. Let $\psi_0(x)$ be a particular complex; we want to find out whether it can be continued in V. Assuming there is a solution $\psi(\xi; V)$ such that $\psi(\xi; 0) \equiv \psi_0(\xi)$ $(= \psi_0(x))$, we expand $\psi(\xi; V)$ in powers of V as

$$\psi(\xi; V) = e^{-i\theta} \big\{ u_0(\xi) + iv_0(\xi) + V[u_1(\xi) + iv_1(\xi)] + V^2[u_2(\xi) + iv_2(\xi)] + \cdots \big\}, \tag{3.2}$$

where the constant phase θ will be chosen at a later stage. We also expand h and γ: $h = h_0 + h_1 V + \cdots$, $\gamma = \gamma_0 + \gamma_1 V + \cdots$. Substituting into (2.3), the order V^1 gives

$$\mathcal{L} \begin{pmatrix} u_1 \\ v_1 \end{pmatrix} = \begin{pmatrix} v_0' \\ -u_0' \end{pmatrix} + \mathcal{B} \begin{pmatrix} u_0 \\ v_0 \end{pmatrix}, \tag{3.3}$$

where the operator $\mathcal{L}$ has the form

$$\mathcal{L} = \begin{pmatrix} -\partial_x^2 + 1 + h_0 \cos 2\theta - 6u_0^2 - 2v_0^2 & \gamma_0 + h_0 \sin 2\theta - 4u_0 v_0 \\ -\gamma_0 + h_0 \sin 2\theta - 4u_0 v_0 & -\partial_x^2 + 1 - h_0 \cos 2\theta - 2u_0^2 - 6v_0^2 \end{pmatrix}, \tag{3.4}$$

the constant matrix $\mathcal{B}$ is given by

$$\mathcal{B} = \begin{pmatrix} -h_1 \cos 2\theta & -\gamma_1 - h_1 \sin 2\theta \\ \gamma_1 - h_1 \sin 2\theta & h_1 \cos 2\theta \end{pmatrix},$$

and the primes over u_0 and v_0 indicate derivatives with respect to x. (In (3.3) and (3.4) we have replaced ξ with x as ξ coincides with x for $V = 0$.) According to Fredholm's alternative, (3.3) has a bounded solution $u_1(x), v_1(x)$ if and only if the vector in the right-hand side is orthogonal to the kernel of the Hermitian-conjugate operator $\mathcal{L}^\dagger$:

$$\int (y, w) \begin{pmatrix} v_0' \\ -u_0' \end{pmatrix} dx + \int (y, w)\, \mathcal{B} \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} dx = 0. \tag{3.5}$$

Here $\vec{y}(x) = (y, w)^T$ is the eigenvector of $\mathcal{L}^\dagger$ associated with the zero eigenvalue: $\mathcal{L}^\dagger \vec{y} = 0$. That the operator $\mathcal{L}^\dagger$ has a zero eigenvalue follows from the fact that the operator $\mathcal{L}$ has one—namely, the translation eigenvalue corresponding to the eigenvector $(u_0', v_0')^T$. Equation (3.5) gives a necessary continuability condition of damped quiescent solitons to nonzero velocities.

3.2. Noncontinuability of the "building blocks." It is quite easy to check that when $\gamma_0 \neq 0$, the individual ψ_+ and ψ_- solitons (the basic "building blocks" of which all complexes are constructed) are *not* continuable to nonzero V. Choosing $\theta = \theta_+$ for the ψ_+ soliton and $\theta = \theta_-$ for the ψ_- (where $\theta_\pm$ are to be computed from the bottom formula in (3.1) with $\gamma = \gamma_0$ and $h = h_0$), we get $v_0(x) = 0$, $\gamma_0 - h_0 \sin 2\theta = 0$ and so the 2×2 matrix $\mathcal{L}$, (3.4), becomes upper triangular. The zero mode of $\mathcal{L}^\dagger$ can now be readily found.

Consider, for instance, the ψ^+ case. The zero mode satisfies

$$\begin{pmatrix} -\partial_x^2 + A_+^2 - 6u_0^2 & 0 \\ 2\gamma_0 & -\partial_x^2 + A_-^2 - 2u_0^2 \end{pmatrix} \begin{pmatrix} y \\ w \end{pmatrix} = 0,$$

and hence $y(x) = u_0'(x)$, and $w(x)$ is found from

$$(-\partial_x^2 + A_-^2 - 2u_0^2) w = -2\gamma u_0'(x). \tag{3.6}$$

Using the explicit expression for $u_0(x)$, $u_0(x) = A_+ \mathrm{sech}(A_+ x)$, the operator in the left-hand side of (3.6) can be written as $A_+^2(L_0 - \epsilon)$, where $\epsilon = 2h_0 \cos(2\theta_+)/A_+^2$, L_0 is given by

$$L_0 = -\partial_X^2 + 1 - 2\mathrm{sech}^2 X,$$

and $X = A_+ x$. The operator L_0 has familiar spectral properties; in particular it has a single discrete eigenvalue $E_0 = 0$ associated with an even eigenfunction $z_0 = \mathrm{sech} X$, while its continuous spectrum occupies the semiaxis $E_k \geq 1$. Consequently, for $0 < \epsilon < 1$ (that is, for $h_0 < \sqrt{1 + \gamma_0^2}$), the operator $L_0 - \epsilon$ is invertible and a bounded solution $w(x)$ of (3.6) exists and is unique. It can be found explicitly, but this is not really necessary for our purposes. All we need to know is that, since L_0 is a parity-preserving operator, $w(x)$ has the same parity as the right-hand side in (3.6), i.e., it is an odd function. For that reason the second integral in (3.5) vanishes and the necessary continuability condition reduces to

$$\gamma \int u_0'(x)(L_0 - \epsilon)^{-1} u_0'(x) dx = 0. \tag{3.7}$$

This quadratic form can be easily evaluated by expanding $u_0'(x)$ over eigenfunctions of the operator L_0:

$$u_0'(x) = \int_{-\infty}^{\infty} U(k) z_k(X) dk,$$

where $L_0 z_k(X) = (1+k^2) z_k(X)$. (The "discrete" eigenfunction $z_0(X)$ does not appear in the expansion as it has the opposite parity to $u_0'(x)$.) Utilizing the orthonormality of the eigenfunctions, the continuability condition (3.7) is transformed into

$$\gamma \int \frac{|U(k)|^2}{k^2 + 1 - \epsilon} dk = 0. \tag{3.8}$$

As $\epsilon < 1$, this condition obviously cannot be satisfied (unless $\gamma = 0$).

In the case of the ψ_- soliton, the analysis is similar. In this case, the continuability condition (3.8) is replaced by

$$\gamma \int \frac{|U(k)|^2}{k^2 + (1-\epsilon)^{-1}} dk = 0,$$

and this cannot be met for the same reason as for (3.8).

3.3. Continuation of the complexes. Turning to the *complexes* of the solitons ψ_+ and ψ_-, the phase of the complex varies with x and therefore the matrix $\mathcal{L}$ cannot be made triangular no matter how we choose the constant θ in (3.2). For this reason, aggravated by the fact that the multisoliton solutions are not available explicitly, the continuability condition (3.5) cannot be verified analytically. Resorting to the help of a computer, we evaluated the eigenfunction $\vec{y}(x)$ associated with the zero eigenvalue of the operator $\mathcal{L}^\dagger$ numerically. (Here we set $\theta = \theta_+$.)

All dissipative soliton complexes found in [29] were symmetric; that is, the corresponding u and v are *even* functions of x. Therefore, the operator $\mathcal{L}^\dagger$, whose potential part is made up of $u(x)$ and $v(x)$, is parity preserving and all its eigenfunctions pertaining to nonrepeated eigenvalues are either even or odd. As we move along a continuous branch of solutions, the parity of the eigenfunction has to change continuously. Since the parity equals either $+1$ (for even functions) or -1 (for odd functions), the only opportunity left to it by the continuity argument is to remain constant on the entire branch. For that reason it is sufficient to determine the parity of the eigenfunction for one specific value of h and then we will know it at all other points. Our numerical calculation shows that the eigenfunction $\vec{y}(x)$ is *odd* on all branches reported in [29]. Consequently, the second term in (3.5) is always zero and we only need to evaluate the first term.

The vanishing of the term involving coefficients h_1 and γ_1 in (3.5) implies that it was not really necessary to expand h and γ in powers of V. This fact has a simple geometric interpretation. As we will see below, for the fixed γ the continuable solutions occur only at isolated values of h and hence they exist only for h and γ lying on continuous curves in the (h, γ)-plane. Each curve results from an intersection of some surface in the three-dimensional (h, γ, V)-space with the $(V = 0)$-plane. The fact that one does not have to alter h and γ when continuing the solution to nonzero V indicates that these surfaces are orthogonal to the $(V = 0)$-plane along their curves of intersection.

Having found the solution $\psi(x) = u(x) + iv(x)$ at representative points along each branch, we obtained the eigenfunction $\vec{y}(x)$ at these points and evaluated what

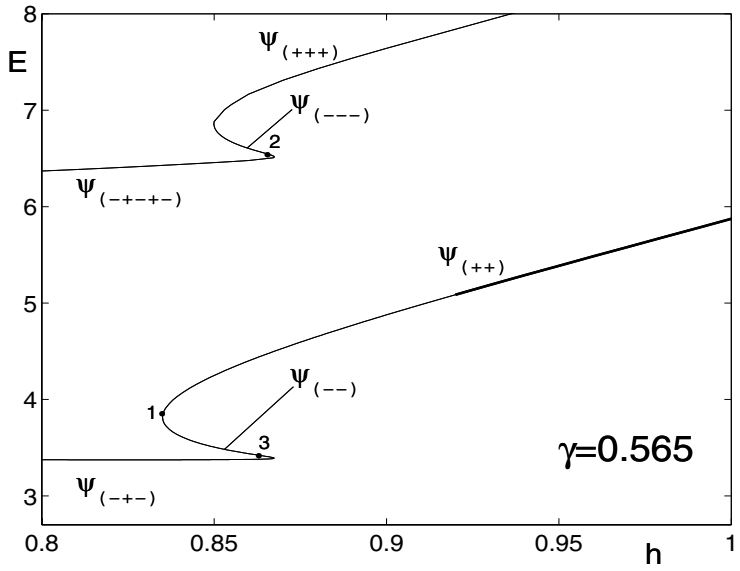


FIG. 3.1. *A fragment of the bifurcation diagram for stationary multisoliton complexes (adapted from* [29]*.) Shown is the energy* (2.6) *of the complex as a function of* h*. The bottom branch pertains to symmetric two-soliton complexes* $\psi_{(++)}$ *and* $\psi_{(--)}$ *and a three-soliton solution* $\psi_{(-+-)}$*; the top branch includes the three-soliton states* $\psi_{(+++)}$ *and* $\psi_{(---)}$*, as well as a five-soliton solution* $\psi_{(-+-+-)}$*. The thick curve corresponds to stable and thin curves to unstable solutions, respectively. The black dots indicate points where the integral* (3.9) *equals zero and therefore moving solitons are allowed to bifurcate off.*

remains of the integral (3.5):

$$\int (yv_0' - wu_0')\,dx \equiv I(h). \tag{3.9}$$

The integral I is a continuous function of h, and it was not difficult to find points on the curve at which it changes from positive to negative values, or vice versa.

We examined two branches of multisoliton solutions obtained previously [29] (Figure 3.1). The integral $I(h)$ was found to change its sign at three points, marked by black dots in the figure. (Although it may seem from the figure that I equals zero right at the turning points, in actual fact zeros of I do not *exactly* coincide with the turning points.) We were indeed able to numerically continue our solutions in V from each of these three points. Results are presented in Figure 3.2, (a)–(c).

The point "1" in Figure 3.1 corresponds to the stationary complex $\psi_{(++)}$ and lies just above the turning point where the $\psi_{(++)}$ turns into $\psi_{(--)}$. (The turning point has $h = 0.83504217$ while $I(h) = 0$ for $h = 0.8353$.) This solution has four positive real eigenvalues in the spectrum of the associated linearized operator and hence is unstable. The $\gamma(V)$ curve which results from the continuation of this solution in V is shown in Figure 3.2(a). As V grows from zero, the solution loses its even symmetry (see the inset to Figure 3.2(a)) while the four positive eigenvalues collide, pairwise, and become two complex conjugate pairs with positive real parts. After reaching a maximum velocity of approximately 0.65, the curve turns back toward $V = 0$, with γ first growing but then also turning toward $\gamma = 0$. The solution transforms into a (strongly overlapped) $\psi_{(+-)}$ complex. As V and γ tend to their zero values,

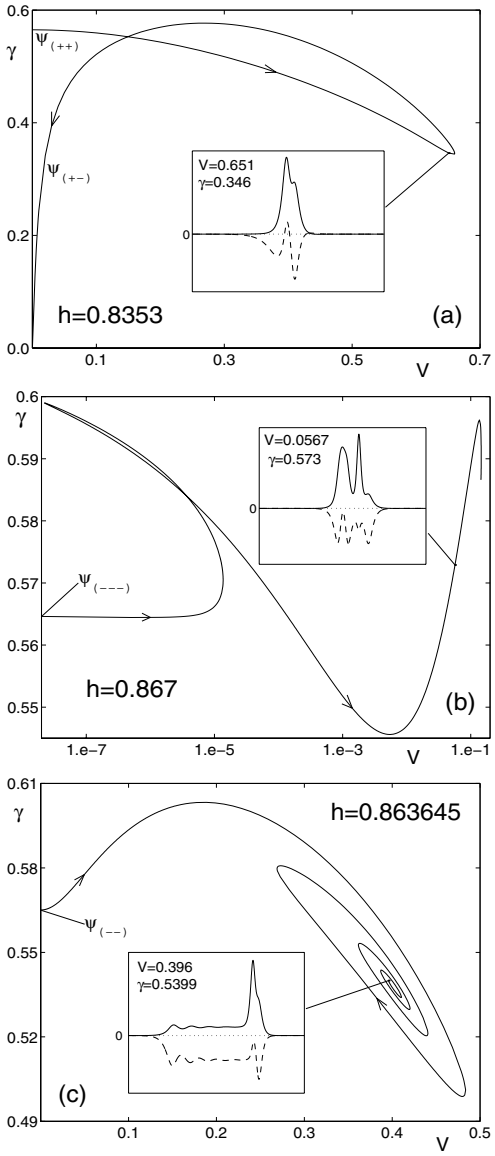


FIG. 3.2. *Bifurcation curves branching off the points marked by black dots in Figure* 3.1. *The curves illustrate the relation between the value of the damping γ and velocity V at which the wave may travel for that γ. Each curve begins at the point $\gamma = 0.565$ on the vertical axis. The insets show representative solutions at internal points of each branch. (Solid line: real part; dashed line: imaginary part.) Note the logarithmic scale of V in* (b). *Here, and in all other diagrams, arrows indicate our direction of continuation.*

the separation between the ψ_+ and ψ_- constituent solitons in the complex grows to infinity. The spectrum becomes the union of the eigenvalues of the individual ψ_+ and ψ_- solitons; in particular, it includes a complex-conjugate pair with a positive real part, and a positive real eigenvalue. Thus the entire branch shown in Figure 3.2(a) is unstable.

One more comment that we need to make here concerns the validity of the

continuation scenario presented in Figure 3.2(a) for other values of h. Note that if we chose a smaller value of γ in Figure 3.1, the value of h corresponding to the point "1" would also be smaller. (For example, for $\gamma = 0.548$ the integral $I(h)$ vanishes at the point $h = 0.82$.) For this smaller h the final product of the continuation turns out to be not a pair of infinitely separated stationary ψ_+ and ψ_- but a totally different complex. This is discussed below in section 5; see also Figures 4.2(c) and 5.1.

Another branch bifurcates off at the point marked "2" in Figure 3.1. Here $h = 0.867$. The corresponding $\gamma(V)$ diagram is displayed in Figure 3.2(b). As we move along the branch departing from $V = 0$, the original stationary complex $\psi_{(---)}$ transforms into a solution displaying three widely separated peaks in its real part: one corresponding to a strongly overlapping complex $\psi_{(-+-)}$, the next one to the ψ_+, and the last one to the ψ_- soliton. After passing a turning point, the curve is reapproaching, tangentially, the $(V = 0)$-axis. However, having reached $V = 2.2 \times 10^{-8}$, it suddenly turns back and the velocity starts to grow again. The separation between the solitons decreases and the solution can now be interpreted as a strongly overlapping four-soliton complex $\psi_{(+++-)}$ (shown in the inset to Figure 3.2(b)). As we continue further, the four constituent solitons regroup into two complexes, $\psi_{(++)}$ and $\psi_{(+-)}$. The distance between the two complexes grows rapidly, and, for certain finite V and γ (at the endpoint of the curve in Figure 3.2(b)), becomes infinite. At this point we have two coexisting solutions, $\psi_{(++)}$ and $\psi_{(+-)}$, and so this point corresponds to the point of self-intersection of the curve shown in Figure 3.2(a). Continuing the two solutions separately, from the endpoint of the curve in Figure 3.2(b), we reproduce the diagram of Figure 3.2(a) for a slightly different value of h (i.e., for $h = 0.867$).

The entire branch shown in Figure 3.2(b) is unstable. The start-off stationary solution $\psi_{(---)}$ has three positive real eigenvalues in its spectrum; one of these persists for all V and γ while the other two collide and form a complex-conjugate pair with a positive real part.

The branch continuing from the point "3" in Figure 3.1, for which $h = 0.863645$, leads to the least expected solutions. The resulting $\gamma(V)$ curve is shown in Figure 3.2(c). For points lying on the "spiral" part of the curve, the function $\psi(x)$ is equal to a constant in a relatively large but finite region, and is zero outside that region. (See the inset to Figure 3.2(c).) The constant is $\psi^{(0)} = (A_-/\sqrt{2})e^{-i\theta_-}$; it defines a stationary spatially uniform solution to (2.1). (This flat background is unstable with respect to the continuous spectrum perturbations. Figuratively speaking, our pulse solution $\psi(x)$ represents a "droplet" of the unstable phase in the stable one.) On one side (at the rear of the pulse) the zero background is connected to the background $\psi^{(0)}$ by a kink-like interface. In the front of the pulse, the interface has the character of a large-amplitude excitation, with the shape resembling the $\psi_{(+-)}$ complex. As the curve $\gamma(V)$ spirals toward its "focus" in Figure 3.2(c), the length of the region where $\psi(x) = \psi^{(0)}$ is growing. The entire branch is unstable; the start-off $\psi_{(--)}$ solution already has two real positive eigenvalues in its spectrum and more appear as we move along the branch. Those additional positive eigenvalues are remnants of the "unstable" interval of the continuous spectrum of the flat nonzero solution $\psi^{(0)}$.

4. Continuation of traveling waves to nonzero dampings.

4.1. Continuability conditions. When $\gamma = 0$, equation (2.3) has a plethora of localized solutions with nonzero V [34], and our second strategy will be to attempt to continue these nondissipative traveling waves to nonzero γ. We start with establishing the necessary and sufficient conditions for such a continuation.

A set of the *necessary* conditions can be derived easily using two integral charac-

teristics of (2.1), the momentum,

$$P = (i/2)\int(\overline{\psi}_x\psi - \psi_x\overline{\psi})dx, \tag{4.1}$$

and energy (2.6). No matter whether γ equals zero or not, the uniformly traveling solitary waves (i.e., solutions of the form (2.2)) satisfy $\dot{P} = \dot{E} = 0$. Using these relations in (1.4) and (2.7) with $\gamma \neq 0$, we get

$$P = 0 \tag{4.2}$$

and

$$E = \int |\psi|^4 dx. \tag{4.3}$$

Equations (4.2) and (4.3) have to be satisfied by the undamped solutions continuable to nonzero γ.

In fact, (4.2) and (4.3) are not independent. Indeed, multiplying (2.3) by $\overline{\psi}$, adding its complex conjugate, and integrating, gives an identity

$$E - \int |\psi|^4 dx = VP. \tag{4.4}$$

Letting $P = 0$ in (4.4), equation (4.3) immediately follows. Thus we can keep $P = 0$ as the *only* necessary condition for the continuability to nonzero γ; (4.3) is satisfied as soon as (4.2) is in place.

It turns out that $P = 0$ is also a *sufficient* condition. To show this, we expand the field $\psi = u + iv$ in powers of γ:

$$u = u_0 + \gamma u_1 + \gamma^2 u_2 + \cdots, \quad v = v_0 + \gamma v_1 + \gamma^2 v_2 + \cdots,$$

substitute into (2.3), and equate coefficients of like powers. (We also could have expanded h and V in γ, but, similarly to the continuation in V described in the previous section, the terms with coefficients h_1 and V_1 cancel out of the resulting continuability condition.) At the order $\mathcal{O}(\gamma^1)$, we obtain

$$\mathcal{H}_0 \begin{pmatrix} u_1 \\ v_1 \end{pmatrix} = \begin{pmatrix} -v_0 \\ u_0 \end{pmatrix}. \tag{4.5}$$

Here the Hermitian operator $\mathcal{H}_0$ is as in (2.5) where we only need to attach zero subscripts to u and v:

$$\mathcal{H}_0 = \begin{pmatrix} -\partial_\xi^2 + 1 + h - 6u_0^2 - 2v_0^2 & -V\partial_\xi - 4u_0v_0 \\ V\partial_\xi - 4u_0v_0 & -\partial_\xi^2 + 1 - h - 2u_0^2 - 6v_0^2 \end{pmatrix}. \tag{4.6}$$

Since the operator $\mathcal{H}_0$ has a zero eigenvalue, with the translation mode as an associated eigenvector, (4.5) is solvable only if its right-hand side is orthogonal to (u_0', v_0'):

$$\int (u_0', v_0') \begin{pmatrix} -v_0 \\ u_0 \end{pmatrix} d\xi \quad = 0. \tag{4.7}$$

(Here the prime indicates the derivative with respect to ξ.) The expression in the left-hand side of (4.7) coincides with (4.1) written in terms of the real and imaginary parts of ψ, and so the solvability condition (4.7) is simply $P = 0$.

Now assume that P is equal to zero so that a bounded solution to (4.5) exists. All traveling waves found in [34] have even real and odd imaginary parts: $u_0(-x) = u_0(x)$, $v_0(-x) = -v_0(x)$. Noticing that the diagonal elements of the operator $\mathcal{H}_0$ are parity-preserving while the off-diagonal elements change their sign under the $\xi \to -\xi$ reflection, we conclude that $u_1(x)$ is odd and $v_1(x)$ is even.

Proceeding to the order $\mathcal{O}(\gamma^2)$, we have

$$\mathcal{H}_0 \begin{pmatrix} u_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} -v_1 + u_0[6u_1^2 + 2v_1^2] + 4v_0u_1v_1 \\ u_1 + v_0[2u_1^2 + 6v_1^2] + 4u_0u_1v_1 \end{pmatrix}. \tag{4.8}$$

The top entry in the right-hand side of (4.8) is even and the bottom one odd; hence the right-hand side is orthogonal to the null vector (u_0', v_0') and a bounded solution $u_2(\xi), v_2(\xi)$ exists. This time the u-component is even and the v-component odd: $u_2(-\xi) = u_2(\xi)$, $v_2(-\xi) = -v_2(\xi)$.

It is not difficult to verify that this parity alternation property guarantees the boundedness of $u_n(\xi)$ and $v_n(\xi)$ for all n. Therefore, (2.3) has a localized solution ($\psi(\xi) \to 0$ as $|\xi| \to \infty$) for sufficiently small γ. Thus if we have an undamped soliton traveling with zero momentum, it can be continued to nonzero values of γ.

4.2. Continuable solutions: The bifurcation diagram of the undamped nonlinear Schrödinger. In this subsection we review the $P(V)$ law for the undamped solitons and solitonic complexes [34]. Of interest, of course, are points where the graph crosses the V-axis, i.e., where $P(V) = 0$.

The simplest solutions arising for $V = 0$ are, obviously, our stationary fundamental solitons ψ_+ and ψ_-. These are given by (3.1), where one only needs to set $\gamma = 0$:

$$\psi_+(x) = A_+ \mathrm{sech}\,(A_+x), \quad \psi_-(x) = iA_- \mathrm{sech}\,(A_-x),$$

with $A_\pm^2 = 1 \pm h$. Both ψ_+ and ψ_- have zero momenta and therefore are continuable to nonzero γ. However, the continuation does not produce any traveling waves in this case; all we get is our static damped solitons $\psi_\pm$, equation (3.1).

Next, both ψ_+ and ψ_- admit continuation to nonzero V (for the fixed $\gamma = 0$) [34]. As V is increased to

$$c = \sqrt{2 + 2\sqrt{1 - h^2}},$$

the width of the soliton ψ_- increases, its amplitude decreases, and the soliton gradually transforms into the trivial solution, $\psi \equiv 0$. On the resulting branch, the momentum vanishes only for $V = 0$ and $V = c$ and therefore, no dissipative branches can bifurcate off the traveling ψ_- soliton.

We now turn to the soliton ψ_+. When $h < 0.28$, its fate is similar to that of the ψ_-: as $V \to c$, the soliton spreads out and merges with the zero solution. The momentum equals zero at only two points, $V = 0$ and $V = c$; for $0 < V < c$, the momentum is positive.

For $h > 0.28$, the transformation of the ψ_+ is more promising from the present viewpoint (see the dashed curve in Figure 4.1). As V is increased from zero, the momentum grows, then the branch turns back toward the $V = 0$ axis. For some $V < 0$ the momentum reaches its maximum and then decreases to zero. The point $V = V_1$ where $P(V_1) = 0$ is of interest to us, as a branch of damped solitons can bifurcate off at this point (and it really does; see subsection 4.4). Continuing beyond

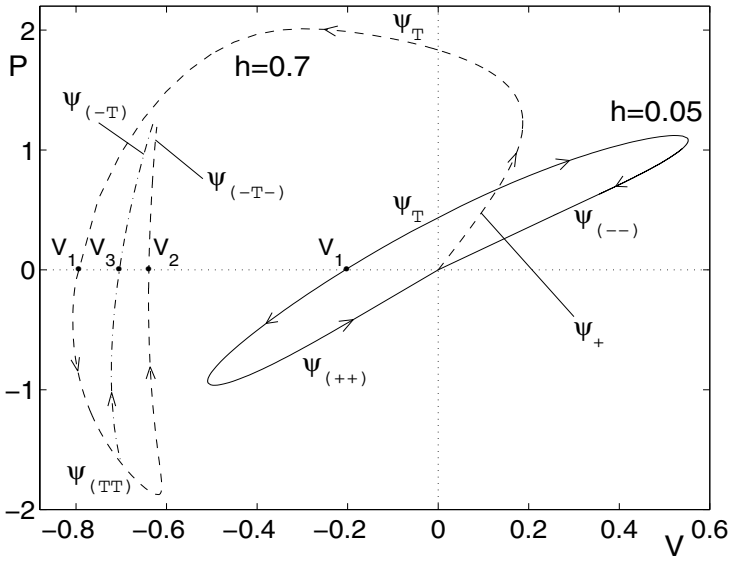


FIG. 4.1. *The momentum of the undamped traveling wave as a function of its velocity (a combined and advanced version of two diagrams from* [34]*). The dashed and dash-dotted curves pertain to the case of large driving strengths (here exemplified by* $h = 0.7$*). The starting point* $P = V = 0$ *of the dashed curve corresponds to the stationary undamped* ψ_+ *soliton, which then transforms to the twist, then to a bound state of two twists, and then to a complex of a twist and two* ψ_- *solitons. (This curve appeared in* [34]*.) The dash-dotted offshoot is our new contribution to the diagram; it corresponds to an asymmetric solution,* $\psi_{(-T)}$*, detaching from the* $\psi_{(TT)}$ *curve. The solid curve pertains to the case of small driving amplitudes (here* $h = 0.05$*). (This curve also appeared in* [34]*.) The points of its intersection with the* P*-axis correspond to stationary twist solitons; continuing each of these counterclockwise gives rise to a bound state of two* ψ_+*'s, while when continued clockwise each twist transforms into a complex of two* ψ_-*'s. More solution curves can be generated by the mapping* $V \to -V$*,* $P \to -P$*.*

V_1, the curve $P(V)$ turns toward $V = 0$ and then, after one more turning point, we have another zero crossing: $P(V_2) = 0$. This is how far we have managed to advance in our previous work [34].

At this point we need to mention that the ψ_+ and ψ_- are not the only quiescent solitons for $\gamma = 0$. The dashed $P(V)$ curve in Figure 4.1 is seen to have one more intersection with the P-axis, apart from the one at the origin. The corresponding solution represents a symmetric strongly overlapping complex of the ψ_+ and ψ_- solitons and was coined "twist" (symbolically ψ_T) in [34]. The twist soliton arises both for h greater and smaller than 0.28. In the former region the twist obtains from the V-continuation of the ψ_+ soliton, while for $h < 0.28$, it is not connected to the ψ_+. (See the solid curve in Figure 4.1.) The continuation of the twist in V in the case $h < 0.28$ gives rise to a new branch of the undamped solutions which has a point of intersection with the $(P = 0)$-axis, at some $V = V_1$. A dissipative traveling wave is bifurcating off at this value of velocity; see the next subsection. We are using the same notation, V_1, in the small- and large-h cases in Figure 4.1 to emphasize the similarity of the resulting $\gamma(V)$ curves in the two cases; see below.

Returning to the case of large h, the entire dashed curve in Figure 4.1 corresponds to symmetric solutions, $\psi(-\xi) = \overline{\psi}(\xi)$. It turns out that there are also nonsymmetric solutions; these were missed in [34]. The real part of a nonsymmetric solution is not

even, and the imaginary part is not odd. In particular, a pair of asymmetric solutions arise in a pitchfork bifurcation of the complex $\psi_{(TT)}$; see the dash-dotted offshoot from the dashed curve in Figure 4.1. (The two asymmetric solutions are related by the transformation $\psi(\xi) \to \overline{\psi}(-\xi)$; they obviously have equal momenta and hence are represented by the same curve.) Continuing the asymmetric branch we have the third zero crossing, at $V = V_3$. When continued to positive P, the asymmetric solution acquires the form of a complex of ψ_- and ψ_T solitons, with the intersoliton separation growing as P is increased. (Note that although the dashed and dash-dotted curves end at nearby points, they are *not* connected.) Our numerical analysis shows that branches of damped solitons do indeed detach at V_1, V_2, and V_3; these will be described in the next two subsections.

4.3. Numerical continuation: Small driving amplitudes. For small h, $h < 0.28$, our continuation departs from the twist soliton moving with the velocity V_1 (the point of intersection of the solid curve with the horizontal axis in Figure 4.1). The real part of this solution is even and the imaginary part odd: $\psi(-x) = \overline{\psi}(x)$. As we continue to nonzero γ, this symmetry is lost; a typical profile at the internal points looks like a nonsymmetric complex of the ψ_- and ψ_+ and is displayed in the inset to Figure 4.2(a). The rest of Figure 4.2(a) shows the resulting $\gamma(V)$ relation. As γ grows, the negative velocity of the traveling wave decreases in modulus. However, the damping cannot be increased beyond a certain limit value; as we reach it, the $\gamma(V)$-curve turns down (Figure 4.2(a)). As V and γ tend to zero, the separation between the ψ_- and ψ_+ solitons in the complex grows without bounds.

These transformations of the solution are reflected by the behavior of the linearized eigenvalues in the eigenvalue problem (2.4). At the point $V = V_1$, $\gamma = 0$ of the $\gamma(V)$ curve, the twist solution has a quadruplet of complex eigenvalues $\pm\lambda, \pm\overline{\lambda}$ which dissociates into two pairs of complex-conjugate eigenvalues $\lambda_1, \overline{\lambda}_1$ and $\lambda_2, \overline{\lambda}_2$ (with $\operatorname{Re}\lambda_1 < 0$ and $\operatorname{Re}\lambda_2 > 0$) as γ deviates from zero. As we move toward the maximum of the curve, the imaginary parts of λ_1 and λ_2 decrease and the four complex eigenvalues move onto the real axis. At the point of maximum, one of the resulting two positive eigenvalues crosses to the negative real axis, but the other one persists all the way to $V = -0$ and $\gamma = +0$. Therefore the spectrum of eigenvalues on the "downhill" portion of the curve is a union of eigenvalues of the ψ_- and ψ_+ solitons. The conclusion of the eigenvalue analysis is that the traveling complex whose bifurcation diagram is exhibited in Figure 4.2(a) is unstable for all V and γ.

4.4. Numerical continuation: Large driving amplitudes. For $h > 0.28$ we have three starting points with $P = 0$ corresponding to two intersections of the dashed curve and one of the dash-dotted curve with the horizontal axis in Figure 4.1.

The $\gamma(V)$ curve emanating out of the point V_1 is the top, arc-shaped, curve in Figure 4.2(b). For $V = V_1$ and $\gamma = 0$ the solution is symmetric and its shape resembles two strongly overlapping twists. The linearized spectrum includes two complex quadruplets. As γ deviates from zero, the symmetry is lost and the solution starts looking like an asymmetric complex of two pulses. The two complex quadruplets become four complex-conjugate pairs of eigenvalues, two with positive and two with negative real parts. Two of these pairs (one with $\operatorname{Re}\lambda > 0$ and one with $\operatorname{Re}\lambda < 0$) move on to the real axis. After that, one positive real eigenvalue crosses to the negative semiaxis, while the complex pair with $\operatorname{Re}\lambda > 0$ crosses into the $\operatorname{Re}\lambda < 0$ half-plane but then returns to $\operatorname{Re}\lambda > 0$. As $V, \gamma \to 0$, the separation between the ψ_- and ψ_+ solitons comprising this complex increases, and eventually the two constituents diverge to infinities. On the "downhill" portion of the curve, the spectrum is a union of

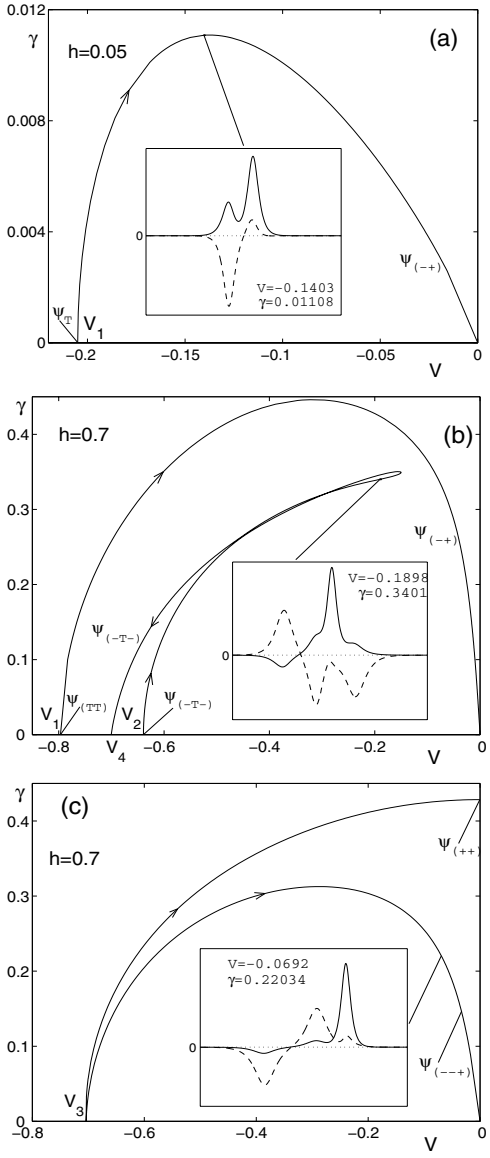


FIG. 4.2. *Results of the numerical continuation of the undamped traveling solitons to nonzero* γ. (a): *small* h; (b),(c): *large* h. *The inset displays a representative solution at one of the internal points of the curve. (Solid line: real part; dashed line: imaginary part.) Each curve shown has a positive-velocity counterpart which arises by the mirror reflection* $V \to -V$ *of the figure.*

the spectra of the individual ψ_- and ψ_+ solitons; in particular, it includes a positive real eigenvalue and a complex quadruplet. Since there are eigenvalues with $\operatorname{Re}\lambda > 0$ for all V, the entire branch is unstable.

The second undamped traveling wave with zero momentum (point V_2 on the bifurcation diagram Figure 4.1) corresponds to a symmetric ($\psi(-x) = \overline{\psi}(x)$) complex of two ψ_- and one twist soliton, symbolically $\psi_{(-T-)}$. The spectrum includes three complex quadruplets. As we continue in γ and V, the symmetry is lost but the

solution still looks like a complex of three solitons (see the inset to Figure 4.2(b)). The bottom, spike-shaped, curve in Figure 4.2(b) depicts the corresponding $\gamma(V)$ relation. Unlike the branch starting at the value $V = V_1$, this solution cannot be continued to zero velocities. Instead, the $\gamma(V)$ curve turns back and, as γ approaches zero from above, V tends to a negative value V_4, with $|V_4| > |V_2|$. For sufficiently small γ the corresponding solution consists of two ψ_- solitons and a twist in between, with the intersoliton separations growing to infinity as $\gamma \to 0$, $V \to V_4$. The associated eigenvalues perform rather complicated movements on the complex plane; skipping the details, it suffices to mention that "unstable" eigenvalues (real positive or complex with positive real parts) are present for all V. Hence the entire branch is unstable.

Finally, the point V_3 on Figure 4.1 represents *two* nonequivalent asymmetric solutions with zero momentum, $\psi_1(\xi)$ and $\psi_2(\xi)$, with $\psi_2(\xi) \equiv \overline{\psi}_1(-\xi)$. Consequently, there are *two* distinct $\gamma(V)$-branches coming out of this point (Figure 4.2(c)). One of these corresponds to a complex of two solitons; when continued to $V = 0$, it gives rise to the symmetric complex $\psi_{(++)}$ with nonzero γ. (See the top curve in Figure 4.2(c).) Continuing the other asymmetric solution to $V = 0$, the corresponding value of γ reaches a maximum at $|V| \sim 0.3$ and then tends to zero (the bottom curve in Figure 4.2(c)). For sufficiently small V and γ this solution represents a complex $\psi_{(--+)}$ (shown in the inset to Figure 4.2(c)). As $V, \gamma \to 0$, the intersoliton separation tends to infinity. Turning to the eigenvalues, the start-off solution at the point V_3 has two complex quadruplets and a real positive eigenvalue in its spectrum. When we continue along the top curve in Figure 4.2(c), two complex eigenvalues move on to the positive real axis, so we end up with three positive eigenvalues. When we continue along the bottom curve, the movements of the eigenvalues are more involved but some of them always remain in the "unstable" half-plane, $\operatorname{Re}\lambda > 0$. The upshot of the eigenvalue analysis is that both curves represent only unstable solutions.

5. Consistency of the two approaches. To complete our classification of damped traveling solitons, we need to comment on what may seem to be an inconsistency between results obtained within the above two complementary approaches. The solution representing the well-separated ψ_+ and ψ_- solitons reported in sections 3 and 4 can be reached by continuing both off the ($\gamma = 0$)- and ($V = 0$)-axes. (This branch connecting to the origin on the (V, γ)-plane appears in both Figures 3.2(a) and 4.2(b).) Although such a curve should obviously not depend on the starting point of the continuation, one notices that the $\psi_{(+-)}$ branches "flowing into the origin" in Figures 3.2(a) and 4.2(b) behave differently when traced backward (i.e., away from $V = \gamma = 0$). While the curve in Figure 3.2(a) intersects the γ-axis, its counterpart in Figure 4.2(b) crosses the other, V-, axis. (Here the reader should not be confused by the fact that the $\psi_{(+-)}$ branch in Figure 3.2(a) is shown for positive values and its counterpart in Figure 4.2(b) for negative values of V. In view of the $\xi \to -\xi$, $V \to -V$ invariance of (2.3), to each γ there correspond *two* traveling waves, one with positive and the other with negative values of V, respectively. Therefore, one should mirror-reflect Figure 4.2(b) prior to comparing it to Figure 3.2(a). This reflection maps the solution $\psi_{(-+)}$ of Figure 4.2(b) to the $\psi_{(+-)}$ of Figure 3.2(a).)

To resolve the paradox, one needs to note that the two figures correspond to different values of h: Figure 3.2(a) to $h = 0.8353$ and Figure 4.2(b) to $h = 0.7$. It turns out that a qualitative change of behavior occurs for h somewhere between these values, more precisely between 0.82 and 0.8275. For $h = 0.82$ and smaller (in particular, for $h = 0.7$) the $\gamma(V)$ curve has the form of an arc shown in Figure 4.2(b) (i.e., it crosses the V-axis as $|V|$ is increased), while for $h = 0.8275$ and greater, the

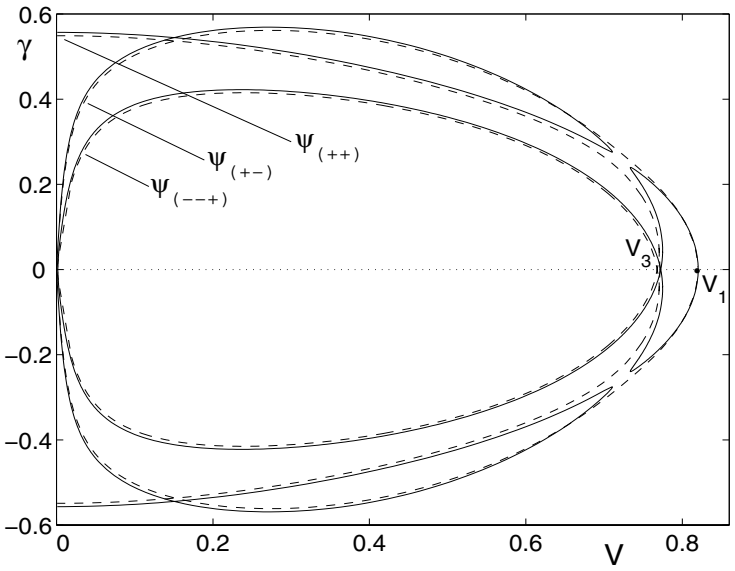


FIG. 5.1. *The comparison of the bifurcation diagrams for* $h = 0.82$ *and* $h = 0.8275$. *Solid curve:* $h = 0.8275$*; dashed curve:* $h = 0.82$. *Although the points of intersection of the dashed and solid curves with the* V*-axis do not coincide, i.e.,* $V_3(0.8275) \neq V_3(0.82)$ *and* $V_1(0.8275) \neq V_1(0.82)$*, they are quite close to each other and hence we mark them as a single point (*V_3 *and* V_1*, respectively). The branches coming out of the point* V_2 *on the* V*-axis are omitted for visual clarity. Note the shape of the dashed and solid curves near* $V \sim 0.7$*,* $\gamma \sim 0.3$*, characteristic of phase portraits of two-dimensional dynamical systems in the neighborhood of saddle points.*

curve is already loop-shaped and does not reach to $\gamma = 0$. This change of behavior, accounting for the above "inconsistency," is illustrated by Figure 5.1 which compares the $\gamma(V)$ relations for $h = 0.82$ and $h = 0.8275$. Figure 5.1 also serves to illustrate the different outcomes of the continuation of the complex $\psi_{(++)}$ for $h = 0.8353$ and smaller h. (We note that for $h = 0.8353$ the continuation of the motionless $\psi_{(++)}$ produces a pair of infinitely separated solitons ψ_+ and ψ_- (Figure 3.2(a)) while for $h = 0.7$, the curve departing from the same type of starting point (i.e., from $\psi_{(++)}$) ends up at the undamped asymmetric solution traveling with nonzero velocity V_3 (Figure 4.2(c)).)

The above differences in behavior result from the presence of a saddle point on the (V, γ)-plane, in the gap between the two lobes of the solid curve in Figure 5.1. Indeed, the dashed and solid curves can be seen as sections of the surface $h = h(\gamma, V)$ by the horizontal planes $h = 0.82$ and $h = 0.8275$, respectively. The gap in the upper solid curve is then accounted for by letting $h = h_0 + x^2 - y^2$ in the vicinity of the gap. Here the constant h_0 lies somewhere between 0.82 and 0.8275, and (x, y) is a pair of suitably chosen orthogonal coordinates on the (V, γ)-plane.

6. Conclusions. One of the conclusions of this work is that by clustering into complexes, solitons (or, equivalently, solitary pulses) can adjust their total momentum to zero. By doing so they can travel with nonzero speed in the presence of damping—without violating the momentum decay law, $\dot{P} = -\gamma P$. Two identical solitons traveling at the same speed in the same direction have equal momenta; therefore, in order to arrange for $P = 0$ the traveling complex inevitably has to include solitons of different varieties (i.e., both ψ_+'s and ψ_-'s). Consequently, the real and

imaginary parts of the traveling complex will always be represented by *asymmetric* functions of $\xi = x - Vt$.

Although the possibility of nondecelerated motion may be out of line with the common perception of the soliton dynamics in weakly damped Hamiltonian equations, moving pulses are not unknown in *strongly* dissipative systems. A suitable example is given by the complex Ginzburg–Landau equation. Asymmetric Ginzburg–Landau pulses, uniformly traveling with nonzero velocities, were reported in [44].

All moving solutions that we have found in this paper turned out to be unstable. This instability admits a simple qualitative explanation—at least, for small dampings. In the undamped situation, the ψ_- solitons are unstable when traveling with small velocities while the ψ_+'s become unstable when moving sufficiently fast [34]. In the presence of dissipation the traveling wave has to include solitons of both varieties; on the other hand, the eigenvalues corresponding to small nonzero γ should remain close to their ($\gamma = 0$)-counterparts. Therefore the spectrum of the traveling complex will "inherit" unstable eigenvalues of either ψ_- (for small velocities) or ψ_+ (for large velocities).

Thus, despite the fact that the parametric driver can sustain the uniform motion of a damped soliton, an additional agent (such as, possibly, the diffusion and/or a nonlinear damping term) is required to make this motion stable. Here it is appropriate to refer, again, to the complex Ginzburg–Landau equation. Stable Ginzburg–Landau pulses arise as a result of a delicate balance of the whole series of terms, including dispersion, cubic and quintic conservative nonlinearity, diffusion, cubic gain, and linear and quintic nonlinear damping [44, 45, 46, 47]. In a similar way, the gain/loss and spreading/steepening balances of the damped-driven traveling solitons could be restored by adding one or several missing agents.

Acknowledgment. We thank Nora Alexeeva for her advice on numerics.

REFERENCES

[1] J.W. Miles, *Parametrically excited solitary waves*, J. Fluid Mech., 148 (1984), pp. 451–460.
[2] C. Elphick and E. Meron, *Localised structures in surface waves*, Phys. Rev. A (3), 40 (1989), pp. 3226–3229.
[3] X.N. Chen and R.J. Wei, *Dynamic behaviour of a non-propagating soliton under a periodically modulated oscillation*, J. Fluid Mech., 259 (1994), pp. 291–303.
[4] W. Zhang and J. Viñals, *Secondary instabilities and spatiotemporal chaos in parametric surface waves*, Phys. Rev. Lett., 74 (1995), pp. 690–693.
[5] W. Wang, X. Wang, J. Wang, and R. Wei, *Dynamical behavior of parametrically excited solitary waves in Faraday's water trough experiment*, Phys. Lett. A, 219 (1996), pp. 74–78.
[6] X. Wang and R. Wei, *Dynamics of multisoliton interactions in parametrically resonant systems*, Phys. Rev. Lett., 78 (1997), pp. 2744–2747.
[7] X. Wang and R. Wei, *Oscillatory patterns composed of the parametrically excited surface-wave solitons*, Phys. Rev. E (3), 57 (1998), pp. 2405–2410.
[8] A. Il'ichev, *Faraday resonance: Asymptotic theory of surface waves*, Phys. D, 119 (1998), pp. 327–351.
[9] G. Miao and R. Wei, *Parametrically excited hydrodynamic solitons*, Phys. Rev. E (3), 59 (1999), pp. 4075–4078.
[10] D. Astruc and S. Fauve, *Parametrically amplified 2-dimensional solitary waves*, in Proceedings of the IUTAM Symposium on Free Surface Flows, Birmingham, UK, 2000, Fluid Mech. Appl. 62, A.C. King and Y.D. Shikhmurzaev, eds., Kluwer, Dordrecht, The Netherlands, 2001, pp. 39–46.
[11] I.H. Deutsch and I. Abram, *Reduction of quantum noise in soliton propagation by phase-sensitive amplification*, J. Opt. Soc. Amer. B Opt. Phys., 11 (1994), pp. 2303–2313.
[12] A. Mecozzi, L. Kath, P. Kumar, and C.G. Goedde, *Long-term storage of a soliton bit stream by use of phase-sensitive amplification*, Opt. Lett., 19 (1994), pp. 2050–2052.

[13] S. Longhi, *Ultrashort-pulse generation in degenerate optical parametric oscillators*, Opt. Lett., 20 (1995), p. 695–697.
[14] V.J. Sánchez-Morcillo, I. Pérez-Arjona, F. Silva, G.J. de Valcárcel, and E. Roldán, *Vectorial Kerr-cavity solitons*, Opt. Lett., 25 (2000), pp. 957–959.
[15] I.V. Barashenkov, M.M. Bogdan, and V.I. Korobov, *Stability diagram of the phase-locked solitons in the parametrically driven, damped nonlinear Schrödinger equation*, Europhys. Lett., 15 (1991), pp. 113–118.
[16] B. Denardo, B. Galvin, A. Greenfield, A. Larraza, S. Putterman, and W. Wright, *Observation of localised structures in nonlinear lattices: Domain walls and kinks*, Phys. Rev. Lett., 68 (1992), pp. 1730–1733.
[17] W.-Z. Chen, *Experimental observation of solitons in a 1D nonlinear lattice*, Phys. Rev. B (3), 49 (1994), pp. 15063–15066.
[18] G. Huang, S.-Y. Lou, and M. Velarde, *Gap solitons, resonant kinks, and intrinsic localised modes in parametrically excited diatomic lattices*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 6 (1996), pp. 1775–1787.
[19] N.V. Alexeeva, I.V. Barashenkov, and G.P. Tsironis, *Impurity-induced stabilization of solitons in arrays of parametrically driven nonlinear oscillators*, Phys. Rev. Lett., 84 (2000), pp. 3053–3056.
[20] W. Chen, B. Hu, and H. Zhang, *Interactions between impurities and nonlinear waves in a driven nonlinear pendulum chain*, Phys. Rev. B (3), 65 (2002), 134302.
[21] D.J. Kaup and A.C. Newell, *Solitons as particles, oscillators, and in slowly changing media: A singular perturbation theory*, Proc. Roy. Soc. London Ser. A, 361 (1978), pp. 413–446.
[22] I.V. Barashenkov and Yu. S. Smirnov, *Existence and stability chart for the ac-driven, damped nonlinear Schrödinger equation*, Phys. Rev. E (3), 54 (1996), pp. 5707–5725.
[23] G. Terrones, D.W. McLaughlin, E.A. Overman, and A. Pearlstein, *Stability and bifurcation of spatially coherent solutions of the damped-driven NLS equation*, SIAM J. Appl. Math., 50 (1990), pp. 791–818.
[24] I.V. Barashenkov and E.V. Zemlyanaya, *Existence threshold for the ac-driven nonlinear Schrödinger solitons*, Phys. D, 132 (1999), pp. 363–373.
[25] B.A. Malomed, *Bound solitons in the nonlinear Schrödinger–Ginzburg–Landau equation*, Phys. Rev. A (3), 44 (1991), pp. 6954–6957.
[26] D. Cai, A.R. Bishop, N. Grønbech-Jensen, and B.A. Malomed, *Bound solitons in the ac-driven, damped nonlinear Schrödinger equation*, Phys. Rev. E (3), 49 (1994), pp. 1677–1679.
[27] I.V. Barashenkov, Yu. S. Smirnov, and N.V. Alexeeva, *Bifurcation to multisoliton complexes in the ac-driven, damped nonlinear Schrödinger equation*, Phys. Rev. E (3), 57 (1998), pp. 2350–2364.
[28] M. Kollmann, H.W. Capel, and T. Bountis, *Breathers and multibreathers in a periodically driven damped discrete nonlinear Schrödinger equation*, Phys. Rev. E (3), 60 (1999), pp. 1195–1211.
[29] I.V. Barashenkov and E.V. Zemlyanaya, *Stable complexes of parametrically driven, damped nonlinear Schrödinger solitons*, Phys. Rev. Lett., 83 (1999), pp. 2568–2571.
[30] D.J. Kaup and A.C. Newell, *Theory of nonlinear oscillating dipolar excitations in one-dimensional condensates*, Phys. Rev. B (3), 18 (1978), pp. 5162–5167.
[31] C. Elphick and E. Meron, *Comment on "Solitary waves generated by subcritical instabilities in dissipative systems,"* Phys. Rev. Lett., 65 (1990), p. 2476.
[32] Yu. S. Kivshar and B.A. Malomed, *Dynamics of solitons in nearly integrable systems*, Rev. Modern Phys., 61 (1989), pp. 763–915.
[33] V.S. Shchesnovich and I.V. Barashenkov, *Soliton-radiation coupling in the parametrically driven, damped nonlinear Schrödinger equation*, Phys. D, 164 (2002), pp. 83–109.
[34] I.V. Barashenkov, E.V. Zemlyanaya, and M. Bär, *Travelling solitons in the parametrically driven nonlinear Schrödinger equation*, Phys. Rev. E (3), 64 (2001), 016603.
[35] P. Coullet, J. Lega, B. Houchmanzadeh, and J. Lajzerowicz, *Breaking chirality in nonequilibrium systems*, Phys. Rev. Lett., 63 (1990), pp. 1352–1355.
[36] P. Coullet, J. Lega, and Y. Pomeau, *Dynamics of Bloch walls in a rotating magnetic field: a model*, Europhys. Lett., 15 (1991), pp. 221–226.
[37] P. Coullet and K. Emilsson, *Strong resonances of spatially distributed oscillators: A laboratory to study patterns and defects*, Phys. D, 61 (1992), pp. 119–131.
[38] B.A. Malomed and A.A. Nepomnyashchy, *Stability limits for arrays of kinks in two-component nonlinear systems*, Europhys. Lett., 27 (1994), pp. 649–653.
[39] C. Elphick, A. Hagberg, B.A. Malomed, and E. Meron, *On the origin of traveling pulses in bistable systems*, Phys. Lett. A, 230 (1997), p. 33–37.

[40] S. Longhi, *Perturbation of parametrically excited solitary waves*, Phys. Rev. E (3), 55 (1997), pp. 1060–1070.

[41] D.V. Skryabin, A. Yulin, D. Michaelis, W.J. Firth, G.-L. Oppo, U. Peschel, and F. Lederer, *Perturbation theory for domain walls in the parametric Ginzburg–Landau equation*, Phys. Rev. E (3), 64 (2001), 056618.

[42] C. Utzny, W. Zimmermann, and M. Bär, *Resonant spatio-temporal forcing of oscillatory media*, Europhys. Lett., 57 (2002), pp. 113–119.

[43] G.J. de Valcárcel, I. Pérez-Arjona, and E. Roldán, *Domain walls and Ising–Bloch transitions in parametrically driven systems*, Phys. Rev. Lett., 89 (2002), 164101.

[44] V.V. Afanasjev, N. Akhmediev, and J.M. Soto-Crespo, *Three forms of localised solutions of the quintic complex Ginzburg–Landau equation*, Phys. Rev. E (3), 53 (1996), pp. 1931–1939.

[45] O. Thual and S. Fauve, *Localized structures generated by subcritical instabilities*, J. Phys. France, 49 (1988), pp. 1829–1833.

[46] S. Fauve and O. Thual, *Solitary waves generated by subcritical instabilities in dissipative systems*, Phys. Rev. Lett., 64 (1990), pp. 282–284.

[47] N. Akhmediev and A. Ankiewicz, *Solitons of the complex Ginzburg–Landau equation*, in Spatial Solitons, Springer Series in Optical Sciences 82, S. Trillo and W. Torruellas, eds., Springer-Verlag, Berlin, 2001.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 819–857

A BOUND ON THE TOTAL VARIATION OF THE CONSERVED QUANTITIES FOR SOLUTIONS OF A GENERAL RESONANT NONLINEAR BALANCE LAW*

JOHN HONG† AND BLAKE TEMPLE‡

Abstract. We introduce a new potential interaction functional and use it to define a new Glimm-type functional that bounds the total variation of the conserved quantities at time $t > 0$ by the total variation at time $t = 0+$ in Glimm approximate solutions of a general resonant nonlinear balance law.

Key words. shock waves, resonance, Glimm scheme, balance laws

AMS subject classification. 35L65

DOI. 10.1137/S0036139902405249

1. Introduction.

In [13], Isaacson and Temple introduced the 2×2 system

$$\begin{aligned} a_t &= 0, \\ u_t + f(a,u)_x &= a' g(a,u) \end{aligned} \tag{1.1}$$

as a general nonlinear balance law that models resonance between a nonlinear wave field and a stationary source (cf. [5, 7, 8, 9, 10, 11, 14, 17, 19, 20, 21, 22, 23, 25, 26]). Here a and u are assumed to be scalar valued, and resonance occurs at states $U_* = (a_*, u_*)$, where the nonlinear wave speed $\lambda = f_u$ vanishes. Assume further that f and g are smooth functions and that the following conditions are satisfied at the state U_*:

$$f_u(U_*) = 0, \tag{1.2}$$

$$g(U_*) - f_a(U_*) \neq 0 \quad \text{(w.l.o.g. assume } g(U_*) - f_a(U_*) > 0), \tag{1.3}$$

$$f_{uu}(U_*) \neq 0 \quad \text{(w.l.o.g. assume } f_{uu}(U_*) < 0), \tag{1.4}$$

and

$$g_u(U_*) \neq 0. \tag{1.5}$$

It was shown in [13] that the generic conditions (1.2)–(1.5) imply that the structure of elementary wave curves (shock waves, rarefaction waves, and standing waves) and the solution of the Riemann problem (the initial value problem when the initial data consists of constant states U_L, U_R, separated by a discontinuity) are canonical[1] in a neighborhood Ω of the state U_*; cf. [16, 13, 24]. (The cases $g_u > 0$ and $g_u < 0$ are

*Received by the editors April 8, 2002; accepted for publication (in revised form) February 22, 2003; published electronically March 11, 2004.
http://www.siam.org/journals/siap/64-3/40524.html
†Department of Mathematics, UCLA, Los Angeles, CA 90095-1555 (jhong@math.ucla.edu).
‡Department of Mathematics, University of California, Davis, Davis, CA 95616 (jbtemple@ucdavis.edu). The research of this author was supported in part by NSF Applied Mathematics grant DMS-010-2493 and by the Institute of Theoretical Dynamics, UC-Davis.

[1]Here we show that the condition $f_a \neq 0$, stated as a condition for genericity in [13], is not required, except by (1.3) in the case $g = 0$; cf. the construction of the zero speed shock curve below.

qualitatively different.) Here $a' \equiv a_x \equiv \frac{da}{dx}$, and $a = a(x)$ is an inhomogeneous term that is treated as a variable so that (1.1) takes the form of a system of two equations that expresses the dependence of the solution on the source a.

In this paper we introduce a new potential interaction functional and use it to construct a nonlinear Glimm functional that is positive decreasing on solutions of (1.1) and bounds the total variation of the conserved quantity u in terms of the initial data for all time $t > 0$. We show that the functional is always locally finite at time $t = 0+$ of the random choice method, and so the limit solution will be of bounded total variation for all time so long as this functional is bounded uniformly at $t = 0+$ as the mesh length $\Delta x \to 0$. This then gives a condition on the initial data that guarantees the solution will be of bounded total variation in u for all time. Moreover, the potential interaction estimate can be interpreted as the best possible estimate for the increase in total variation in u that can occur due to the interaction of an initial set of waves, taking no account of the initial distances between the waves or the times at which pairs of waves will interact. As part of our proof, we show that the only potential for increase of total variation is due to the interaction of rarefaction waves and standing waves. An immediate consequence of this is a proof that *the total variation of u at any $t > 0$ will be uniformly bounded by a constant times the total variation of u at $t = 0+$ in any weak solution of* (1.1) *generated by the generalized Glimm method, which initially consists entirely of shock waves and standing waves.*

The lack of a total variation estimate in the conserved quantities is the main obstacle to extending the results in [25, 13] to systems of equations (that is, when u is a vector instead of a scalar), and this is the primary motivation for our work. An important example of a system of form (1.1) is given by the equations for compressible Euler flow in a variable area duct:

$$
\begin{aligned}
&a_t = 0,\\
&\rho_t + (\rho u)_x = -\frac{a'}{a}\rho u,\\
&(\rho u)_t + (\rho u^2 + p)_x = -\frac{a'}{a}\rho u^2,\\
&(\rho E)_t + (\rho E u + p u)_x = -\frac{a'}{a}(\rho E u + p u),
\end{aligned}
\tag{1.6}
$$

where ρ is the density, p is the pressure, E is the energy density, and $a(x)$ is the diameter of the duct at position x [2]. It is a mathematical open problem to show that wave strengths remain bounded in the time evolution of solutions of (1.6) in a neighborhood of a point of resonance U_* when the flow is transonic; cf. [1]. The main thrust of this paper is thus to establish total variation estimates for (1.1) that can be extended to a general class of systems of form (1.1), which includes (1.6). Now the total variation in the conserved quantity u at time $t > 0$ in a solution of (1.1) is not in general bounded by any uniform constant times the total variation of u at time zero in the presence of resonance. In fact, solutions of the linearization of (1.1) about $U = U_*$ grow unboundedly as $t \to \infty$ [13]. In [25, 13] a time independent bound on the supnorm and global existence of weak solutions is demonstrated based on obtaining a time independent total variation estimate for solutions in the coordinate system of Riemann invariants,[2] which is related to the conserved variables $U = (a, u)$ by a *singular* coordinate transformation. These estimates do not carry over naturally to

[2]The fact that solutions are bounded at all is thus a purely nonlinear effect.

systems like (1.6), in which u is a vector. Indeed, Glimm's method indicates that a time independent bound on the total variation of the conserved quantities is needed to extend the analysis to systems. To establish a bound on the total variation of the conserved quantity U, we introduce a singular transformation *of the coordinate system of Riemann invariants* and give essentially the best possible bound on the total variation at time $t > 0$ in terms of the initial data in these coordinates, which are regular with respect to the coordinates of conserved quantities. Our method of analysis is then to adapt the linear functional introduced in [25, 13] over to these new coordinates (which requires a correction term for the wave strengths of certain standing waves in order to make the linear part of the functional continuous) and then to add a potential interaction term for rarefaction wave-standing wave interactions to account for the fact that the functional is not contractive (decreasing in time) in these new coordinates. The total variation bounds on the solutions imply supnorm bounds, and these bounds help explain why, as waves interact due to the nonlinearity of wave speeds, solutions of the nonlinear problem (1.1) do not blow up like the resonant linear equation but rather decay to time asymptotic wave patterns given by the solutions of the Riemann problem.

We use the notation $U = (a, u),\ \mathcal{F} = (0, f),\ G = (0, a'g)$ so that the initial value problem for (1.1) is a special case of the general initial value problem,

$$\begin{aligned} &U_t + \mathcal{F}(U)_x = G(U), \\ &U(x, 0) = U_0(x). \end{aligned} \tag{1.7}$$

The advantage of treating systems in the form (1.1) instead of general systems of form (1.7) is that for system (1.1) we can define a generalized Riemann problem and analyze solutions by Glimm-type methods that can be applied, in principle, to systems of equations. The point of incorporating the a' term in front of g on the right-hand side of (1.1) is that it ensures that standing waves can be rescaled into discontinuities [13, 6]. It was shown in [6] that in the strictly hyperbolic regime, general source terms can be treated like contact discontinuities in such a way that the Riemann problem of Lax, and the random choice method of Glimm, both extend *virtually unchanged* to systems of the form (1.1)—that is, general systems with sources can be treated numerically just as the source-free equations. Of course, since the right-hand side of (1.1) involves the derivative a', there is no classical weak formulation of (1.1) when a is discontinuous—you cannot multiply a delta function by a discontinuous function in the classical theory of distributions; cf. [3]. Thus, the generalized Riemann problems used to construct the Glimm approximates are *weaker than weak* solutions of the equations; cf. [6]. To justify the method, it is important to show that the limits of approximate solutions of the generalized Glimm method are veritable weak solutions of (1.1) *when system* (1.1) *has a weak formulation*, namely, when $a(x)$ is Lipschitz continuous. This is accomplished in [6].[3] The interesting point to make here is that because the Riemann problems are based on approximating $a(x)$ by piecewise constant states, it follows that the Glimm scheme approximates can give only a C^0 and not a C^1 approximation of $a(x)$, and thus a' is not well approximated in L^1. Even so, Hong showed in [6] that for any test function ϕ, the residual and, in particular, $\int_{t\geq 0} a'g(a, u)\phi(x, t)dxdt$ converges not by L^1 convergence (as in Glimm's original results) but weakly, by *oscillation*, when a is Lipschitz continuous; cf. [21]. This argument, appropriately modified for the resonant

[3]Note also that every a of bounded variation can be approximated by a Lipschitz continuous function.

case considered here, is presented in section 6 below. Interestingly, three mollification parameters are needed to conclude the proof of convergence of the residual in section 6.

In section 2 we review the results in [13]; we define the regular transformation $(a, u) \to (a, w)$ and the linear functional $L_w(J)$ and compare these to the singular transformation $(a, u) \to (a, z)$ and linear functional $L_z(J)$ defined in [25, 13].[4] We then review the solution of the Riemann problem and construct the admissible solution $[U_L, U_R]$ based on an L_w minimization principle that is finer than the L_z minimization principle introduced in [13]. The L_w minimization is required for the subsequent analysis. The nonuniqueness of solutions of the Riemann problem even in the presence of the classical entropy condition for the nonlinear waves reflects an interesting instability in the time asymptotics of solutions of (1.1).

In section 3 we construct the approximate solutions $U_{\Delta x}$ by the generalized Glimm method. For a given approximate solution, the functionals $L_w(J)$ and $L_z(J)$ both sum the strengths of waves that cross an I-curve J with weight factors according to whether the wave is a nonlinear wave, a *weak* standing wave, or a *strong* standing wave, respectively; cf. [25].[5] The purpose of the weight factors is to make $L_z([U_L, U_R])$ and $L_w([U_L, U_R])$ continuous functions of U_L and U_R for the admissible solution of the Riemann problem $[U_L, U_R]$ (cf. [16, 24] and (2.1) below). Now it was shown in [13] that the weight factors $1, 2$, and 4 on nonlinear waves, weak standing waves, and strong standing waves, respectively, suffice to make L_z continuous (these weights were introduced in [25]). We show here that in the case $g = 0$, the weight factors $1, 2, 4$ also suffice to make $L_w([U_L, U_R])$ continuous functions of U_L and U_R. However, when $g_u \neq 0$, we must adjust the definition of strength for the standing waves in order to preserve continuity when the standing wave curves diverge from the zero speed shock curves; cf. [13]. It was shown in [13] that the functional L_z is positive and nondecreasing across interaction diamonds Δ that lie between successive I-curves J_1 and J_2 in an approximate Glimm scheme solution, and $L_z(J)$ bounds the total variation in (a, z) of the solution along J [24, 4]. On the other hand, $L_w(J)$ bounds the total variation in (a, w) (and hence also the total variation in (a, u)) along an I-curve J but does not decrease across interaction diamonds.

In section 4 we define the interaction potential $d(\gamma_0, \gamma_r)$ between a rarefaction wave γ_r and a standing wave γ_0, and in section 5 we define the nonlocal Glimm functional, $P(J) = \sum_{(\alpha,\beta)\in App(J)} d(\gamma_0^\alpha, \gamma_r^\beta)$, and prove that the functional $F(J) = L_w(J) + P(J)$ decreases across interaction diamonds Δ, where the sum is taken over all approaching waves that cross J in a Glimm approximate solution. From this we establish the total variation bound for the generalized Glimm approximates and thus conclude the main total variation bound in the conserved variables (a, u) for solutions of the resonant nonlinear system (1.1). It is fortunate that at the transitions between regions where the structure of the admissible solution Riemann problem changes, the Riemann problem *never* involves rarefaction waves. Moreover, rarefaction waves are never created by interaction, and thus, since the potential interaction functional P only requires the potential for rarefaction waves to interact with standing waves, it follows that the continuity of both P and F is also maintained as states cross

[4] The functionals L_z were labeled "F" in [25] and [13], but we refer to these as L_z here because they contain no potential interaction term and are therefore linear on sequences of elementary waves; cf. [4].

[5] A standing wave is *strong* if the jump in u across the wave has the same sign as the jump in u across a shock wave and *weak* if the jump has the same sign as the jump in u across a rarefaction wave; cf. [13].

transitional boundaries between different regions of the Riemann problem.

In section 6 we modify the argument in [6] and prove the convergence of the residual when a is Lipschitz continuous.

2. Review of the Riemann problem. The Riemann problem is the initial value problem with initial data given at $t = 0$ by the jump discontinuity

$$U_0(x) = \begin{cases} U_L = (a_L, u_L) & \text{if } x < 0, \\ U_R = (a_L, u_L) & \text{if } x > 0. \end{cases} \tag{2.1}$$

The solution of the Riemann problem for (1.1), assuming (1.2)–(1.5), was first described in [13]. The solutions that minimize L_z were constructed within the class of shock waves, rarefaction waves, and standing waves, and the solution was thereby shown to have a canonical structure for pairs of states U_L and U_R in a sufficiently small neighborhood of U_*. In this section we review the solution of the Riemann problem and define the functionals L_z and L_w.

To motivate this, we note that by [13], near a point of resonance U_* of system (1.1), solutions of (2.1) have an interesting multiplicity of solutions even when the standard entropy condition for shocks is imposed on the nonlinear waves. An additional admissibility condition is required to fix a unique solution. For system (1.1) in the case $g = 0$, uniqueness is implied by the Lax entropy condition for shocks, together with the condition that the wave curves for the waves that solve the Riemann problem should lie between the values of a on the left and right; cf. [13]. This is a natural condition if one views the discretization of a as approximating a smooth duct—the time asymptotic wave pattern will depend on the interior structure of the duct as well as the left and right most diameters. However, when $g \neq 0$, system (1.1) has a more interesting and nontrivial multiplicity of solutions of the Riemann problem: in certain cases, there is a multiplicity of three distinct solutions of the Riemann problem that preserve the bounds in a from the left and right, and these reduce to two possible solutions at boundary cases. The main purpose of this section is to define the functional L_w and show that the following admissibility condition is sufficient to pick out a unique solution of the Riemann problem (except of course for a dual ambiguity at the boundary regions where the qualitative wave structure changes).

DEFINITION 2.1. *A solution of the Riemann problem* (2.1) *is called* admissible *if it minimizes L_w among all other solutions of the Riemann problem that preserve the bounds in a and contain only Lax entropy shocks.*[6]

In contrast, the admissibility criterion in [13], which requires that L_z be minimized, still leaves some ambiguity in cases where there are three solutions. We let $[U_L, U_R]$ denote the admissible solution of the Riemann problem, and we will show that $[U_L, U_R]$ always consists of three elementary waves: a negative speed nonlinear wave followed by a single standing wave followed by a positive speed nonlinear wave. However, in two cases diagrammed in Figures 15 and 17, the standing wave must be taken to be what we call a *triple composite standing wave*, a wave that consists of a standing wave followed by a zero speed shock wave followed by a second standing wave.

To start, let γ denote an arbitrary elementary wave, and let subscripts $q = 0, r, s$ identify the wave as a standing wave, rarefaction wave, or shock wave, respectively.

[6] Unlike L_z, L_w is *not* minimized on solutions among all connected sequences of elementary waves that take U_L to U_R, and if it were, $F(J) = L_w(J)$ would decrease on solutions, and no potential interaction term would be required in our analysis; cf. [13].

A wave γ is determined by its left and right states, and we say that $\gamma_a \cdots \gamma_b$ is a *connected* sequence of elementary waves that takes U_L to U_R if the right state of any wave in the list is equal to the left state of its successor in the list, and U_L, U_R is the left, right state of the first, last wave in the list, respectively. Thus the admissible solution of the Riemann problem $[U_L, U_R]$ is just a particular connected sequence of elementary waves that takes U_L to U_R. Two connected sequences of elementary waves $\gamma_a \cdots \gamma_b$ and $\bar{\gamma}_a \cdots \bar{\gamma}_b$ are said to be *similar* if they both take U_L to U_R, in which case we write $\gamma_a \cdots \gamma_b \sim \bar{\gamma}_a \cdots \bar{\gamma}_b$. For a nonlinear wave γ that takes U_L to U_R, we say that $\gamma \sim \gamma^a \gamma^b$ is a *partition* of the wave γ if the state U_M, the right state of γ^a and the left state of γ^b, lies strictly between U_L and U_R in the (a, u)-plane. (We allow both rarefaction waves and shock waves to be partitioned.)

To begin the review of the Riemann problem, we first remind the reader that system (1.1) has standing wave solutions that can be rescaled into discontinuities so that the standing waves can be treated like a family of contact discontinuities in the theory of hyperbolic conservation laws [13, 6]. Indeed, let $(a(x), u(x))$ be a standing wave (i.e., time independent) solution of (1.1). Then

$$\frac{d}{dx} f = a' g,$$

which is equivalent to

$$f_a da + f_u du = g da.$$

We rewrite this as

$$(f_a - g) da + f_u du = 0. \tag{2.2}$$

The nondegeneracy assumption (1.3) implies that $f_a - g \neq 0$ in a neighborhood of U_*, and therefore (2.2) is equivalent to the autonomous ODE

$$\frac{da}{du} = \frac{f_u}{g - f_a}. \tag{2.3}$$

This equation has a unique solution through each point in a neighborhood of U_* in the (a, u)-plane. Thus, for any solution $a = a_s(u)$ of (2.3) and any smooth function $\varphi(x)$, the curve $u = \varphi(x), a = a_s(\varphi(x))$ is a standing wave solution of (1.1). Moreover, if $a_L = a_s(u_L)$ and $a_R = a_s(u_R)$, then the standing wave discontinuity

$$U(x,t) = \begin{cases} (a_L, u_L) & \text{if } x < 0, \\ (a_R, u_R) & \text{if } x > 0 \end{cases} \tag{2.4}$$

is obtained as a limit of smooth solutions; specifically, if $\varphi_\epsilon(x) \to \varphi_0(x)$, where

$$\varphi_0(x) = \begin{cases} u_L & \text{if } x < 0, \\ \\ u_R & \text{if } x > 0, \end{cases}$$

then $U_\epsilon = (a_s(\varphi_\epsilon(x)), \varphi_\epsilon(x)) \to U(x,t)$. Thus we can view the standing wave discontinuities defined in (2.4) as a family of elementary waves for system (1.1), similar to a family of contact discontinuities.

The standing wave curves define solutions of (2.2). Note that for a standing wave,

$$\frac{da}{du} = 0 \text{ if and only if } f_u = 0. \tag{2.5}$$

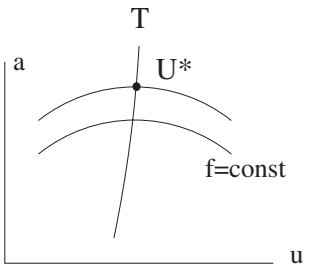


FIG. 1.

Moreover, if $da/du = 0$, then

$$\frac{d^2a}{du^2} = \frac{f_{uu}}{g - f_a} < 0. \tag{2.6}$$

Thus, $d^2a/du^2 < 0$ in a neighborhood of U_*.

DEFINITION 2.2. *The* transition curve $\mathcal{T}$ *associated with system* (1.1) *is the set*

$$\mathcal{T} = \{(a, u) : \lambda \equiv f_u = 0\}. \tag{2.7}$$

Since $f_{uu} \neq 0$, the implicit function theorem implies that (in a neighborhood of U_*) $\mathcal{T}$ is a smooth curve passing through U_*, which we denote by

$$u = u_{\mathcal{T}}(a). \tag{2.8}$$

The curve $\mathcal{T}$ comprises the states near U_* for which the nonlinear wave speed $\lambda \equiv f_u$ is zero. By (2.5) and (2.6), the standing wave curves $u \mapsto (a_s(u), u)$ are convex down, cross $\mathcal{T}$ transversally, and maximize a on $\mathcal{T}$ in some neighborhood of U_*. (The notation comes from [7]. See Figure 1.)

We now define the zero speed shock curve corresponding to a given standing wave curve. By our choice of signs ($f_{uu} < 0$ and $g - f_a > 0$), the entropy shock waves (see [24]) for the nonlinear scalar conservation law $u_t + f(a, u)_x = 0$ jump always from left to right in the (x, t)-plane and (a, u)-plane simultaneously; thus, by the Rankine–Hugoniot jump relation for shocks,

$$s[u] = [f],$$

the zero speed shocks ($s = 0$) cross $\mathcal{T}$ from left to right at a constant value of f.

Now, for a given standing wave $a = a_s(u)$ and a given state (a, u) on this standing wave, define $\bar{u}$ to be the value of u such that the state $(a, \bar{u})$ lies on the opposite side of $\mathcal{T}$ at the same a-level and on the same standing wave curve as the given state (a, u). If the state $U = (a, u)$ lies on the left-hand side of $\mathcal{T}$ (we write $U < \mathcal{T}$), then define $\tilde{u}$ to be the value of u such that the state $(a, \tilde{u})$ lies on the right-hand side of $\mathcal{T}$ and at the same level a, but on the same constant f curve as the given state (a, u). That is, for $U < \mathcal{T}$, $\bar{u}$ satisfies

$$a_s(\bar{u}) = a_s(u), \tag{2.9}$$

and $\tilde{u}$ satisfies

$$f(a, \tilde{u}) = f(a, u) \tag{2.10}$$

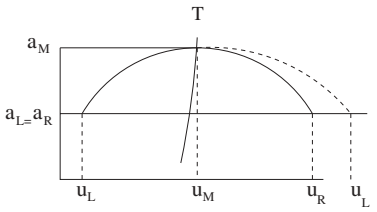


FIG. 2.

(see Figure 1).

DEFINITION 2.3. *Let $a = a_s(u)$ be a standing wave curve. Then, (assuming $f_{uu} < 0$) the zero speed shock curve corresponding to standing wave curve a_s is the curve (lying to the right of $\mathcal{T}$) defined by*

$$\{\tilde{u} : f(a, \tilde{u}) = f(a, u) \text{ where } u \le u_{\mathcal{T}}(a) \text{ and } \tilde{u} \ge u_{\mathcal{T}}(a)\}.$$

(When $f_{uu} > 0$, we change to $u \ge u_{\mathcal{T}}(a)$ and $\tilde{u} \le u_{\mathcal{T}}(a)$.)

LEMMA 2.4. *If $g_u < 0$, then for each standing wave curve $a = a_s(u)$, the corresponding zero speed shock curve lies to the right of the standing wave curve in the (a, u)-plane. That is, if (a, u) satisfies $a = a_s(u)$ with $u < u_{\mathcal{T}}(a)$, then*

$$f(a, \bar{u}) < f(a, \tilde{u}) = f(a, u). \tag{2.11}$$

If $g_u > 0$, then the corresponding zero speed shock curve lies to the left of the standing wave curve in the (a, u)-plane. That is,

$$f(a, \bar{u}) > f(a, \tilde{u}) = f(a, u). \tag{2.12}$$

For example, in the case $g_u < 0$, Lemma 2.4 implies that the zero speed shock curve lies above and to the right of the standing wave curve $a_s(u)$ (see Figure 2). (For a proof of Lemma 2.4, see [13, Lemma 2.4, p. 13], and note that the condition $f_a \neq 0$ was not required.)

We now define the nonsingular coordinate w and functional L_w and formulate the L_w minimization principle to select a unique admissible solution of the Riemann problem. To construct L_w, we first construct w and a functional L_w^* that is analogous to the construction of the singular coordinate z and functional L_z defined in [25, 13], and then we obtain L_z by modifying L_w^* so that $L_w^*[U_L, U_R]$ depends continuously on U_L and U_R. To start, we first review the construction of z and L_z for system (1.1).

The coordinate z is based on the singular coordinate system of nonlinear hyperbolic wave curves ($a =$constant) and standing wave curves ($a = a_s(u)$) as observed in the (a, u)-plane and is defined as follows. For each point (a, u), let $(a_{\mathcal{T}}, u_{\mathcal{T}})$ denote the unique point where the standing wave curve through (a, u) crosses $\mathcal{T}$, and set

$$z(a, u) = \operatorname{sgn}(u - u_{\mathcal{T}})|a - a_{\mathcal{T}}|.$$

Using this, define the strength $|\gamma|_z$ of an elementary wave γ by

$$|\gamma|_z = \begin{cases} |z(U_R) - z(U_L)| & \text{if } \gamma \text{ is a nonlinear wave,} \\ 2|z(U_R) - z(U_L)| & \text{if } \gamma \text{ is a weak standing wave,} \\ 4|z(U_R) - z(U_L)| & \text{if } \gamma \text{ is a strong standing wave.} \end{cases} \tag{2.13}$$

Here a standing wave is *weak* if the jump in u across the wave is in the direction of a rarefaction wave ($u_R < u_L$ since we assume $f_{uu} < 0$) and is *strong* if the jump in u across the wave is in the direction of a shock wave ($u_R > u_L$ when $f_{uu} < 0$); cf. [25, 21]. For a sequence of elementary waves $\gamma_1, \dots, \gamma_n$, define

$$L_z[\gamma_1, \dots, \gamma_n] = \sum_{i=1}^{n} |\gamma_i|_z. \tag{2.14}$$

Analogously, define the nonsingular coordinate w by

$$w(a,u) = \begin{cases} u - u_{\mathcal{T}} & \text{if } u < \mathcal{T}, \\ u_{\mathcal{T}} - \bar{u} & \text{if } u > \mathcal{T} \end{cases}$$

and the strength $|\gamma|_w$ of an elementary wave γ by

$$|\gamma|_w^* = \begin{cases} |w(U_R) - w(U_L)| & \text{if } \gamma \text{ is a nonlinear wave,} \\ 2|w(U_R) - w(U_L)| & \text{if } \gamma \text{ is a weak standing wave,} \\ 4|w(U_R) - w(U_L)| & \text{if } \gamma \text{ is a strong standing wave.} \end{cases} \tag{2.15}$$

For a sequence of elementary waves $\gamma_1, \dots, \gamma_n$, define

$$L_w^*[\gamma_1, \dots, \gamma_n] = \sum_{i=1}^{n} |\gamma_i|_w^*. \tag{2.16}$$

We next show that the change in w across an elementary wave bounds the change in u across the wave in any neighborhood Ω of U_* that is sufficiently small.[7] This is guaranteed by the simpler condition stated in the following lemma. (Since the change in u across a nonlinear wave is equal to the change in w across the wave, the only issue is with the standing waves.)

LEMMA 2.5. *Let γ_0 denote a standing wave with left state U_L and right state U_R, both states lying on one side of the transition curve. Then for Ω sufficiently small, there exists a constant $c > 1$ such that the condition $U_L, U_R \in \Omega$ implies that the absolute change in u across γ_0 between a_L and a_R is always that constant times larger than the absolute change in u along the transition curve $\mathcal{T}$ between a_L and a_R.*

Proof. We verify the lemma in the case diagrammed in Figure 3 (other cases are similar). Thus we show that for Ω sufficiently small, there exists $c > 1$ such that if $U_L, U_R \in \Omega$, then $|DF| > c|GG'|$. (We use the notation that $|DF|$ denotes the absolute change in u between states D and F.) But $|DF| = |DC|$ is the change in u across the wave γ_0. Thus, by construction of the standing wave curves, we know that

$$\frac{du}{da} = \frac{g - f_a}{f_u}$$

along a standing wave curve, so by the mean value theorem

$$|DC| = \frac{g - f_a}{f_u} |a_R - a_L|, \tag{2.17}$$

[7] We treat the local problem here because it demonstrates that the analysis is generic in a neighborhood of any state U_*, but all of this can be globalized to apply to any neighborhood Ω where the solution of the Riemann problem has the canonical structure described in section 2 such that Lemma 2.5 applies.

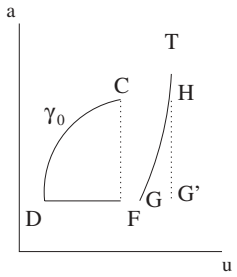


FIG. 3.

where it is understood that $\frac{g-f_a}{f_u}$ is evaluated at some point in Ω. Also $|GG'| = |GH|$ is the change in u along $\mathcal{T}$ between a_L and a_R. Since G and H lie on $\mathcal{T}$, we have $f_u(H) = f_u(G) = 0$, and so differentiating and using the mean value theorem we obtain that

$$|GH| = \frac{f_{ua}}{f_{uu}}|a_R - a_L|. \tag{2.18}$$

Since f_u can be taken arbitrarily small in a neighborhood of $\mathcal{T}$, it follows from (2.17) and (2.18) that there exists a constant $c > 1$ such that $|DC| > c|GH|$ so long as $U_L, U_R \in \Omega$. □

COROLLARY 2.6. *Assume that Ω is sufficiently small so that Lemma 2.5 holds. Then there exists a constant $c > 0$ such that, if $U_L, U_R \in \Omega$, then*

$$c^{-1}|u_L - u_R| < |\gamma|_* < c|u_L - u_R|. \tag{2.19}$$

Proof. The second inequality in (2.19) is clear by construction. We verify the first inequality in (2.19) in the case of a standing wave $|\gamma_0|$ diagrammed in Figure 3. (Again, there is no issue for nonlinear waves, and the cases for other standing waves are similar, because we always assume that standing waves do not cross $\mathcal{T}$). In the case of Figure 3, $|\gamma_0| = 4\{|w(C) - w(D)|\}$. However,

$$\begin{aligned}\frac{1}{4}|\gamma_0| = |w(C) - w(D)| &= ||CH| - |DG|| = ||FG'| - |DG|| = ||DF| - |GG'|| \\ &\geq \left||DF| - c^{-1}|DF|\right| = \left(1 - \frac{1}{c}\right)|DF| = \left(1 - \frac{1}{c}\right)|u(F) - u(D)|\end{aligned}$$

for the $c > 1$ of Lemma 2.5. It follows that

$$|u_L - u_R| \leq \frac{1}{4}\left(1 - \frac{1}{c}\right)^{-1}|\gamma_0|,$$

which proves the corollary. □

From here on out, we always assume that all states lie in a region Ω where lemma 2.5 and Corollary 2.6 apply.

In order to deduce the solution of the Riemann problem from a minimization principle, we will use the following property of the functional L^*_w.

LEMMA 2.7. *Let points A, B, C, D denote the vertices of a region in U-space bounded on either side by standing wave curves and above and below by nonlinear wave curves such that the region lies entirely on one side of the transition curve.*

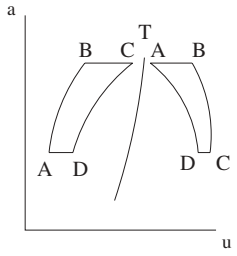


FIG. 4.

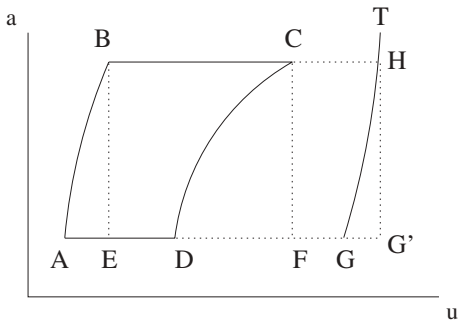


FIG. 5.

Assume that the vertices of the two possible such regions of this type are labeled with the orientation shown in Figure 4. *Then*

$$L_w^*(A \to B \to C) \le L_w^*(A \to D \to C), \tag{2.20}$$

$$L_w^*(D \to A \to B) \le L_w^*(D \to C \to B). \tag{2.21}$$

(Again, we use the convention that an elementary wave can be denoted by the left and right states of the wave separated by an arrow.)

Proof. We verify (2.20) in the case diagrammed in Figure 5, which is similar to Figure 3 of Lemma 2.5. (The other cases are similar.) Referring to Figure 5, we can estimate

$$\begin{aligned} L_w^*(A \to D \to C) - L_w^*(A \to B \to C) \\ &= |AD| + 4|w(C) - w(D)| - |EF| - 4|w(B) - w(A)| \\ &= |AE| - |DF| + 4||CH| - |DG|| - 4||BH| - |AG||. \end{aligned} \tag{2.22}$$

But by Lemma 2.5,

$$\begin{aligned} ||CH| - |DG|| &= |DF| - |GG'|, \\ ||BH| - |AG|| &= |AE| - |GG'|. \end{aligned}$$

Substituting these into (2.22) gives

$$\begin{aligned} L_w^*(A \to D \to C) - L_w^*(A \to B \to C) \\ &= 3|DF| - 4|GG'| + |AE| - 4|AE| + 4|GG'| \\ &= 3|DF| - 3|AE| \ge 0, \end{aligned}$$

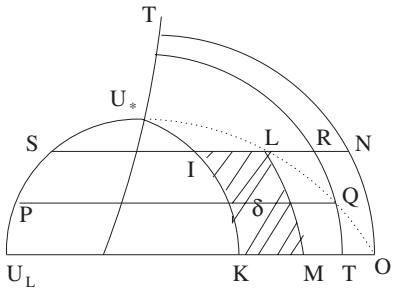


FIG. 6.

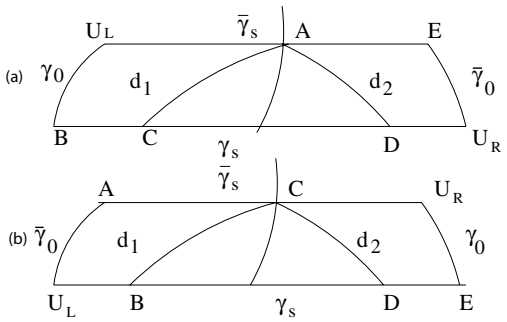


FIG. 7.

where we have used $|AE| \geq |GG'|$ by Lemma 2.5 to conclude the last line. □

The following is a simple corollary of Lemma 2.7.

COROLLARY 2.8. *If γ_a and γ_b are two standing waves on the same side of $\mathcal{T}$ that pass between the same values of a, then $L_w^*(\gamma_a) > L_w^*(\gamma_b)$ if γ_a is the wave closer to the transition curve $\mathcal{T}$.*

The next lemma provides an important continuity property of the functional L_w^* for waves that cross the transition curve.

LEMMA 2.9. *Consider the interaction $\bar{\gamma}_0 + \bar{\gamma}_s \to \gamma_s + \gamma_0$ diagrammed in Figure* 7(a)*. Then, referring to the points referenced in that diagram, we have*

$$L_w^*(U_L \to A \to U_R) = L_w^*(U_L \to E \to U_R) \tag{2.23}$$

and

$$\begin{aligned} d_1 &\equiv L_w^*(U_L \to B \to C) - L_w^*(U_L \to A \to C) \\ &= L_w^*(D \to C \to U_R) - L_w^*(D \to E \to U_R) \equiv d_2. \end{aligned} \tag{2.24}$$

Moreover, statement (2.23) *also holds for the analogous points diagrammed in Figure* 7(b)*, together with*

$$\begin{aligned} d_1 &\equiv L_w^*(U_L \to A \to C) - L_w^*(U_L \to B \to C) \\ &= L_w^*(A \to D \to U_R) - L_w^*(A \to E \to U_R) \equiv d_2. \end{aligned} \tag{2.25}$$

Proof. We verify (2.23) and (2.25). For (2.23), let F and G denote the points such that $L_w^*(A \to U_R) = L_w^*(F \to G)$. Then by the $1, 2, 4$ weightings on wave strengths, it

follows that $|\bar{\gamma}_0|_* = L_w^*(U_L \to A) = L_w^*(U_L \to F) + L_w^*(G \to E) + L_w^*(E \to U_R)$. This is enough to verify (2.23). For (2.25), note that $L_w^*(U_L \to A \to U_R) = d_1 + L_w^*(U_L \to B \to C \to U_R) = d_1 + L_w^*(U_L \to D \to C \to U_R) = d_1 - d_2 + L_w^*(U_L \to E \to U_R)$, so by (2.23), $d_1 = d_2$. □

We now define L_w in terms of L_w^*. To this end, note that because of Lemma 2.7, the functional $L_w^*([U_L, U_R])$ will be a continuous function of U_L and U_R on the admissible solution of the Riemann problem *only in the case* when $g_u \equiv 0$, and in this case, we can take $L_w \equiv L_w^*$. However, when $g_u \neq 0$, we show below that the functional $L_w([U_L, U_R])$ will not be continuous everywhere (for any choice of admissible solution of the Riemann problem) due to the divergence of the zero speed shock curve from the standing wave curves when $g_u \neq 0$. Moreover, we must modify the definition of wave strength for the triple composite standing waves, (described by the wave $U_L \to P \to Q \to R$ in Figure 15 and Figure 17, when $g_u < 0$ and $g_u > 0$, respectively) in order to insure that L_w is minimized on a triple composite standing wave. The idea is to first modify the strength of a triple composite standing wave to be equal to the strength of the two waves (a positive speed shock wave followed by a standing wave on the right when $g_u < 0$, and a standing wave on the left followed by a negative speed shock wave when $g_u > 0$) that *would* solve the same Riemann problem in the case $g_u = 0$. We call these two waves the *projection* of the triple composite wave γ_0, and label it $P(\gamma_0)$. By so changing the wave strength, we introduce a new discontinuity in the functional L_w^* that must be corrected for. Thus, to modify L_w^* into a continuous functional L_w, we must further add a compensating term $\delta(\gamma_0)$ to each standing wave γ_0 on the right, left when $g_u < 0$, $g_u > 0$, respectively. (We label a triple composite standing wave as being on the left, right of $\mathcal{T}$ according to the side of $\mathcal{T}$ on which the standing wave in $P(\gamma_0)$ falls. Thus, triple composite standing waves lie on the right, left of $\mathcal{T}$ when $g_u < 0$, $g_u > 0$, respectively.) Thus, the strategy for modifying L_w^* into a continuous functional L_w at triple composite standing waves is to redefine the strength of a triple composite standing wave $|\gamma_0| = |P(\gamma_0)|_* + \delta(\gamma_0)$, where $P(\gamma_0)$ and $\delta(\gamma_0)$ are appropriately defined below.

So assume first that $g_u < 0$. We first show that L_w^* is discontinuous under perturbation of a zero speed shock wave followed by a strong standing wave on the right of $\mathcal{T}$; cf. Figure 8. Indeed, referring to Figure 8, the elementary waves defined by $U_L \to U_M \to U_R$ and $U_L \to E \to U_R$ both must serve as admissible solutions of the Riemann problem, but $L_w^*(U_L \to U_M \to U_R) \neq L_w^*(U_L \to I \to K \to U_R) = L_w^*(U_L \to E \to U_R)$. We correct for this in the case $g_u < 0$ by modifying the definition of wave strength for strong standing waves ($u_L < u_R$) on the *right* of $\mathcal{T}$ by exactly the amount required to make L_w^* continuous.

To make this precise, let U_L and U_R denote the left and right states of a strong standing wave γ_0 on the right of $\mathcal{T}$. Let $f(a, u) = f(a_L, u_R)$ define the unique zero speed shock curve that passes through the state U_L, and for our purposes here, let U_* denote the unique point where this zero speed shock curve intersects the transition curve $\mathcal{T}$. The state $U_* = (a_*, u_*)$ is determined by the conditions $f(a_*, u_*) = f(a_L, u_L)$ and $u_* = u_{\mathcal{T}}(a_*)$; cf. Figure 9. Let $a_s(u)$ denote the unique standing wave curve that emanates from the point U_*. The curve a_s lies to the left of the standing shock curve emanating from U_* because $g_u < 0$. Now define the points I and K that lie on the standing wave curve a_s to the right of $\mathcal{T}$, at levels a_L and a_R, respectively (again see Figure 9). Since I and K are determined by γ_0 alone, we can define

$$\delta(\gamma_0) = L_w^*(I \to K \to U_R) - L_w^*(I \to U_L \to U_R) \tag{2.26}$$

for any strong standing wave γ_0 lying to the right of $\mathcal{T}$ in the case $g_u < 0$. (Note

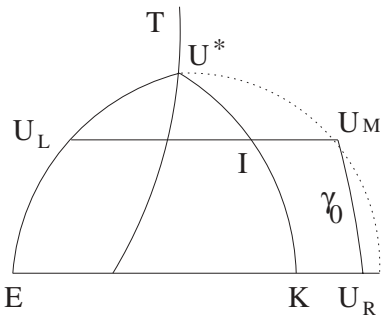


FIG. 8.

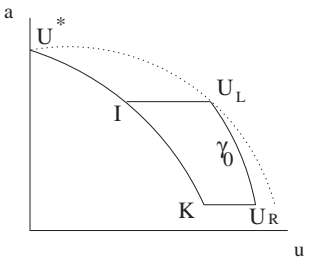


FIG. 9.

that $\delta(\gamma_0)$ depends only on U_L and U_R across the standing wave and is exactly the deficit between $L_w^*(U_L \to U_M \to U_R)$ and $L_w^*(U_L \to E \to U_R)$ encountered in Figure 8. Note also that $\delta(\gamma_0) = 0$ when $U_L \in \mathcal{T}$, because $|\gamma_0|$ reduces to $|\gamma_0|_*$ in this limit.) Thus, in the case $g_u < 0$, we define the modified strength $|\gamma_0|$ of a *strong standing wave on the right of* $\mathcal{T}$ by the rule

$$|\gamma_0|_w = |\gamma_0|_w^* + \delta(\gamma_0), \tag{2.27}$$

where $d(\gamma_0)$ is defined in (2.26).

Consider next the triple composite standing waves in the case $g_u < 0$. The main examples are given by $\gamma_0 \equiv U_L \to P \to Q \to R$ in Figures 15 and 6, the general case isolated in Figure 6. In both diagrams, $R = U_R$ denotes the right state of the triple composite standing wave γ_0. In these cases, the projection $P(\gamma_0)$ is given by $P(\gamma_0) = U_L \to T \to R$. We now show that the value of $L_w^*(P(\gamma_0))$ is discontinuous as $U_L = R$ varies from R to I along the line segment SN in Figure 6. Indeed, as $P(\gamma_0)$ varies from $U_L \to T \to R$ to $U_L \to M \to L$, the solution of the Riemann problem changes to $U_L \to S \to L$ and then to $U_L \to K \to I$. Thus for continuity, we require that $L_w(U_L \to S \to L) = L_w(U_L \to M \to L)$. But $L_w^*(U_L \to S \to L) = L_w^*(U_L \to K \to I \to L) = L_w^*(U_L \to M \to L) + \delta$, where $\delta = \delta(\gamma_0) = L_w^*(K \to I \to L) - L_w^*(K \to M \to L)$. Thus, for the general weak standing wave $U_L \to U_R$ on the right of $\mathcal{T}$ when $g_u < 0$, diagrammed in Figure 10, define

$$\delta = \delta(\gamma_0) = L_w^*(K \to I \to U_R) - L_w^*(K \to U_L \to U_R). \tag{2.28}$$

We take this as defining $\delta(\gamma_0)$ for any weak standing wave on the right of $\mathcal{T}$ that takes U_L to U_R, where for triple composite waves, (2.28) is assumed to apply to the weak standing wave on the right in $P(\gamma_0)$. (Note that the points K and I in Figure

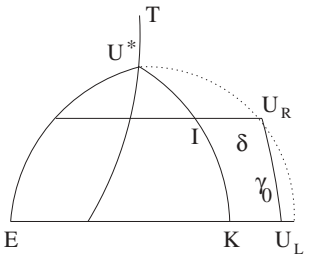


FIG. 10.

10 are determined by U_L and U_R alone.) To put this all together, let $P(\gamma_0) = \gamma_0$ for any standing wave that is not triple composite, and let $\delta(\gamma_0)$ be defined in (2.26) and (2.28) for strong and weak standing waves on the right of $\mathcal{T}$. Then we define the modified strength $|\gamma|$ of an elementary wave γ in the case $g_u < 0$ by

$$|\gamma|_w = \begin{cases} |P(\gamma)|^*_w + \delta(\gamma) & \text{if } \gamma \text{ is a standing wave on the right of } \mathcal{T}, \\ |\gamma|^*_w & \text{otherwise,} \end{cases} \tag{2.29}$$

where $d(\gamma)$ is defined in (2.26) and (2.28). For a sequence of elementary waves $\gamma_1, \ldots, \gamma_n$, we define the modified linear functional

$$L_w[\gamma_1, \ldots, \gamma_n] = \sum_{i=1}^{n} |\gamma_i|_w. \tag{2.30}$$

This completes the definition of L_w in the case $g_u < 0$. We now define the modified linear functional L_w in the case $g_u > 0$.

So assume now that $g_u > 0$. We show first that L^*_w is discontinuous under perturbation of a strong standing wave on the left of $\mathcal{T}$ followed by a zero speed shock wave; cf. Figure 11. Referring to Figure 11, we see that both $U_L \to U_M \to U_R$ and $U_L \to E \to U_R$ both solve the Riemann problem, but $L^*_w(U_L \to U_M \to U_R) = L^*_w(U_L \to K \to I \to U_R) \neq L^*_w(U_L \to E \to U_R)$. To correct for this in the case $g_u > 0$, we modify the definition of wave strength for strong standing waves on the *left* of $\mathcal{T}$ by exactly the amount required to make L^*_w continuous.

To make this precise, let U_L and U_R denote the left and right states of a strong ($u_L < u_R$) standing wave γ_0 on the left of $\mathcal{T}$. In this case, let $a_s(u)$ denote the unique standing wave curve that passes through the states U_L and U_R, and let $U_* = (a_*, u_*)$ denote the unique point at which this standing curve a_s intersects the transition curve $\mathcal{T}$. Let $f(a, u) = f(a_*, u_*)$ define the unique zero speed shock curve that passes through the state U_*, defined to the right of $\mathcal{T}$, and let $I = (a_\#, u_\#)$ denote the state on this zero speed shock curve at level a_R; cf. Figure 12. Thus, I is determined by the condition that $I > \mathcal{T}$, together with $a_\# = a_R$, and $f(a_*, u_*) = f(a_R, u_\#)$. (Note that the zero speed shock curve emanating from U_* lies to the left of the standing wave curve emanating from U_* because $g_u > 0$.) Now define the state K to be the state at level a_L on the standing wave curve through I lying on the right-hand side of the transition curve $\mathcal{T}$ on the opposite side from U_L (see Figure 12). Since I and K are determined by γ_0 alone, we can define

$$\delta(\gamma_0) = L^*_w(\bar{U}_L \to K \to I) - L^*_w(\bar{U}_L \to \bar{U}_R \to I), \tag{2.31}$$

which is defined for any strong standing wave γ_0 lying to the left of $\mathcal{T}$ in the case $g_u > 0$. (Note that this is exactly the deficit between $L^*_w(U_L \to K \to I)$ and $L^*_w(U_L \to$

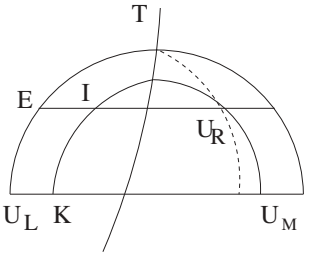


FIG. 11.

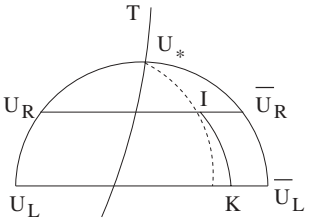


FIG. 12.

$U_R \to I$) in Figure 12. Note also that as before, $\delta(\gamma_0) = 0$ when $U_R \in \mathcal{T}$, $|\gamma_0|$ reduces to $|\gamma_0|_*$ in this limit.) Thus, in the case $g_u > 0$, we define the modified strength $|\gamma_0|$ of a strong standing wave on the left of $\mathcal{T}$ by the rule $|\gamma_0|_w = |\gamma_0|^*_w + \delta(\gamma_0)$.

Consider finally the triple composite standing wave $\gamma_0 \equiv U_L \to P \to Q \to R$ in Figure 17, isolated in Figure 13, for the case $g_u > 0$. In both diagrams, $R = U_R$ denotes the right state of the triple composite standing wave γ_0. In this case, the projection $P(\gamma_0)$ is given by $P(\gamma_0) = U_L \to T \to R$. We now show that the value of $L^*_w(P(\gamma_0))$ is discontinuous as $U_L = R$ varies from R to I along the line segment SN in Figure 13. Indeed, as $P(\gamma_0)$ varies from $U_L \to T \to R$ to $U_L \to M \to L$, the solution of the Riemann problem changes to $U_L \to M \to L$ and then to $U_L \to K \to I$. Thus, for continuity, we require that $L_w(U_L \to M \to L) = L_w(U_L \to T \to L)$. But $L^*_w(U_L \to M \to L) + \delta = L^*_w(U_L \to T \to L)$, where $\delta = \delta(\gamma_0) = L^*_w(M \to K \to I) - L^*_w(M \to L \to I)$. Thus, for the general weak standing wave $U_L \to U_R$ on the left of $\mathcal{T}$ when $g_u > 0$, diagrammed in Figure 14, define

$$\delta = \delta(\gamma_0) = L^*_w(I \to K \to \bar{U}_R) - L^*_w(I \to \bar{U}_L \to \bar{U}_R). \tag{2.32}$$

We take this as defining $\delta(\gamma_0)$ for any weak standing wave on the left of $\mathcal{T}$ that takes U_L to U_R, where for triple composite waves, (2.32) is assumed to apply to the weak standing wave on the left in $P(\gamma_0)$. (Again, note that the points K and I in Figure 14 are determined by U_L and U_R alone.) To put this together, let $P(\gamma_0) = \gamma_0$ for any standing wave that is not triple composite, and let $\delta(\gamma_0)$ be defined in (2.31) and (2.32) for strong and weak standing waves on the right of $\mathcal{T}$. Then we define the modified strength $|\gamma|$ of an elementary wave γ in the case $g_u > 0$ by

$$|\gamma|_w = \begin{cases} |P(\gamma)|^*_w + \delta(\gamma) & \text{if } \gamma \text{ is a standing wave on the left of } \mathcal{T}, \\ |\gamma|^*_w & \text{otherwise,} \end{cases} \tag{2.33}$$

where $d(\gamma)$ is defined in (2.31) and (2.32). For a sequence of elementary waves

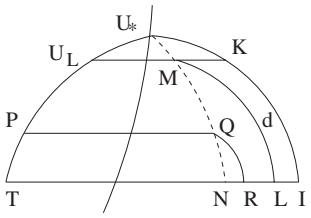


FIG. 13.

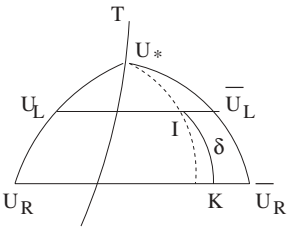


FIG. 14.

$\gamma_1, \ldots, \gamma_n$, again define the modified linear functional

$$L_w[\gamma_1, \ldots, \gamma_n] = \sum_{i=1}^{n} |\gamma_i|_w. \tag{2.34}$$

This completes the definition of L_w for the case $g_u > 0$ and so completes the definition of L_w in general.

We can now present in detail the admissible solution of the Riemann problem based on the L_w minimization principle. The solutions $[U_L, U_R]$ that are admissible by Definition 2.1 are diagrammed in Figures 15–18 for the cases $g_u < 0, g_u > 0$ and U_L to the left of $\mathcal{T}$, U_L to the right of $\mathcal{T}$. The solutions that minimize L_z are pointed out for comparison.[8] The cases $g_u < 0$ and $g_u > 0$ are qualitatively different because of the location of the zero speed shock curve. To read the diagrams, start at U_L and follow the arrows to an arbitrary state U_R. The wave curves traversed then give the elementary waves in the solution of the Riemann problem going from left to right in the (x,t)-plane. In the limit as g tends to zero, these diagrams reduce to those for the resonant homogeneous system $u_t + f(a,u)_x = 0$ [10, 12].

In Figures 15–18, the solid convex down curves denote standing wave curves, and the dotted curve to the right of $\mathcal{T}$ denotes the zero speed shock curve corresponding to the standing wave curve through U_L. In Figures 15 and 16, the dotted line falls to the right of the standing wave curve through U_L because $g_u < 0$. Similarly, in Figures 17 and 18, it falls to the left because $g_u > 0$. We discuss the multiplicity of solutions in Figures 15–17 below. In Figure 18, solutions are unique.

In each of Figures 15–17, there is a region of right states U_R for which there are multiple solutions of the Riemann problem that minimize the total variation in a.

[8] In [13] it was shown that the solutions of the Riemann problem that minimize L_z actually minimize L_z over all sequences of connected elementary waves that connect U_L to U_R. This essentially implies that L_z is nonincreasing on solutions. On the other hand, this is not the case for the solutions of the Riemann problem that minimize L_w, and this explains why a potential interaction term is required to construct a decreasing functional that incorporates L_w.

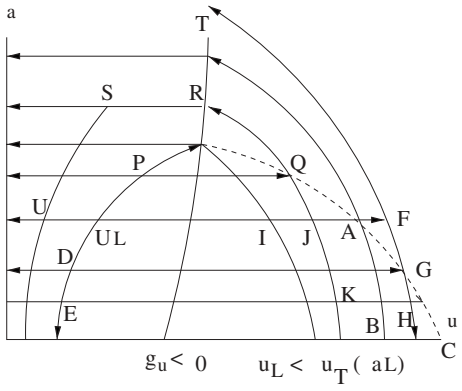


FIG. 15.

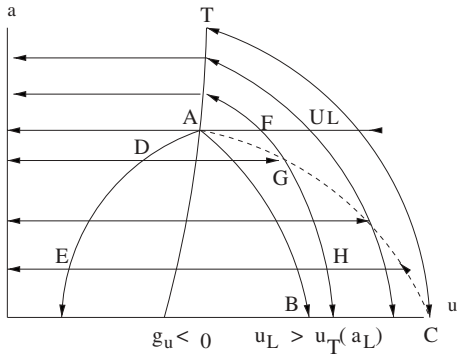


FIG. 16.

In the region of multiple solutions, there is always a multiplicity of three solution in the interior of the region, but this reduces to a multiplicity of two on the boundary of the region. The L_w minimization principle rules out every four wave solution, *except for* the two special cases labeled by $U_L \to P \to Q \to R$ in Figures 15 and 17. However, in both cases, the zero speed wave in the solution of the Riemann problem always consists of a standing wave followed by a zero speed shock wave followed by another standing wave (all zero speed) and the monotonicity in a is preserved across triple composite standing waves. From the point of view of wave interactions, such composite waves interact like a single wave, and so in our discussion below, we will treat triple composite standing waves as *a single standing wave*. With this convention, (and allowing waves to have zero strength), the admissible solution of the Riemann problem always consists of three elementary waves: a negative speed nonlinear wave followed by a single standing wave, followed by a positive speed nonlinear wave.

Discussion of Figure 15 [$g_u < 0$; U_L to the left of $\mathcal{T}$]. A multiplicity of solutions occurs when U_R lies within the interior of the region ABC, e.g., $U_R = H$. The three solutions are: $U_L \to F \to H$, $U_L \to D \to G \to H$, and $U_L \to E \to H$. (Here, e.g., $U_L \to F$ denotes the elementary shock wave taking U_L on the left to F on the right. Since F lies to the right of the zero speed shock curve (the dotted line), and since $f_{uu} < 0$, $U_L \to F$ is a shock wave of negative speed.) All of these solutions have the same L_z-value, but only the solution $U_L \to F \to H$ minimizes

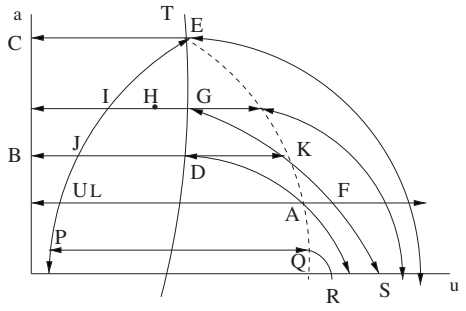


FIG. 17.

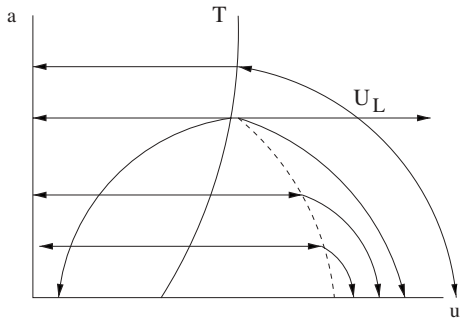


FIG. 18.

L_w. Indeed, consider region U_L, F, H, E of Figure 15, which is described in Figure 19. $L_w(U_L \to F \to H) - L_w(U_L \to E \to H) = L_w^*(U_L \to F \to H) + \delta(F \to H) - L_w^*(U_L \to E \to H) = L_w^*(U_L \to F \to H) + \delta(F \to H) - L_w^*(U_L \to I \to K \to H) = L_w^*(I \to F \to H) - L_w^*(I \to K \to H) + \delta(F \to H) = L_w^*(I \to F \to H) - L_w^*(M \to F \to H) + L_w^*(M \to N \to H) - L_w^*(I \to K \to H) = L_w^*(I \to M \to N) - L_w^*(I \to K \to N) < 0$ by Lemma 2.7. Similarly we can show that $L_w(U_L \to F \to H) - L_w(U_L \to D \to G \to H) < 0$. It follows that $U_L \to F \to H$ is the unique solution of the Riemann problem selected by the L_w minimization principle.

Discussion of Figure 16 [$g_u < 0$; U_L to the right of $\mathcal{T}$]. A multiplicity of solutions that minimize the total variation in a (but do not necessarily minimize L_z) occurs when U_R lies within the interior of the region ABC, e.g., $U_R = H$. The three solutions are $U_L \to F \to H$, $U_L \to A \to E \to H$, and $U_L \to A \to D \to G \to H$. The L_z-value is minimized only on the first of these, and thus in this case the L_z minimization principle selects a unique admissible solution. The functional L_w is also minimized on the solution $U_L \to F \to H$. For example, referring to Figure 20, $L_w(U_L \to F \to H) - L_w(U_L \to A \to E \to H) = L_w^*(F \to H) + \delta(F \to H) - L_w^*(F \to A \to B \to H)$. But $\delta(F \to H) = L_w^*(I \to K) + L_w^*(K \to H) - L_w^*(F \to H) - L_w^*(F \to I)$, $L_w^*(K \to H) - L_w^*(B \to H) = -L_w^*(B \to K)$, and $L_w^*(I \to K) \le L_w^*(A \to B)$ (by the Corollary to Lemma 2.7). Substituting these as inequalities into the previous line gives $L_w(U_L \to F \to H) - L_w(U_L \to A \to E \to H) \le -[L_w^*(F \to I) + L_w^*(F \to A) + L_w^*(B \to K)] < 0$. The case for $L_w(U_L \to F \to H) - L_w(U_L \to A \to D \to G \to H) < 0$ is similar.

Discussion of Figure 17 [$g_u > 0$; U_L to the left of $\mathcal{T}$]. A multiplicity of solutions that minimize the total variation in a occurs when U_R lies within the

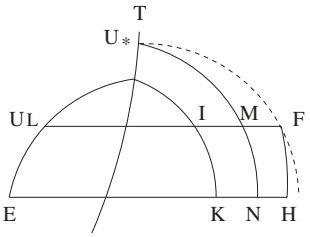


FIG. 19.

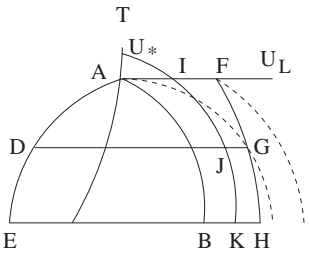


FIG. 20.

interior of the region CEADB, e.g., $U_R = H$. The three solutions are $U_L \to I \to H$, $U_L \to F \to G \to H$, and $U_L \to J \to K \to G \to H$. In this case the L_z and L_w minimization principles both pick out the unique solution $U_L \to I \to H$. Note that on the boundary, say $U_R = M$, where the wave structure changes, there is a multiplicity of *two* distinct solutions, $U_L \to I \to M$ and $U_L \to N \to M$, and at this boundary, the L_z and L_w values of both solutions are equal, a requirement for the continuity of the functionals with respect to U_L and U_R. (cf. the paragraph preceding (2.29)). As an example, we verify that $L_w(U_L \to I \to H) - L_w(U_L \to F \to G \to H) < 0$, (see Figure 21). To this end, write $L_w(U_L \to I \to H) - L_w(U_L \to F \to G \to H) = L_w^*(U_L \to I \to H) + \delta(U_L \to I) - L_w^*(U_L \to L \to G \to H) = L_w^*(U_L \to I) + L_w^*(I \to H) + [L_w^*(U_L \to P) + L_w^*(P \to Q) - L_w^*(U_L \to I) - L_w^*(I \to Q)] - L_w^*(U_L \to L) - L_w^*(L \to G) - L_w^*(G \to H) = L_w^*(P \to Q \to H) - L_w^*(P \to L \to G) - L_w^*(G \to H) < 0$ by Lemma 2.7.

Discussion of Figure 18 [$g_u > 0$; U_L to the right of $\mathcal{T}$]. In this case the solution that minimizes the total variation in a is unique.

We now summarize the main results regarding the solution of the Riemann problem.

PROPOSITION 2.10. *The admissible solution of the Riemann problem $[U_L, U_R]$ always consists of a sequence of three connected waves, a negative speed nonlinear wave γ_1 followed by a standing wave γ_0 followed by a positive speed nonlinear wave γ_2, where we allow $\gamma_i = 0$, and we treat the composite zero speed waves of type $U_L \to P \to Q \to R$ in Figure 15 as a single wave γ_0. We write*

$$[U_L, U_R] = \gamma_1\gamma_0\gamma_2. \tag{2.35}$$

PROPOSITION 2.11. *The functional $L_w([U_L, U_R])$ is a continuous function of U_L and U_R throughout the domain Ω.*

PROPOSITION 2.12. *The convex side, (i.e., lower side when $f_{uu} < 0$), of each standing wave curve is an invariant region for admissible solutions of the Riemann*

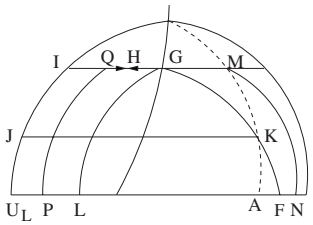


FIG. 21.

problem.

Proof. Proposition 2.10 is clear by construction. The proof of Propositions 2.11 has been indicated above and follows directly via a case by case inspection of the Riemann problem. Note that the continuity of L_w follows upon verifying that the only places where $L_w^*([U_L, U_R])$ is discontinuous were identified in the discussion following Corollary 2.8, and the deficit was accounted for by the correction term $\delta(\gamma)$. Proposition 2.12 follows immediately from the fact that all intermediate states in the admissible solution of the Riemann problem lie on the convex side of the outer of the two standing wave curves through U_L and U_R. □

Proposition 2.12 implies an L^∞ bound on Glimm approximate solutions generated by the admissible solution of the Riemann problem.

3. The generalized Glimm method. In this section we construct the approximate Glimm scheme solutions $U_{\Delta x}(x,t)$ and prove the compactness of approximate solutions under the assumption that the initial data is of bounded total variation in (a,z) (which implies that the initial data is of bounded total variation in (a,w)). We call this a generalized Glimm scheme because the standing waves are generalized weak solutions of system (1.1) due to the presence of the a' in the term $a'g$ on the right-hand side of (1.1). The proof of convergence of the residual must be modified because a piecewise constant approximation of $a(x)$ does not give an L^∞ approximation of $a'g$, and so the residual does not converge strongly, but rather weakly. This argument was first given in the strictly hyperbolic case in [6], and for completeness, we include the argument adapted to the problem here (in section 6).

To begin, assume that the initial data $U_0(x)$ takes values in a neighborhood Ω which lies below a standing wave curve and above a curve $a = \text{const}$ contained within a neighborhood of U_*, where the unique solution of the Riemann problem exists as constructed in the previous section, and small enough so that Lemmas 2.4–2.7 and Propositions 2.10–2.12 hold throughout Ω. Since Ω is an invariant region for Riemann problems, it follows that Ω is also an invariant region for the Glimm scheme, which is therefore defined for all time for any mesh length. To construct the approximate solutions, first discretize $R \times [0,\infty)$ by spatial mesh length Δx and time mesh length Δt such that

$$\frac{\Delta x}{\Delta t} = \lambda, \tag{3.1}$$

where

$$\lambda \equiv 2 \sup_{(a,u)\in\Omega} \left\{ \left| \frac{\partial f}{\partial u} \right| \right\}. \tag{3.2}$$

We let $x_i = i\Delta x$, $t_j = j\Delta t$ so that (x_i, t_j) denote the mesh points of the approximate

solution. Define

$$S_j = \{(x,t) : t_j \le t < t_{j+1}\}.$$

The approximate solution $U_{\Delta x}$ generated by the Glimm scheme is defined as follows. First, fix a sample sequence $\theta = \{\theta_{ij}\} \in \Theta$, where Θ denotes the infinite product of intervals $(0,1)$ indexed by mesh points (with Lebesgue measure) so that $\Theta = \Pi(0,1)_{ij}$ and $\theta_{ij} \in (0,1)$, $-\infty < i < \infty$, $j \ge 0$ [4, 24]. (We randomize in space and time to facilitate the proof of convergence of the residual; cf. [25]). To initiate the scheme at $j = 0$, approximate the initial data by piecewise constant states by setting

$$U_i^0 = U_{\Delta x}(x,0) = U_0(x_i + \theta_{i0}\Delta x), \quad x_i < x < x_{i+1}. \tag{3.3}$$

Assuming that $U_{\Delta x}(x,t)$ has been constructed for $(x,t) \in \bigcup_{j=0}^{j-1} S_j$, then define $U_{\Delta x}$ in S_j as the solution of (1.1) with the initial values

$$U_i^j = U_{\Delta x}(x,t_j+) = U_{\Delta x}(x_i + \theta_{ij}\Delta x, t_j-), \quad x_i < x < x_{i+1}. \tag{3.4}$$

In other words, at each time t_j, a piecewise constant approximation $U_{\Delta x}(x,t_j+)$ is obtained by sampling the solution $U_{\Delta x}(x,t_j-)$ in each interval of the mesh at time level t_j, so that the solution in S_j can be constructed by solving the Riemann problems $[U_{i-1}^j, U_i^j]$ posed at each point of discontinuity $(x_i,t_j), i \in Z$. The Courant–Friedrichs–Levy restriction (3.1) ensures that the Riemann problem solutions in each S_j do not interact before time t_{j+1} [13].

We need to define the I-curves for the analysis of the nonlocal functional F defined below; cf. [4]. An I-curve J is a continuous space-like piecewise linear curve in the (x,t)-plane that connects adjacent mesh points of type $(x_i + \theta_j\Delta x, t_j)$ to ones of type $(x_i, t_{j+1/2})$, where $(x_i, t_{j+1/2}) = (i\Delta x, (j+1/2)\Delta t)$. Given an I-curve J_1 that extends from $i = -\infty$ to $i = +\infty$, we obtain a successor J_2 of J_1 by lifting the point $(x_i, t_{j-1/2})$ to the point $(x_i, t_{j+1/2})$ when the points $(x_{i-1} + \theta_j\Delta x, t_j)$ and $(x_i + \theta_j\Delta x, t_j)$ both lie on J_1. We call the region $(x_i, t_{j-1/2}), (x_i, t_{j+1/2}), (x_{i-1} + \theta_j\Delta x, t_j), (x_i + \theta_j\Delta x, t_j)$ between J_1 and J_2 the interaction diamond Δ. We let J^j denote the I-curve that contains all of the sample points $(x_i + \theta_j\Delta x, t_j)$ at time level t_j. The I-curve J^j crosses all of the waves in the Riemann problems posed in $U_{\Delta x}$ at time level t_j, and the I-curve J^j can be obtained by a sequence of successive I-curves. (Note that lifting the mesh point $(x_i + \theta_j\Delta x, t_j)$ to $(x_i + \theta_j\Delta x, t_{j+1})$ when mesh points $(x_{i-1}, t_{j+1/2})$ and $(x_i, t_{j+1/2})$ both lie on J, does not change the waves that J crosses, and so we can consider these to be equivalent.) It follows that to show that a functional F satisfies $F(J^j) \le F(J^0)$, it suffices only to prove that $F(J_2) \le F(J_1)$ for any pair of successive mesh curves J_1 and J_2 [4].

We have the following theorem; cf. [13].

THEOREM 3.1. *If the neighborhood Ω containing U_* is chosen to be small enough, then the Glimm approximate solutions $U_{\Delta x}(x,t)$ are defined for all time. Moreover,*

$$L_z(J^{j+1}) \le L_z(J^j) \tag{3.5}$$

for each $j \ge 0$, where J^j identifies the sequence of elementary waves appearing in the approximate solution $U_{\Delta x}$ in the strip S_j, and L_z is defined in (2.14).

Proof. The proof of (3.5) was given in [13]. The supnorm bound on solutions follows from Proposition 2.11 which asserts the existence of convex invariant regions for Riemann problems in a neighborhood of U_*. The main point in the proof of (3.5)

is that the solutions of the Riemann problems used in the construction of the Glimm approximate solutions are admissible solutions of the Riemann problem, and so were selected to minimize the L_z-value of the elementary waves among all possible solutions of the Riemann problem. But L_z has the further property of being minimized on the solution of the Riemann problem among all connected sequences of elementary waves that take U_L to U_R. (This was proven in [13].) Using this, estimate (3.5) follows because L_z decreases across any interaction diamond Δ_{ij} lying between the two successive I-curves J_1 and J_2 with interaction diamond centered on (x_i, t_j). Indeed, the Glimm scheme replaces the sequence of waves that take U^j_{i-1} to U^j_i at time level t_j- by the waves that solve the Riemann problem $[U^j_{i-1}, U^j_i]$ at $t = t_j +$. □

Theorem 3.1 leads directly to the following compactness result for approximate solutions generated by the Glimm method.

THEOREM 3.2. *Assume that the initial data $U_0(x) \in \Omega$ satisfies the condition $\mathrm{Var}_z\{U_0(\cdot)\} = V_z < \infty$ and $\mathrm{Var}\{a(\cdot)\} = V_a < \infty$. Then $U_{\Delta x}(x,t) \in \Omega$ for all $x, t \geq 0$, $\mathrm{Var}_z\{U_{\Delta x}(\cdot,t)\} < 4V_z$ for all $t \geq 0$, and a subsequence of $\{U_{\Delta x}\}$ converges boundedly, almost everywhere, to a bounded measurable function $U(x,t)$ as Δx tends to zero.*

Proof. See Theorem 3.2 [18]. □

From here on out we assume that $U_{\Delta x}(x,t)$ is a sequence of Glimm approximate solutions that converges boundedly, pointwise almost everywhere to a function $U(x,t)$, and satisfies the estimate

$$\mathrm{Var}_z\{U_{\Delta x}(\cdot,t)\} < 4V_z. \tag{3.6}$$

In section 6 we conclude this argument by showing that the limit function $U(x,t)$ is a classical weak solution of (1.1) when a' has no delta function singularities.

4. The interaction potential $d(\gamma_0, \gamma_r)$. Assume that $U_{\Delta x}(x,t)$ is an approximate Glimm scheme solution starting from initial data $U_0(x)$ of bounded total variation in (a,u) and hence (a,w) as well. Then the total variation in (a,z) of $U_{\Delta x}(\cdot,0)$ is uniformly bounded, and hence the existence theory of section 3 applies. Thus, without loss of generality, we can assume that $U_{\Delta x} \to U$, where $U(x,t)$ is a weak solution of (1.1) of bounded total variation in z at each fixed time. (The convergence is in L^1_{loc} at each fixed time, uniformly on compact sets.) We now estimate the growth of the total variation in w (and hence in u) in the approximate solutions $U_{\Delta x}(x,t)$.

Our idea is to use the functional L_w to estimate the total variation in w at each time in an approximate solution $U_{\Delta x}(x,t)$. The problem of estimating L_w is more difficult than the problem of estimating L_z because in the case of L_w, it is *not* true that $L_w(J_{j+1}) \leq L_w(J_j)$ across interactions. The point is that L_w is minimized on the admissible solution of the Riemann problem among all solutions of the Riemann problem, but it is *not* minimized on the admissible solution of the Riemann problem among all connected sequences of elementary waves that take U_L to U_R, even if there is just a single standing wave within the sequence. Indeed, if a fast rarefaction wave followed by a slow standing wave interacts to produce a slow standing wave followed by a fast rarefaction wave, then L_w increases across this interaction. This is because rarefaction wave-standing wave interactions, in which incoming and outgoing waves all lie on one side of $\mathcal{T}$, always have the effect of moving the standing wave closer to the transition curve—this increases the L_w because it shifts the total variation in u from the nonlinear waves to the standing waves, which are weighted by the larger factors of 2 and 4 over the weight on the nonlinear waves. We verify this in two examples below. The remarkable fact that the functional L_w increases *only* on rarefaction

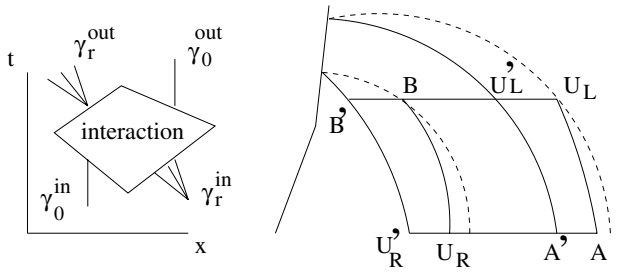


FIG. 22.

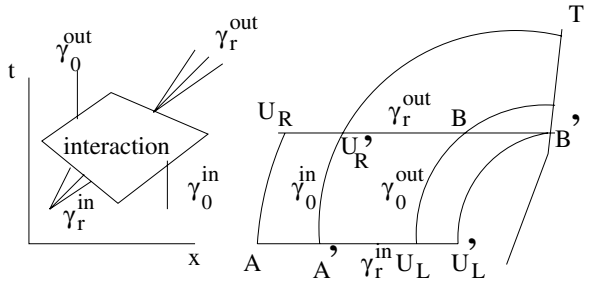


FIG. 23.

wave/standing wave interactions, and is nonincreasing on *all other* interactions, is discussed after the examples. Our strategy is then to define a potential for the increase in L_w due to the interaction of a standing wave and a rarefaction wave and to prove that L_w plus the sum of all potential interaction terms define a nonlocal functional F that bounds the total variation in w and decrease across interactions.

We begin by verifying that L_w increases on rarefaction wave/standing wave interactions in two salient examples: the case when $g_u < 0$ and the standing wave is a strong standing wave on the right of $\mathcal{T}$, and the case when $g_u > 0$ and the standing wave is a strong standing wave on the left of $\mathcal{T}$. These two examples clarify the problem of bounding the increase in L_w on interactions. So consider first the interaction diagrammed in Figure 22, the case when $g_u < 0$, and a standing wave γ_0^{IN} starting to the left of a negative speed rarefaction wave γ_r^{IN} interacts to produce a negative speed rarefaction wave γ_r^{OUT} followed by a standing wave γ_0^{OUT}. (For simplicity we assume in this example that all waves lie to the right of $\mathcal{T}$.) Then $[U_L, U_R] = (\gamma_r^{OUT}, \gamma_0^{OUT})$, but $L_w(\gamma_r^{OUT}, \gamma_0^{OUT}) - L_w(\gamma_r^{IN}, \gamma_0^{IN}) = L_w^*(\gamma_r^{OUT}) + L_w^*(\gamma_0^{OUT}) + \delta(\gamma_0^{OUT}) - L_w^*(\gamma_r^{IN}) - L_w^*(\gamma_0^{IN}) - \delta(\gamma_0^{IN}) = L_w^*(U_L \to B \to U_R) + \delta(\gamma_0^{OUT}) - L_w^*(U_L \to A \to U_R) - \delta(\gamma_0^{IN}) = L_w^*(U_L' \to B' \to U_R') - L_w^*(U_L' \to A' \to U_R') > 0$ by Lemma 2.7.

Consider next the case of the interaction diagrammed in Figure 23, the case when $g_u > 0$, and a positive speed rarefaction wave γ_r^{IN} starts to the left of a standing wave γ_0^{IN} and interacts to produce a standing wave γ_0^{OUT} followed by (that is, to the left of) a positive speed rarefaction wave γ_r^{OUT}. (Again, for simplicity, we assume in this example that all waves lie to the left of $\mathcal{T}$.) Then $[U_L, U_R] = (\gamma_0^{OUT}, \gamma_r^{OUT})$, but $L_w(\gamma_0^{OUT}, \gamma_r^{OUT}) - L_w(\gamma_0^{IN}, \gamma_r^{IN}) = L_w^*(\gamma_0^{OUT}) + \delta(\gamma_0^{OUT}) + L_w^*(\gamma_r^{OUT}) - L_w^*(\gamma_0^{IN}) -$

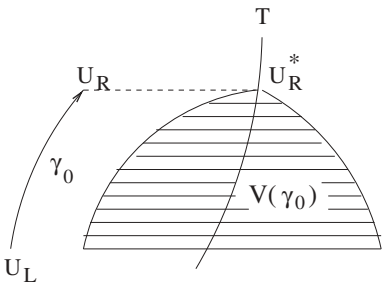


FIG. 24.

$\delta(\gamma_0^{IN}) - L_w^*(\gamma_r^{IN}) = L_w^*(U_L \to B \to U_R) + \delta(\gamma_0^{OUT}) - L_w^*(U_L \to A \to U_R) - \delta(\gamma_0^{IN}) = L_w^*(U_L' \to B' \to U_R') - L_w^*(U_L' \to A' \to U_R') > 0$ by Lemma 2.7. One can verify that L_w is nonincreasing on shock wave-standing wave interactions that lie on one side of $\mathcal{T}$ by similar examples. This concludes the examples.

What is remarkable is that the increase in L_w due to rarefaction wave-standing wave interactions that are not transonic (that is, all waves in the interaction lie entirely on the same side of $\mathcal{T}$) accounts for *all* of the ways L_w can increase, even for complicated transonic wave interactions that carry waves across the transition curve. The proof that we need only a potential interaction term for nontransonic rarefaction wave-standing wave interactions is a consequence of our proof below that the nonlocal functional F is nonincreasing on all interactions, but in the proof it is difficult to see the reason for the decrease in the functional in the complicated case when the interactions are transonic. To motivate the argument, consider a standing wave γ_0 that takes $U_L = (a_L, u_L)$ to $U_R = (a_R, u_R)$. Then this wave lies entirely on one side of $\mathcal{T}$, or else it is a composite wave of type $U_L \to P \to Q \to R$ of Figure 15. Let $a_* = \max\{a_L, a_R\}$, and let $U_* = (a_*, u_*)$ denote the point on $\mathcal{T}$ that lies at level $a = a_*$. Consider now the region $V(\gamma_0)$ that lies below the standing wave curves on the left and right of $\mathcal{T}$ that pass through the state $U = U_*$; cf. Figure 24. The claim then is that any rarefaction wave that lies in the region $V(\gamma_0)$ in an approximate solution that contains the wave γ_0 cannot interact with γ_0 in such a way as to produce an increase in L_w. For example, one can verify that when the connected sequence of waves $\gamma_r\gamma_0$ or $\gamma_0\gamma_r$ interact to produce the waves in the Riemann problem $[U_L, U_R]$, L_w will be nonincreasing and the wave γ_r will be eliminated by the interaction when γ_r in $V(\gamma_0)$. This helps explain why we needn't include such portions of the rarefaction wave in the definition of the interaction potential $d(\gamma_0, \gamma_r)$ below.

We now define $d(\gamma_0, \gamma_r)$, the potential for the increase in L_w due to the interaction of a standing wave γ_0 that *approaches* a rarefaction wave γ_r; cf. [4]. (Although there is an ordering of the waves in the (x,t)-plane implied by the condition that two waves approach, we assume no ordering in d, so that $d(\gamma_0, \gamma_r) \equiv d(\gamma_r, \gamma_0)$.) So assume that γ_0 and γ_r are waves that cross the same I-curve J in an approximate Glimm scheme solution $U_{\Delta x}$. We say that γ_0 and γ_r *approach* on J if the faster of the two waves is positioned to the left of the slower wave on J in the (x,t)-plane. Any two such waves will interact at a later time in the approximate solution $U_{\Delta x}$. Note that standing waves always have zero speed, and to make the definition of approaching unambiguous, assume that all rarefaction waves have purely positive or negative speed by treating any rarefaction wave that crosses $\mathcal{T}$ as two separate waves by partitioning such a rarefaction wave into its positive and negative speed parts. (In this case, the wave

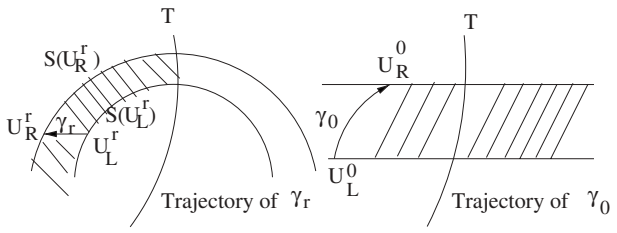


FIG. 25.

will be partitioned at the point where the wave crosses $\mathcal{T}$ since this is the curve of zero characteristic speed.) If α and β are indices that identify two waves that cross J, then we write $(\alpha, \beta) \in App(J)$ if γ^α approaches γ^β on J.

In order to define $d(\gamma_0, \gamma_r)$ for two approaching waves γ_0 and γ_r, we first define what we call the interaction region $\Delta(\gamma_0, \gamma_r)$, the region in U-space where the interaction of γ_r and γ_0 will take place (assuming the rarefaction wave is not canceled out before the interaction occurs). To this end, we first define what we call the *trajectory* of the waves γ_r and γ_0. If the waves interact, then the interaction will occur within the region determined by the intersection of the two trajectories. Since the standing wave curves and nonlinear wave curves act like Riemann invariants for the system (1.1), it follows that when a rarefaction wave interacts with a standing wave, the standing wave is just translated along the nonlinear wave curves and the rarefaction wave is translated along the standing wave curves. Thus let $\gamma_0 = [U_L^0, U_R^0]$ and $\gamma_r = [U_L^r, U_R^r]$ denote a standing wave and a rarefaction wave, respectively. In the case when the standing wave γ_0 is a composite wave of type $U_L \to P \to Q \to R$ of Figure 15, we define

$$d(\gamma_0, \gamma_r) = d(\gamma_0', \gamma_r), \tag{4.1}$$

where γ_0' denotes the standing wave in the projection $P(\gamma_0)$ (e.g., $\gamma_0' = \mathcal{T} \to R$ in Figure 15). Thus to define $d(\gamma_0, \gamma_r)$, it suffices to assume that the standing wave γ_0 lies entirely on one side of $\mathcal{T}$ (admissible, noncomposite standing waves do not cross the transition curve), and we can assume that the rarefaction wave γ_r lies entirely on one side of $\mathcal{T}$ because rarefaction waves are partitioned so as to have unambiguous positive or negative speed. For the rarefaction wave γ_r let $\mathcal{S}(U_L^r), \mathcal{S}(U_R^r)$ denote the standing wave curves that pass through states U_L^r, U_R^r, respectively. We can now define the *trajectory* of a rarefaction wave γ_r and a standing wave γ_0; cf. Figure 25.

DEFINITION 4.1. *Let $\gamma_r = [U_L^r, U_R^r]$ denote a rarefaction wave that lies entirely one side of the transition curve, say $\gamma_r \le \mathcal{T}$ or $\gamma_r \ge \mathcal{T}$. Then the* trajectory $Traj(\gamma_r)$ *of γ_r is the region in U-space between the two standing wave curves $\mathcal{S}(U_L^r)$ and $\mathcal{S}(U_R^r)$, intersected with $u \le \mathcal{T}$ or $u \ge \mathcal{T}$, according to whether $\gamma_r \le \mathcal{T}$ or $\gamma_r \ge \mathcal{T}$, respectively.*

DEFINITION 4.2. *The trajectory $Traj(\gamma_0)$ of a standing wave γ_0 is the region between the curves $a = a_L^0$ and $a = a_R^0$, i.e., the region between the two nonlinear wave curves through U_L^0 and U_R^0, respectively.*

We note that $Traj(\gamma_r)$ includes only the region on the side of $\mathcal{T}$ that contains the wave γ_r because a rarefaction wave cannot cross $\mathcal{T}$ without being canceled out by a shock wave, but $Traj(\gamma_0)$ contains the region on both sides of $\mathcal{T}$ because a standing wave can cross $\mathcal{T}$ as a result of interaction. It follows that the interaction of γ_0 and γ_r can only take place on the side of $\mathcal{T}$ that contains γ_r.

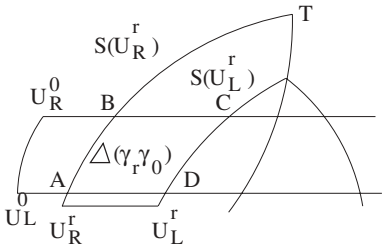


FIG. 26.

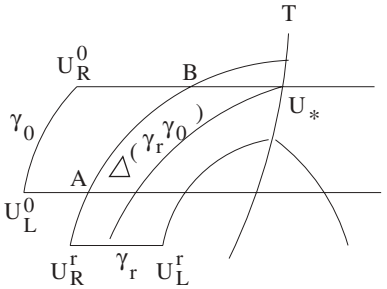


FIG. 27.

We now define the interaction region $\Delta(\gamma_0, \gamma_r)$. To this end, consider the region equal to the intersection between $Traj(\gamma_r)$ and $Traj(\gamma_0)$, which lies entirely on the same side of $\mathcal{T}$ as the rarefaction wave γ_r. If there exists a full set of four intersection points between the curves $\mathcal{S}(U_L^r), \mathcal{S}(U_R^r)$ and $a = a_L^0$, $a = a_R^0$ that all lie on the same side of $\mathcal{T}$ as the wave γ_r (diagrammed A, B, C, D in Figure 26), then define the interaction region $\Delta(\gamma_0, \gamma_r)$ to be the region $ABCD$, which is exactly equal to the intersection of the trajectory of γ_0 and the trajectory of γ_r. If the curves $\mathcal{S}(U_L^r), \mathcal{S}(U_R^r)$ and $a = a_L^0$, $a = a_R^0$ do not intersect in four distinct points on the same side of $\mathcal{T}$ as γ_r, we must modify the definition of $\Delta(\gamma_0, \gamma_r)$ to account for the fact that portions of the rarefaction wave γ_r will be canceled out before γ_0 can interact with the standing wave γ_0. To this end, let U_* denote the highest point on $\mathcal{T}$ where the trajectory of γ_0 intersects $\mathcal{T}$, i.e., let $U_* = (a_{max}, u_{\mathcal{T}}(a_{max}))$, where $a_{max} = \max\{a_L^0, a_R^0\}$; see Figure 27. Consider then the standing wave $\mathcal{S}(U_*)$ that passes through the point U_*, and ask whether $\mathcal{S}(U_*)$ lies within the trajectory of γ_r. If it does not (which means the trajectory of γ_r lies below U_*), then we say that the interaction region $\Delta(\gamma_0, \gamma_r) = \phi$, the empty set; that is, there is no potential for interaction of the waves γ_r and γ_0. If $\mathcal{S}(U_*)$ does lie within the trajectory of γ_r, then let $\Delta(\gamma_0, \gamma_r)$ denote the intersection of the trajectory of γ_0 with the trajectory of γ_r and *take away* all points U that lie below the standing wave curve $\mathcal{S}(U_*)$. In this case, $\Delta(\gamma_0, \gamma_r) = ABU_*D$, as diagrammed in Figure 27. This completes the definition of $\Delta(\gamma_0, \gamma_r)$. Note that in every case, $\Delta(\gamma_0, \gamma_r)$ consists of a region on the side of the transition curve that contains γ_r, bounded on the right and left by standing wave curves and above and below by nonlinear wave curves, determined by four vertices, which we label $ABCD$ as in Figure 28.

Now for any approaching waves γ_r and γ_0 (assuming rarefaction waves are partitioned at points where they cross $\mathcal{T}$), define $d(\gamma_0, \gamma_r)$ in terms of $\Delta(\gamma_0, \gamma_r)$ as follows.

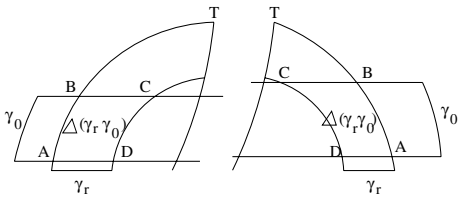


FIG. 28.

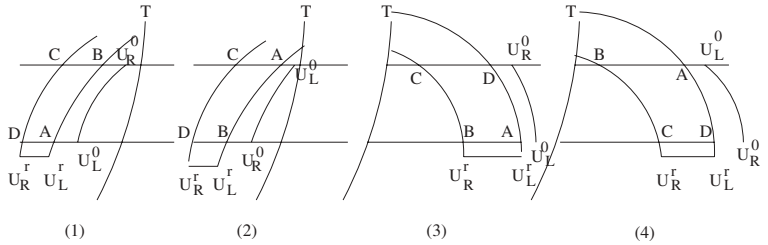


FIG. 29.

The interaction potential $d(\gamma_0, \gamma_r)$ is equal to the change in L_w between the waves that enter and the waves that leave the interaction region $\Delta(\gamma_0, \gamma_r)$, as determined by the orientation of the original waves γ_0 and γ_r. That is, there is only one way to project the waves γ_r and γ_0 to incoming waves on the boundary of $\Delta(\gamma_0, \gamma_r)$ so that γ_r is projected to a rarefaction wave, γ_0 is projected to a standing wave that preserves the increasing/decreasing of a across the wave, and the projected waves define a connected sequence of waves that preserve the left/right orientation of the original waves γ_r and γ_0. Thus, there are four cases in which γ_r and γ_0 can approach, labeled in Figure 29. These are determined by whether a increases or decreases across the standing wave γ_0 and whether the wave γ_r lies to the left or right of $\mathcal{T}$. In the four cases (1)–(4) labeled in Figure 29, $d(\gamma_0, \gamma_r)$ in each case is defined by

$$d(\gamma_0, \gamma_r) = L_w(A \to B \to C) - L_w(A \to D \to C). \tag{4.2}$$

Therefore, assuming that all rarefaction waves have been partitioned at points on $\mathcal{T}$, equation (4.2) defines $d(\gamma_0, \gamma_r)$ for any pair of approaching waves γ_r and γ_0, and we set $d(\gamma_0, \gamma_r) = 0$ for any pair of nonapproaching waves. For our arguments below, we wish to index the waves in an approximate Glimm scheme solution as they are given in the solution of the Riemann problems themselves, without further partitioning. Thus for a rarefaction wave γ_r that crosses $\mathcal{T}$ and is partitioned into $\gamma_r = \gamma_r^a \gamma_r^b$ at the point where it crosses $\mathcal{T}$, we say γ_r approaches a standing wave γ_0 if γ_r^a approaches γ_0 or γ_r^b approaches γ_0, and we define $d(\gamma_0, \gamma_r) = d(\gamma_0, \gamma_r^a) + d(\gamma_0, \gamma_r^b)$. It follows that for *any partitioning* of a rarefaction wave $\gamma_r = \gamma_r^a \cdots \gamma_r^b$, we have that $d(\gamma_0, \gamma_r) = d(\gamma_0, \gamma_r^a) + \cdots + d(\gamma_0, \gamma_r^b)$. This completes the definition of $d(\gamma_0, \gamma_r)$ for any pair of waves γ_0, γ_r that crosses an I-curve J in an approximate Glimm scheme solution of (1.1).

We note that the potential $d(\gamma_0, \gamma_r)$ is symmetric, $d(\gamma_0, \gamma_r) = d(\gamma_r, \gamma_0)$, and is constructed so that if a standing wave γ_0 is displaced to $\hat{\gamma}_0$ by interaction with a nonlinear wave and a rarefaction wave γ_r is displaced to $\hat{\gamma}_r$ by interaction with a standing wave, then (assuming no cancellation of shock and rarefaction waves) $d(\gamma_0, \gamma_r) = d(\hat{\gamma}_0, \hat{\gamma}_r)$. Thus, $d(\gamma_0, \gamma_r)$ is invariant under such interactions even though $|\gamma_0|_w \neq |\hat{\gamma}_0|_w$ and

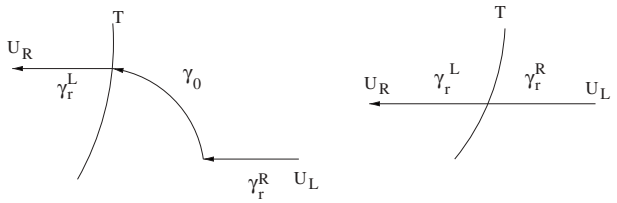


FIG. 30.

$|\gamma_r|_w \neq |\hat{\gamma}_r|_w$. Remarkably, this statement holds even when γ_0 is a composite wave of form $U_L \to P \to Q \to R$ of Figure 15. Therefore, even though wave strengths change as waves evolve in the solution, the potential interaction between waves is constructed so as to be an invariant of interactions (assuming no cancellation of rarefaction waves by shock waves).

The following proposition gives the main property that tells how rarefaction wave trajectories change when waves interact. To state the proposition, note that the rarefaction waves in any admissible solution of the Riemann problem $[U_L, U_R]$ can always be partitioned into a positive speed rarefaction wave γ_r^L on the left of $\mathcal{T}$ and a negative speed rarefaction wave γ_r^L on the right of $\mathcal{T}$. The fact that one can always uniquely identify exactly two such waves for every choice of U_L and U_R (allowing for one or both of the waves to be zero) can be verified directly in Figures 15–18. (Recall that a rarefaction wave cannot cross $\mathcal{T}$ in any admissible solution of the Riemann problem unless the standing wave is zero, in which case the solution is a single rarefaction wave γ_r that can be partitioned into $\gamma_r \sim \gamma_r^R \gamma_r^L$; cf. Figure 30.)

PROPOSITION 4.3. *Let γ_r^L and γ_r^R denote the left and right rarefaction waves in the solution of the Riemann problem $[U_L, U_R]$, and let $\bar{\gamma}_1\bar{\gamma}_0\bar{\gamma}_2$ be any connected sequence of elementary waves that takes U_L to U_R such that $\bar{\gamma}_1$, $\bar{\gamma}_2$ are nonlinear waves and $\bar{\gamma}_0$ is a standing wave. Then $Traj(\gamma_r^L) \subseteq Traj(\bar{\gamma}_r^L)$ and $Traj(\gamma_r^R) \subseteq Traj(\bar{\gamma}_r^R)$, where $\bar{\gamma}_r^L$ and $\bar{\gamma}_r^R$ denote the union of all left and right rarefaction waves, respectively, among $\bar{\gamma}_i$, i=1, 2.*

Proof. The proof of Proposition 4.3, which can be verified case by case from the admissible solution of the Riemann problem, is postponed until the appendix. ∎

5. The nonlocal functional. In this section we define the nonlocal functional $F(J)$ that bounds the total variation in w for the waves that cross an I-curve J in an approximate Glimm scheme solution. We then prove that F is nonincreasing on approximate solutions. To start, let J denote a fixed I-curve, and for notational convenience let Λ be an index set such that γ_q^α, $\alpha \in \Lambda$, $q \in \{0, r, s\}$, lists all of waves that cross J. Here $q = 0, r, s$ means that the wave is a standing wave, rarefaction wave, or shock wave, respectively, so that, for example, $\{\gamma_0^\alpha\}_{\alpha\in\Lambda}$ denotes the set of all standing waves that cross J, etc. (To achieve such an indexing just allow for arbitrarily many zero waves.) Thus the local functional $L_w(J)$ is defined by

$$L_w(J) = \sum_{\alpha\in\Lambda, q\in\{0,r,s\}} |\gamma_q^\alpha|_w. \tag{5.1}$$

Define the functional $F(J)$ by

$$F(J) = L_w(J) + P(J), \tag{5.2}$$

where $P(J)$ is the nonlocal potential interaction functional defined by

$$P(J) = \sum_{(\alpha,\beta)\in App(J)} d(\gamma_0^\alpha, \gamma_r^\beta), \tag{5.3}$$

where we use the notation $(\alpha, \beta) \in App(J)$ if and only if γ_0^α approaches γ_r^β on J. Since $F(J)$ is determined by the connected sequence of waves that cross J, we can similarly define $F(\gamma_a \cdots \gamma_b)$ for any connected sequence of elementary waves. (Two waves in the sequence *approach* if the left wave is faster than the right wave in the sequence, etc.) Then, for example, $F([U_L, U_R]) = L_w([U_L, U_R])$ because the solution of the Riemann problem contains no pairs of approaching waves. We now prove the following theorem.

THEOREM 5.1. *If J_2 is a successor of J_1 in an approximate Glimm scheme solution of system* (1.1)*, then*

$$F(J_2) - F(J_1) \le 0. \tag{5.4}$$

The proof of Theorem 5.1 is a consequence of the following lemma. (We think of bar, tilde, and hat as identifying incoming waves and unbarred waves as representing outgoing waves, and $[U_L, U_R]$ denotes the admissible solution of the Riemann problem, where the strength of each zero speed composite wave γ_0 has a strength equal to the strength of the waves in its projection $P(\gamma_0)$.)

PROPOSITION 5.2. *Let $\bar\gamma_1\bar\gamma_0\bar\gamma_2$ denote any connected sequence of three elementary waves that takes U_L to U_R such that $\bar\gamma_1, \bar\gamma_2$ are nonlinear waves, and $\bar\gamma_0$ is a standing wave.*

$$F(\gamma_1\gamma_0\gamma_2) \le F(\bar\gamma_1\bar\gamma_0\bar\gamma_2), \tag{5.5}$$

where $\gamma_1\gamma_0\gamma_2 = [U_L, U_R]$.

The proofs of Propositions 4.3 and 5.2 involve a case by case study of the Riemann problem and will be dealt with together in the appendix. Assuming Propositions 4.3 and 5.2, we now give the following proof.

Proof of Theorem 5.1. Assume that Propositions 4.3 and 5.2 hold, and assume that J_2 is an immediate successor of J_1 in the approximate Glimm scheme solution $U_{\Delta x}$ of system (1.1). We show that $F(J_2) \le F(J_1)$. Let Δ denote the interaction diamond between J_1 and J_2, let J_1', J_2' denote the restriction of J_1, J_2 to the region Δ, respectively, and let J_0 denote the restrictions of J_1 and J_2 to the region *outside* Δ; cf. [4, 24]. Thus we write $J_1 = J_0 \cup J_1'$ and $J_2 = J_0 \cup J_2'$. Note that since we use an unstaggered grid, the states U_L and U_R that lie at the right and left vertices of Δ are consecutive sample points at some time level t_j in the approximate solution $U_{\Delta x}$, and thus there is at most one standing wave between U_L and U_R on both J_1' and J_2'. It follows that there are at most five incoming waves that cross J_1', i.e., at most two nonlinear waves $\bar\gamma_1^a$ and $\bar\gamma_1^b$, followed by a standing wave $\bar\gamma_0$, followed by at most two nonlinear waves $\bar\gamma_2^a\bar\gamma_2^b$. (Subscripts $1, 2$ denote nonlinear waves, and subscript 0 denotes a standing wave.) Thus let $\bar\gamma_1^a\bar\gamma_1^b\bar\gamma_0\bar\gamma_2^a\bar\gamma_2^b$ denote the connected sequence of elementary waves that take U_L to U_R and cross the curve J_1', the incoming waves for the interaction diamond Δ. The waves that leave the interaction diamond Δ cross J_2' and hence solve the Riemann problem $[U_L, U_R]$.

Now first let $\bar\gamma_1$ and $\bar\gamma_2$ denote the nonlinear waves such that $\bar\gamma_1 \sim \bar\gamma_1^a\bar\gamma_1^b$ and $\bar\gamma_2 \sim \bar\gamma_2^a\bar\gamma_2^b$. Define $\bar J_1'$ to be the I-curve obtained by replacing the waves on J_1' by the

waves $\bar\gamma_1\bar\gamma_0\bar\gamma_2$, and set $\bar J_1 = J_0 \cup \bar J_1'$. Then the proof of Theorem 5.1 is complete once we prove the following claim.

CLAIM. *The following inequalities hold:*

$$F(J_2) \le F(\bar J_1) \le F(J_1). \tag{5.6}$$

Proof of claim. The second inequality holds because in replacing the nonlinear waves $\bar\gamma_i^a\bar\gamma_i^b$ by $\bar\gamma_i$, $i = 1, 2$, there can be no increase in wave strength—only a cancellation of wave strength can occur, this happening when one of $\bar\gamma_i^a, \bar\gamma_i^b$ is a shock wave and the other is a rarefaction wave. Thus $L_w(\bar J_1) - L_w(J_1) \le 0$. Moreover, since the potential $d(\gamma_0, \gamma_r)$ is in general independent of the partitioning of the nonlinear wave γ_r, it follows that $d(\gamma_0, \bar\gamma_i) \le d(\gamma_0, \bar\gamma_i^a) + d(\gamma_0, \bar\gamma_i^b)$ for any standing wave γ_0 on J_0, $i = 1, 2$. From this it readily follows that $P(\bar J_1) - P(J_1) \le 0$, and hence $F(\bar J_1) - F(J_1) = L_w(\bar J_1) - L_w(J_1) + P(\bar J_1) - P(J_1) \le 0$.

To verify that $F(J_2) \le F(\bar J_1)$, write

$$F(J_2) - F(\bar J_1) = F(J_2') - F(\bar J_1') + P(J_2', J_0) - P(\bar J_1', J_0) + P(J_2', J_2') - P(\bar J_1', \bar J_1'). \tag{5.7}$$

(Here we use the notation that if $J_a, J_b \subset J$, then $P(J_a, J_b) = \sum d(\gamma^\alpha, \gamma^\beta)$, where the sum is taken over all approaching waves on J such that $\gamma^\alpha \in J_a$, $\gamma^\beta \in J_b$.) But $P(J_2', J_2') = 0$ because the solution of the Riemann problem contains no approaching waves, and by Proposition 5.2, $F(J_2') - F(\bar J_1') \le 0$. Moreover, $P(J_2', J_0) \le P(\bar J_1', J_0)$ because, by Proposition 4.3, the trajectories of the rarefaction waves on $\bar J_1'$ contains the trajectories of the rarefactions waves on J_2'; hence there will be an interaction potential between rarefaction waves in $\bar J_1'$ and standing waves in J_0 that cancels any interaction potential between rarefaction waves in J_2' and standing waves in J_0. Thus (5.7), $F(J_2) - F(\bar J_1) \le 0$, and the proof of the claim is complete. □

The final theorem follows directly from Theorem 5.1.

THEOREM 5.3. *If the initial I-curve $J_{t=0}$ satisfies $F(J_{t=0}) < \infty$ in a Glimm approximate solution $U_{\Delta x}$, then the total variation of $U_{\Delta x}(\cdot, t) <$ const.$F(J_{t=0})$ for all $t > 0$.*

6. Convergence of the residual. In this section we give the proof of convergence of the residual for the approximate Glimm scheme solution constructed in section 3. The residual for system (1.1) is defined by

$$R(a, u, \varphi) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \{u\varphi_t + f\varphi_x + a'g\varphi\}\ dx\,dt + \int_{-\infty}^{+\infty} u_0(x)\varphi(x, 0)\ dx. \tag{6.1}$$

Then (a, u) is a *weak solution* of (1.1) if and only if $R(a, u, \varphi) = 0$ for all compactly supported smooth test functions $\varphi = \varphi(x, t)$. Assume that $U_{\Delta x}$ is a sequence of Glimm approximate solutions that satisfy

$$\mathrm{Var}_z U_{\Delta x}(\cdot, t) < V_z \tag{6.2}$$

for some constant V_z independent of Δx (cf. (3.6)), and assume $U_{\Delta x}(x, t) = (a_{\Delta x}(x), u_{\Delta x}(x, t)) \to U(x, t) = (a(x), u(x, t))$ piecewise a.e. and in L^1_{loc} at each fixed time, uniformly on compact sets (the conclusion of the Oleinik compactness argument; cf. [25]). Note that for fixed initial data, $U_{\Delta x}$ is a function of both Δx and the sample sequence $\theta = \{\theta_{ij}\} \in \Theta$. Assume that $a(x)$ is Lipschitz continuous, so that there exists a constant M such that

$$|a(x) - a(y)| \le M|x - y| \text{ for all } x, y \in \mathbf{R}, \tag{6.3}$$

$$|a(x) - a_{\Delta x}(x)| \le M\Delta x \text{ for all } x \in \mathbf{R}. \tag{6.4}$$

For an approximate solution $U_{\Delta x}$, define

$$R_{\Delta x}(a_{\Delta x}, u_{\Delta x}, \varphi) \equiv \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \{u_{\Delta x}\varphi_t + f(a_{\Delta x}, u_{\Delta x})\varphi_x + a' g(a_{\Delta x}, u_{\Delta x})\varphi\}\, dx\, dt + \int_{-\infty}^{+\infty} u_{\Delta x}(x,0)\varphi(x,0)\, dx \tag{6.5}$$

(obtained by replacing U by $U_{\Delta x}$ in (6.1) everywhere except at a'). We prove the following theorem; cf. [4].

THEOREM 6.1. *There exists a set $\mathcal{N}$ of measure zero in Θ such that, if $\theta \in \Theta/\mathcal{N}$, then*

$$R(a, u, \varphi) = \lim_{\Delta x \to 0} R_{\Delta x}(a_{\Delta x}, u_{\Delta x}, \varphi) = 0 \tag{6.6}$$

for all test functions φ of compact support in $-\infty < x < \infty$, $t \geq 0$. Thus, in particular, passing the limit through the integral sign, we conclude that $U(x,t)$ is a weak solution of (1.7).

Proof of Theorem 6.1. To start, let γ_{ij}^1 and γ_{ij}^2 denote the negative and positive speed waves positioned at mesh point (x_i, t_j) in the approximate solution $U_{\Delta x}$. Let $U_{ij}(x,t)$ denote the approximate solution $U_{\Delta x}$ restricted to the mesh rectangle $\mathcal{R}_{ij} = [x_i, x_{i+1}) \times [t_j, t_{j+1})$, and let $\mathrm{Var}_z U_{ij}$ and $\mathrm{Var}_u U_{ij}$ denote the total variation of U_{ij} in x at fixed time $t \in (t_j, t_{j+1})$, $x_i \leq x < x_{i+1}$. For the proof of Theorem 6.1, we introduce three regularization parameters $\epsilon, \hat{\epsilon}$, and δ, whose values will be chosen at the end: ϵ is a regularization parameter for the standing waves described below; $\hat{\epsilon}$ measures distance to the transition curve so that

$$S(\hat{\epsilon}) \equiv \{U : |U - \mathcal{T}| \leq \hat{\epsilon}\};$$

and δ is a mollification parameter for $g_{\Delta x}$ (so that we can integrate the source term in (6.6) by parts),

$$(g \cdot U_{\Delta x})_\delta = (g \cdot U_{\Delta x}) * \psi_\delta,$$

where $\psi_\delta = (\frac{1}{\delta^2})\psi(\frac{x}{\delta}, \frac{t}{\delta})$ denotes the standard convolution kernel supported on $|(x,t)| \leq \delta$.

For the mollification of the standing waves, let $U_{\Delta x}^\epsilon(x,t) \equiv (a_\epsilon(x), u_{\Delta x}^\epsilon(x,t))$ denote the regularization of $U_{\Delta x}$ obtained by translating γ_{ij}^1 (respectively, γ_{ij}^2) $\epsilon\Delta x$ units to the left (respectively, right) at each mesh point (x_i, t_j) and then replacing each standing wave discontinuity γ_{ij}^0 at (x_i, t_j) by the smoothed out standing wave on the interval $x_i - \epsilon\Delta x < x < x_i + \epsilon\Delta x$, as described in the discussion after (2.4). Thus, states on the smoothed out standing wave $\gamma_{ij}^{0,\epsilon}$ lie on the standing wave curve between the same left and right states as γ_{ij}^0 so that $\mathrm{Var}_z U_{ij}^\epsilon = \mathrm{Var}_z U_{ij}$ and $\mathrm{Var}_u U_{ij}^\epsilon = \mathrm{Var}_u U_{ij}$. (Indeed, recall that in Section 2, standing wave discontinuities were constructed as limits of smooth standing waves under rescaling into discontinuities; cf. [6].) Since U^ϵ satisfies the same total variation bounds as $U_{\Delta x}$, by taking appropriate subsequences, we can assume that at each $\epsilon > 0$, $\lim_{\Delta x \to 0} U_{\Delta x}^\epsilon = U^\epsilon$, where convergence is in the same sense as $U_{\Delta x} \to U$. (We are forced to introduce $U_{\Delta x}^\epsilon$ because our approximate solutions are constructed to (formally) meet (6.6) with $a'_{\Delta x}$, not a'.)

We use the following lemmas.

LEMMA 6.2. *There exists a constant $C_0 > 0$ and a function $\hat{K}(\hat{\epsilon})$ independent of Δx such that*

$$\mathrm{Var}_u U_{\Delta x}(\cdot,t) \le C_0\hat{\epsilon} + \hat{K}(\hat{\epsilon})\mathrm{Var}_z U_{\Delta x}(\cdot,t), \tag{6.7}$$

$$\int_E |U^\epsilon_{\Delta x} - U_{\Delta x}|\,dx\,dt \le C_0|E|\epsilon, \tag{6.8}$$

$$\left|\frac{\partial}{\partial x}(g\cdot U^\epsilon_{\Delta x})_\delta\right| \le \frac{C_0}{\delta}. \tag{6.9}$$

By (6.8) we know that $\int_E |U^\epsilon - U|\,dxdt < O(1)\epsilon$ for each compact set E.[9]

Proof. Estimate (6.7) follows from the fact that the mapping $(a,z) \to (a,u)$ is one-to-one and regular except at the transition curve $\mathcal{T}$ and any wave that lies entirely within $S(\hat{\epsilon})$ has amplitude order $\hat{\epsilon}$; cf. [25]. For (6.8), observe that $meas\{(x,t) \in E : U^\epsilon_{\Delta x} \ne U_{\Delta x}\} = O(1)|E|\epsilon$, where $|E|$ denotes the measure of the set E. Estimate (6.9) follows directly from the definition of convolution. ∎

LEMMA 6.3. *For every compact set E in $-\infty < x < \infty$, $t \ge 0$, there exists a function $K(\epsilon)$ independent of δ such that*

$$\int_E |(g\cdot U^\epsilon_{\Delta x})_\delta - (g\cdot U^\epsilon_{\Delta x})|\,dxdt = \int_E |(g\cdot U^\epsilon)_\delta - (g\cdot U^\epsilon)|\,dxdt + o(\Delta x)K(\epsilon), \tag{6.10}$$

$$\int_E |(g\cdot U^\epsilon_{\Delta x})_\delta - (g\cdot U_{\Delta x})|\,dxdt = \int_E |(g\cdot U^\epsilon)_\delta - (g\cdot U)|\,dxdt + o(\Delta x)K(\epsilon). \tag{6.11}$$

Here we mean that $o(\Delta x)$ is independent of $\epsilon, \hat{\epsilon}$, and δ, and $\lim_{\Delta x\to 0} o(\Delta x) = 0$.

Proof. Both (6.10) and (6.11) follow directly from the convergence of $U^\epsilon_{\Delta x} \to U^\epsilon$ and $U_{\Delta x} \to U$. ∎

The next lemma is the main step in the proof of Theorem 6.1.

LEMMA 6.4. *Let*

$$R^\epsilon_\phi \equiv R(a_\epsilon, u^\epsilon_{\Delta x}, \varphi) = \int\int_{t\ge 0} U^\epsilon_{\Delta x}\phi_t + f(U^\epsilon_{\Delta x})\phi_x + a'_\epsilon g(U^\epsilon_{\Delta x})\phi\,dxdt + \int_{-\infty}^{+\infty} U_0(x)\phi(x,0)dx, \tag{6.12}$$

and write $R^\epsilon_\phi \equiv R^\epsilon_\phi(\theta)$ to express the dependence on $\theta \in \Theta$ when Δx and ϕ are fixed. Then there exists a constant C_1 such that

$$\int_\Theta \left(R^\epsilon_\phi\right)^2 d\theta \le O(1)\left\{\hat{\epsilon} + \hat{K}(\hat{\epsilon})\Delta x + \epsilon(C_0\hat{\epsilon} + \hat{K}(\hat{\epsilon}))^2\right\}. \tag{6.13}$$

Proof of Lemma 6.4. Since $U^\epsilon_{\Delta x}$ is an exact solution in each strip $t_j < t < t_{j+1}$, integrating (6.13) over each mesh rectangle $\mathcal{R}_{ij}$ gives

$$R^\epsilon_\phi = \sum_{i,j} D^\epsilon_{ij}(\theta, \Delta x, \phi), \tag{6.14}$$

[9] We use the notation that C_0, C_1 denote constants that can depend on the equations and the initial data but are independent of $\epsilon, \hat{\epsilon}, \delta, \Delta x$, and the test function ϕ, while $O(1)$ denotes a constant that is independent of $\epsilon, \hat{\epsilon}, \delta$, and Δx, the convergence parameters.

where for $j > 0$,

$$D^\epsilon_{ij}(\theta, \Delta x, \phi) = \int_{x_i}^{x_{i+1}} \left\{U^\epsilon_{\theta,\Delta x}(x, t_j+) - U^\epsilon_{\theta,\Delta x}(x, t_j-)\right\} \phi(x, t_j-) dx$$
$$\equiv \int_{x_i}^{x_{i+1}} [U^\epsilon_{ij}]\phi| dx \tag{6.15}$$

and

$$D^\epsilon_{i0}(\theta, \Delta x, \phi) = \int_{x_i}^{x_{i+1}} \left\{U^\epsilon_{\theta,\Delta x}(x, 0) - U_0(x)\right\} \phi(x, 0) dx$$
$$\equiv \int_{x_i}^{x_{i+1}} [U^\epsilon_{i0}]\phi| dx. \tag{6.16}$$

(We take definitions (6.14)–(6.16) as applying also at $\epsilon = 0$, $U^0_{\Delta x} = U_{\Delta x}$.) It follows directly from (6.15) and (6.16) that

$$\left|D^\epsilon_{ij}(\theta, \Delta x, \phi)\right| \le |\mathrm{Supp}(\phi)|\ ||\phi||_\infty \Delta x \mathrm{Var}_u U^\epsilon_{ij}. \tag{6.17}$$

Now let $O(1)$ denote a constant that is independent of $\epsilon, \hat{\epsilon}, \delta$, and Δx.

Claim. *The following estimate holds:*

$$\left|\int_\Theta D^\epsilon_{ij} D^\epsilon_{kl} d\theta\right| \equiv |\langle D^\epsilon_{ij}, D^\epsilon_{kl}\rangle| \le O(1)\epsilon \Delta x^2 \cdot \mathrm{Var}_u U^\epsilon_{ij} \cdot \mathrm{Var}_u U^\epsilon_{kl}. \tag{6.18}$$

Proof of claim. First, neglecting higher order terms in Δx, we can assume without loss of generality that ϕ is constant on mesh rectangles, $\phi = \phi_{ij} = \text{const}$ on $\mathcal{R}_{ij}$. Following the argument in [6], we first note that if $j < l$, then U_{ij} is independent of a_{kl}, and so we can pass da_k through the integral to the factor $\int_0^1 \int_{x_k}^{x_{k+1}} D_{kl} dx da_k$, which is equal to zero as in Glimm's original argument. Thus,

$$\int_\Theta D^\epsilon_{ij} D_{kl} d\theta = 0. \tag{6.19}$$

Therefore,

$$\begin{aligned} \left|\langle D^\epsilon_{ij}, D^\epsilon_{kl}\rangle\right| &= \left|\langle D^\epsilon_{ij}, D^\epsilon_{kl}\rangle - \langle D^\epsilon_{ij}, D_{kl}\rangle\right| \\ &= \left|\langle D^\epsilon_{ij}, D^\epsilon_{kl} - D_{kl}\rangle\right| \le ||D^\epsilon_{ij}||_\infty ||D^\epsilon_{kl} - D_{kl}||_\infty \\ &\le O(1)\mathrm{Var}_u U^\epsilon_{ij} ||D^\epsilon_{kl} - D_{kl}||_\infty \Delta x. \end{aligned} \tag{6.20}$$

But

$$||D^\epsilon_{kl} - D_{kl}||_\infty \le \int_{x_k}^{x_{k+1}} |[U^\epsilon_{kl}] - [U_{kl}]|\ |\phi| dx \le O(1)\epsilon \Delta x \mathrm{Var}_u U_{kl}, \tag{6.21}$$

and using this in (6.20) gives

$$\left|\langle D^\epsilon_{ij}, D^\epsilon_{kl}\rangle\right| \le O(1)\epsilon \Delta x^2 \cdot \mathrm{Var}_u U_{ij} \cdot \mathrm{Var}_u U_{ij} \tag{6.22}$$

as claimed.

Thus we can estimate

$$\int_\Theta \left(R_\phi^\epsilon\right)^2 d\theta = \int_\Theta \left(\sum_{ij} D_{ij}^\epsilon\right)^2 d\theta$$

$$= \sum_i \int_\Theta \left(D_{ij}^\epsilon\right)^2 d\theta + \sum_{ij,kl} \int_\Theta D_{ij}^\epsilon D_{kl}^\epsilon d\theta \tag{6.23}$$

$$= I + II, \tag{6.24}$$

where

$$|I| \le \sum_{ij} \left(\int_{x_i}^{x_{i+1}} [U_{ij}^\epsilon]\phi dx\right)^2 \le O(1) \sum_{ij} \Delta x^2 \left(\mathrm{Var}_u U_{ij}^\epsilon\right)^2$$

$$\le O(1)\left\{C_0\hat{\epsilon} + \hat{K}(\hat{\epsilon})\right\}\Delta x \tag{6.25}$$

and

$$|II| \le \sum_{ij,kl} O(1)\epsilon\Delta x^2 \cdot \mathrm{Var}_u U_{ij} \cdot \mathrm{Var}_u U_{kl}$$

$$\le \sum_{ij,kl} O(1)\epsilon\Delta x^2 \cdot (C_0\hat{\epsilon} + \hat{K}(\hat{\epsilon}))^2. \tag{6.26}$$

Thus

$$|I| + |II| \le O(1)\left\{\hat{\epsilon} + \hat{K}(\hat{\epsilon})\Delta x + \epsilon(C_0\hat{\epsilon} + \hat{K}(\hat{\epsilon}))^2\right\}, \tag{6.27}$$

which verifies (6.13) of Lemma 6.4. □

Now that we have an estimate for R_ϕ^ϵ in Lemma 6.4; we obtain an estimate for $R_\phi \equiv R_{\Delta x}(a_{\Delta x}, u_{\Delta x}, \varphi)$ by estimating the difference $|R_\phi^\epsilon - R_\phi|$,

$$|R_\phi| \le \left|R_\phi^\epsilon - R_\phi\right| + \left|R_\phi^\epsilon\right|. \tag{6.28}$$

LEMMA 6.5. *The following estimate holds:*

$$\left|R_\phi^\epsilon - R_\phi\right| \le O(1)\left\{\epsilon + o(\Delta x)K(\epsilon) + \frac{\Delta x}{\delta}\right.$$
$$+ \int\int_E |(g\cdot U^\epsilon)_\delta - g(U^\epsilon)|\, dx\, dt$$
$$\left. + \int\int_E |(g\cdot U^\epsilon)_\delta - g(U)|\, dx\, dt\right\}, \tag{6.29}$$

where E denotes the support of ϕ and $O(1)$ denotes a constant independent of $\epsilon, \hat{\epsilon}, \delta$, and Δx.

Proof of Lemma 6.5. Starting with (6.5) and (6.12), we obtain

$$\left|R_\phi^\epsilon - R_\phi\right| \le \int\int_{t\ge 0} |U_{\Delta x}^\epsilon - U_{\Delta x}|\, |\phi_t|\, dx\, dt$$
$$+ \int\int_{t\ge 0} |f(U_{\Delta x}^\epsilon) - f(U_{\Delta x})|\, |\phi_x|\, dx\, dt$$
$$+ \left|\int\int_{t\ge 0} \{a_\epsilon' g(U_{\Delta x}^\epsilon) - a' g(U_{\Delta x})\}\, \phi\, dx\, dt\right|$$

$$= I_1 + I_2 + I_3. \tag{6.30}$$

It follows from (6.9) that

$$|I_1| \le O(1)\epsilon, \tag{6.31}$$

$$|I_2| \le O(1)\epsilon, \tag{6.32}$$

and it remains to estimate I_3. But

$$\begin{aligned} |I_3| &\le \int\int_{t\ge 0} |a'_\epsilon|\,|g(U^\epsilon_{\Delta x}) - (g\cdot U^\epsilon_{\Delta x})_\delta|\,|\phi|\,dx\,dt \\ &\qquad + \int\int_{t\ge 0} |a'_\epsilon|\,|(g\cdot U^\epsilon_{\Delta x})_\delta - g(U_{\Delta x})|\,|\phi|\,dx\,dt \\ &\qquad\qquad + \left|\int\int_{t\ge 0} \frac{d}{dx}(a_\epsilon - a)\,(g\cdot U^\epsilon_{\Delta x})_\delta\,\phi\,dx\,dt\right| \\ &= I_{3a} + I_{3b} + I_{3c}, \end{aligned}$$

and using Lemma 6.3 we obtain

$$|I_{3a}| \le O(1)\left\{\frac{1}{\epsilon}\int\int_{t\ge 0} |(g\cdot U^\epsilon)_\delta - g(U^\epsilon)|\;dx\,dt + o(\Delta x)K(\epsilon)\right\} \tag{6.33}$$

and

$$|I_{3b}| \le O(1)\left\{\int\int_{t\ge 0} |(g\cdot U^\epsilon)_\delta - g(U)|\;dx\,dt + o(\Delta x)K(\epsilon)\right\}. \tag{6.34}$$

Finally, integrating I_{3c} by parts and using (6.9) we obtain

$$|I_{3c}| \le O(1)\int\int_{t\ge 0} |a_\epsilon - a|\left|\frac{d}{dx}(g\cdot U^\epsilon)_\delta\right|\,dx\,dt \le O(1)\frac{\Delta x}{\delta}. \tag{6.35}$$

Putting (6.31)–(6.35) into (6.30) yields (6.29) of Lemma 6.5. □

We can now give the following proof.

Proof of Theorem 6.1. To establish (6.6) for $R_{\Delta x}(a_{\Delta x}, u_{\Delta x}, \varphi) \equiv R_\phi$, we show that

$$\lim_{\Delta x\to 0}\int_\Theta R_\phi^2\,d\theta = 0. \tag{6.36}$$

To this end, using (6.28) we can write

$$\begin{aligned} \int_\Theta R_\phi^2\,d\theta &\le 2\int_\Theta \left(R^\epsilon_\phi\right)^2\,d\theta + 2\int_\Theta \left|R^\epsilon_\phi - R_\phi\right|^2\,d\theta \\ &\le O(1)\left\{\hat\epsilon + \left[\hat K(\hat\epsilon)\Delta x\right]_1 + \left[\epsilon(\hat\epsilon + \hat K(\hat\epsilon)\right]_2\right\} \\ &\quad + O(1)\left\{\left[\epsilon + \int\int_E |(g\cdot U^\epsilon)_\delta - g(U^\epsilon)|\;dx\,dt \right.\right. \\ &\qquad\qquad + \left.\int\int_E |(g\cdot U^\epsilon)_\delta - g(U)|\;dx\,dt\right]_3 \\ &\qquad\qquad\qquad + \left.\left[o(\Delta x)K(\epsilon) + \frac{\Delta x}{\delta}\right]_4\right\}^2, \end{aligned} \tag{6.37}$$

where we have applied (6.13) and (6.29). Now let τ be any small positive number. Then, to make $\int_\Theta R_\phi^2\,d\theta < \tau$, choose $\hat\epsilon, \epsilon, \delta$ and $\Delta x \ll 1$ in order as follows (the brackets and $O(1)$ refer to quantities defined in (6.37)). First choose $\hat\epsilon \ll 1$ so that

$$O(1)\hat\epsilon < \frac{\tau}{4}; \tag{6.38}$$

choose $\epsilon < \epsilon_0 \ll 1$ so that

$$O(1)\,[\cdot]_2 < \frac{\tau}{4}; \tag{6.39}$$

choose $\epsilon \ll \epsilon_0$ and $\delta \ll 1$ so that

$$[\cdot]_3 < \frac{1}{2}\sqrt{\left\{\frac{\tau/4}{O(1)}\right\}}; \tag{6.40}$$

finally, choose $\Delta x \ll 1$ so that

$$[\cdot]_1 < \frac{\tau}{4} \quad \text{and} \quad [\cdot]_4 < \frac{1}{2}\sqrt{\left\{\frac{\tau/4}{O(1)}\right\}}. \tag{6.41}$$

Putting (6.38)–(6.41) into (6.37), we obtain

$$\int_\Theta R_\phi^2\,d\theta < \frac{\tau}{4} + \frac{\tau}{4} + \frac{\tau}{4} + \frac{1}{2}\left(\frac{\tau}{4} + \frac{\tau}{4}\right) < \tau. \tag{6.42}$$

From (6.42) we conclude (6.36), from which we conclude that $R_\phi \to 0$ off a set of measure zero in Θ. Theorem 6.1 now follows by taking a countable dense set of test functions, extracting a set of measure for each one, and taking $\theta \in \Theta/\mathcal{N}$, where $\mathcal{N}$ is the union of the measure zero sets for each of the countable list of test functions; cf. [4]. This completes the proof of Theorem 6.1. □

7. Appendix. In this appendix, we verify Propositions 4.3 and 5.2.

Proof of Proposition 4.3. Let $\bar\gamma_1\bar\gamma_0\bar\gamma_2$ be any connected sequence of incoming waves that take $U_L \to U_R$, and let $[U_L, U_R] = \gamma_1\gamma_0\gamma_2$. To verify the proposition, we can list the sixteen possibilities for $\bar\gamma_1\bar\gamma_0\bar\gamma_2$ according to whether γ_i are shock waves or rarefaction waves ($i = 1, 2$, four cases), whether γ_0 lies to the left or right of $\mathcal{T}$, and whether a increases or decreases across γ_0. (Since the issue involves only the location of the standing wave curves, it is not important whether $g_u > 0$ or $g_u < 0$.) In each case it is easy to verify that the rarefaction waves in the solution of the Riemann problem lie within the standing wave curves that bound the rarefaction waves among the incoming waves $\bar\gamma_1\bar\gamma_0\bar\gamma_2$. It follows that $Traj(\gamma_r^L) \subseteq Traj(\bar\gamma_r^L)$ and $Traj(\gamma_r^R) \subseteq Traj(\bar\gamma_r^R)$ in each case. The details are omitted. □

Proof of Proposition 5.2. We show that $F(\gamma_1\gamma_0\gamma_2) \le F(\bar\gamma_1\bar\gamma_0\bar\gamma_2)$ for any connected sequence of incoming waves $\bar\gamma_1\bar\gamma_0\bar\gamma_2$ that take U_L to U_R, where the outgoing waves $\gamma_1\gamma_0\gamma_2 = [U_L, U_R]_P$. (Recall that $[U_L, U_R]_P$ is obtained from $[U_L, U_R]$ by replacing every triple composite wave by its projection. Note that no wave can precede or follow, a triple composite wave in $[U_L, U_R]$ when $g_u < 0$ or $g_u > 0$, respectively, so that $[U_L, U_R]_P$ always consists of three waves $\gamma_1\gamma_0\gamma_2$.) We verify $F(\gamma_1\gamma_0\gamma_2) \le F(\bar\gamma_1\bar\gamma_0\bar\gamma_2)$ in four salient cases diagrammed in Figure 31. All other cases follow by a concatenation of these cases.

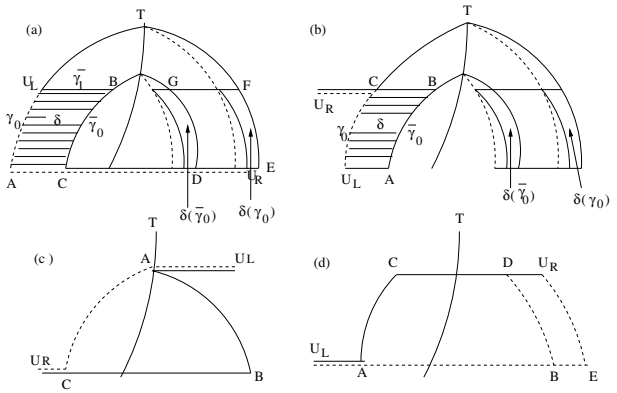


FIG. 31.

Cases (a) and (b) of Figure 31 deal with regular interactions in which the standing waves lie on the same side of $\mathcal{T}$ before and after interaction (the case $\gamma_0, \bar\gamma_0 < \mathcal{T}$ is considered). The point here is that when $\bar\gamma_0$ interacts with a shock wave (Case (a)) F decreases because L_w^* decreases on standing wave–shock wave interactions, and this decrease dominates the change in the corrective terms $\delta(\bar\gamma_0)$ and $\delta(\gamma_0)$ which were added to make L_w continuous. Cases (c) and (d) deal with the case when the standing waves $\bar\gamma_0$ and γ_0 lie on opposite sides of the transition curve, and the crossing occurs by rarefaction wave and shock wave, respectively. We now discuss the cases (a)–(d) of Figure 31 in detail.

Case (a). In this case, $F([U_L, U_R]_P) - F(\bar\gamma_1\bar\gamma_0\bar\gamma_2) = L_w(U_L \to A \to C) - L_w(U_L \to B \to C)$. When $g_u < 0$, $\delta(\bar\gamma_0) = 0 = \delta(\gamma_0)$ because these correction terms are added to standing waves on the right of $\mathcal{T}$ in this case. So when $g_u < 0$, $L_w(U_L \to A \to C) - L_w(U_L \to B \to C) = L_w^*(U_L \to A \to C) - L_w^*(U_L \to B \to C) < 0$ by Lemma 2.7. On the other hand, when $g_u > 0$, we have $L_w(U_L \to A \to C) - L_w(U_L \to B \to C) = L_w^*(U_L \to A \to C) - L_w^*(U_L \to B \to C) + \delta(\gamma_0) - \delta(\bar\gamma_0) - \delta < 0$, because $\delta + \delta(\gamma_0) - \delta(\bar\gamma_0) < 0$ by Lemma 2.7. (That is, the decrease $-\delta$ in L_w^* due to interaction with the shock wave $A \to C$ dominates the change $\delta(\gamma_0) - \delta(\bar\gamma_0)$.)

Case (b). In this case, $F([U_L, U_R]_P) - F(\bar\gamma_1\bar\gamma_0\bar\gamma_2) = L_w(U_L \to C) - L_w(U_L \to A \to B \to C)$. (Here, $d_r(\bar\gamma_0, \bar\gamma_2) = 0$ because the waves $\bar\gamma_0$ and $\bar\gamma_2$ do not approach.) Now if $g_u < 0$, then $\delta(\gamma_0) = 0 = \delta(\bar\gamma_0)$ (because these correction terms are added to the waves on the right of $\mathcal{T}$ when $g_u < 0$), so we have $L_w(U_L \to C) - L_w(U_L \to A \to B \to C) = L_w^*(U_L \to C) - L_w^*(U_L \to A \to B \to C) \le 0$ by Lemma 2.7. On the other hand, when $g_u > 0$ we have $L_w(U_L \to C) = L_w^*(U_L \to C) + \delta(\gamma_0)$ and $L_w(U_L \to A \to B \to C) = L_w^*(U_L \to A \to B \to C) + \delta(\bar\gamma_0)$. But $L_w^*(\gamma_0) + L_w^*(C \to B) - L_w^*(U_L \to A \to B) = -\delta < 0$, and so $L_w^*(\gamma_0) - L_w^*(U_L \to A \to B \to C) \le -\delta - L_w^*(C \to B)$. Thus, $L_w(U_L \to C) - L_w(U_L \to A \to B \to C) = -\delta - L_w^*(C \to B) + \delta(\gamma_0) - \delta(\bar\gamma_0) \le 0$ as claimed.

Case (c). In this case, $F([U_L, U_R]_P) - F(\bar\gamma_1\bar\gamma_0\bar\gamma_2) = L_w(A \to C) - L_w(A \to B \to U_R) \le L_w^*(A \to B \to C) - L_w^*(A \to B \to U_R) = -L_w^*(C \to U_R) < 0$ by Lemma 2.7. (Here, $d_r(\bar\gamma_0, \bar\gamma_2) = 0$ because $B \to C$ lies below the standing wave curve through $\bar\gamma_0$, and $C \to U_R$ does not approach $\bar\gamma_0$.)

Case (d). In this case, choose D between C and U_R and B between A and E such that A, C, B, D lie on the same standing wave curve and $L_w(A \to C \to D) = L_w(A \to B \to D)$. Then $F([U_L, U_R]_P) - F(\bar\gamma_1\bar\gamma_0\bar\gamma_2) = L_w(A \to E \to U_R) - L_w(A \to$

$C \to U_R) = L_w(A \to E \to U_R) - L_w(A \to B \to D \to U_R) = L_w(B \to E \to U_R) - L_w(B \to D \to U_R) \le 0$ by the analysis of Case (a) (that is, we reduced the problem to the case of regular interaction on the right of $\mathcal{T}$).

REFERENCES

[1] G. Chen and J. Glimm, *Global solution to the compressible Euler equations with geometrical structure*, Comm. Math. Phys., 179 (1996), pp. 153–193.

[2] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, John Wiley, New York, 1948.

[3] G. Dal Maso, P. G. LeFloch, and F. Murat, *Definition and weak stability of nonconservative products*, J. Math. Pures Appl., 74 (1995), pp. 483–548.

[4] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[5] S. K. Godunov, *A difference method for numerical calculations of discontinuous solutions of the equations of hydrodynamics*, Mat. Sb., 47 (1959), pp. 271–306 (in Russian).

[6] J. Hong, *The Glimm Scheme Extended to Inhomogeneous Systems*, Ph.D. thesis, University of California, Davis, Davis, CA, 2000.

[7] E. Isaacson, *Global Solution of a Riemann Problem for a Non-strictly Hyperbolic System of Conservation Laws Arising in Enhanced Oil Recovery*, preprint, Rockefeller University, New York, 1981.

[8] E. Isaacson, D. Marchesin, B. Plohr, and B. Temple, *The Riemann problem near a hyperbolic singularity: The classification of solutions of quadratic Riemann problems* I, SIAM J. Appl. Math., 48 (1988), pp. 1009–1032.

[9] E. Isaacson and B. Temple, *The structure of asymptotic states in a singular system of conservation laws*, Adv. in Appl. Math., 11 (1990), pp. 205–219.

[10] E. Isaacson and B. Temple, *Analysis of a singular hyperbolic system of conservation laws*, J. Differential Equations, 65 (1986), pp. 250–268.

[11] E. Isaacson and B. Temple, *Examples and classification of non-strictly hyperbolic systems of conservation laws*, Abstracts Amer. Math. Soc., January 1985.

[12] E. Isaacson and B. Temple, *Nonlinear resonance in systems of conservation laws*, SIAM J. Appl. Math., 52 (1992), pp. 1260–1278.

[13] E. Isaacson and B. Temple, *Convergence of the* 2×2 *Godunov method for a general resonant nonlinear balance law*, SIAM J. Appl. Math., 55 (1995), pp. 625–640.

[14] B. Keyfitz and H. Kranzer, *A system of non-strictly hyperbolic conservation laws arising in elasticity theory*, Arch. Ration. Mech. Anal., 72 (1980), pp. 219–241.

[15] S. N. Kruzkov, *First order quasilinear equations with several space variables*, Mat. USSR-Sb., 10 (1970), pp. 217–243.

[16] P. D. Lax, *Hyperbolic systems of conservation laws,* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[17] P. D. Lax and B. Wendroff, *Systems of conservation laws*, Comm. Pure Appl. Math., 13 (1960), pp. 217–237.

[18] L. Lin, J. B. Temple, and J. Wang, *A comparison of convergence rates for Godunov's method and Glimm's method in resonant nonlinear systems of conservation laws*, SIAM J. Numer. Anal., 32 (1995), pp. 824–840.

[19] L. Longwei, B. Temple, and W. Jinghua, *Suppression of oscillations in Godunov's method for a resonant non-strictly hyperbolic system*, SIAM J. Numer. Anal., 32 (1995), pp. 841–864.

[20] T. P. Liu, *Quasilinear hyperbolic systems*, Comm. Math. Phys., 68 (1979), pp. 141–172.

[21] T. P. Liu, *Resonance for a quasilinear hyperbolic equation*, J. Math. Phys., 28 (1987), pp. 2593–2602.

[22] D. Marchesin and P. J. Paes-Leme, *A Riemann Problem in Gas Dynamics with Bifurcation*, PUC Report MAT 02/84, Rio de Janeiro, Brazil, 1984.

[23] O. A. Oleinik, *Discontinuous solutions of non-linear differential equations*, Uspehi Mat. Nauk (N.S.), 12 (1957), pp. 3–73 (in Russian); Amer. Math. Soc. Transl. (2), 26 (1993), pp. 95–172 (in English).

[24] J. Smoller, *Shock waves and reaction diffusion equations*, Springer-Verlag, Berlin, New York, 1983.

[25] B. Temple, *Global solution of the Cauchy problem for a class of* 2×2 *nonstrictly hyperbolic conservation laws*, Adv. in Appl. Math., 3 (1982), pp. 335–375.

[26] A. Tveito and R. Winther, *Existence, Uniqueness and Continuous Dependence for a System of Hyperbolic Conservation Laws Modelling Polymer Flooding*, preprint, Department of Informatics, University of Oslo, Norway, 1990.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 858–877

A PARAMETER IDENTIFICATION PROBLEM OF MIXED TYPE RELATED TO THE MANUFACTURE OF CAR WINDSHIELDS*

PHILIPP KÜGLER†

Abstract. We study the identification of a parameter in a fourth-order elliptic partial differential equation that models the optimal design of car windshields to be manufactured by the sagging process. Considered as a second-order equation for the unknown parameter, the problem is of mixed type, i.e., changing between elliptic and hyperbolic. Numerical routines for directly solving this equation are not available. In this paper we both theoretically and numerically show that the inverse problem can instead be solved in a stable way by means of a (derivative free) iterative regularization method. The course of the iteration nevertheless depends markedly on the mixed type of the second-order equation.

Key words. nonlinear parameter identification, iterative regularization, PDE of mixed type

AMS subject classifications. 47A52, 35J60

DOI. 10.1137/S0036139903423339

1. Introduction.

Transforming automotive glass from a flat sheet into a curved car windshield is a challenging subject for industrial and academic research. The resulting change in the glass surface area leads to deformations that may cause optical distortions of unacceptable refractive and reflective quality. A related problem that the manufacturer faces is the limited formability of the glass. Hence, not every shape designed at the manufacturer's drawing table can be (immediately) realized in practice and thus costly shape corrections may become necessary.

One industrial method favored for the manufacture of car windshields is the sag bending process: A sheet of glass is put over a rigid frame with the desired edge curvature and is heated from above. The glass becomes viscous and sags under its own weight; the final shape of the glass depends on the viscosity distribution of the glass obtained by varying the temperature. Hence, given a desired target shape, the task is to find the appropriate temperature distribution in order to achieve that goal; see, e.g., [11].

Although the sag bending process operates in the viscous regime, the viscous-elastic analogy allows us to consider the Young's modulus E, a spatially varying glass material parameter, to be proportional to the viscosity; see [12]. Then, since the latter is a function of the temperature (see [8]), the sag bending process can in a first approximation also be controlled in terms of E, where the bending of the glass sheet is described by means of the linear elastic plate theory. Hence, our inverse problem is to identify the parameter E for a given target shape $\hat{w}$, where its solution can finally be used in order to compute the appropriate temperature distribution.

In section 2, we discuss the fourth-order elliptic direct bending problem, for which we shall consider two types of boundary conditions on the fastening of the glass sheet. In section 3, we show that the inverse parameter identification problem can be solved by a (recently developed derivative free) iterative regularization method, whose

*Received by the editors February 14, 2003; accepted for publication (in revised form) June 17, 2003; published electronically March 11, 2004. This work was supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung, projects F013-08 and P13478-INF.
http://www.siam.org/journals/siap/64-3/42333.html
†Industrial Mathematics Institute, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria (kuegler@indmath.uni-linz.ac.at).

convergence can be established under rather natural assumptions. We also discuss the so-called direct approach, which leads to a second-order partial differential equation of mixed type for the unknown parameter E. Due to the changes between elliptic and hyperbolic regions in dependence on the desired target shape $\hat{w}$, it is an open problem how to directly solve this parameter PDE. However, both the convergence rate analysis of the derivative free iterative method and the numerical tests in section 4 show that this mixed PDE type is reflected in the iterative algorithm. Taking care of this special feature is of independent mathematical interest but also might help to improve existing routines for the control of the sag bending process.

2. The direct problem of bending a plate. In a first model based on the viscous-elastic analogy and the linearized elasticity theory (see, e.g., [12], [2], and [8]), the sag bending process can be controlled in terms of the Young's modulus E, a spatially varying and positive glass material parameter. Then, the displacements w of the glass sheet are described by the fourth-order elliptic PDE

$$\frac{t^3}{12(1-\nu^2)}\{(E(w_{xx}+\nu w_{yy}))_{xx}+(E(w_{yy}+\nu w_{xx}))_{yy} + 2(1-\nu)(Ew_{xy})_{xy}\} = f \text{ in } \Omega, \tag{2.1}$$

where t denotes the thickness of the glass plate, $\Omega \subset R^2$ represents its midplane, $\nu \in (0, 0.5)$ is the glass Poisson ratio, and f denotes gravity. As boundary conditions on w we consider either

$$w|_{\partial\Omega} = 0, \quad \frac{\partial w}{\partial n} = 0 \tag{2.2}$$

for a clamped plate, or

$$w|_{\partial\Omega} = 0, \quad M_n = 0 \text{ on } \partial\Omega \tag{2.3}$$

for a simply supported plate; i.e., the moment M_n vanishes such that the plate is allowed to freely rotate around the tangent to $\partial\Omega$. In the case of a rectangular frame, the second condition in (2.3) simplifies to

$$w_{xx} + \nu w_{yy} = 0 \quad \text{along the edges with } y = \text{constant}, \tag{2.4}$$

$$w_{yy} + \nu w_{xx} = 0 \quad \text{along the edges with } x = \text{constant} \tag{2.5}$$

due to the positivity of E; see [9]. For our further discussion, a Hilbert space setup is of advantage; hence we next turn to the weak formulation of (2.1). Denoting by Y_0 a closed subspace of the Hilbert space $Y = H^2(\Omega)$ and considering f as an element of the dual space Y_0^*, the displacement $w \in Y_0$ can be sought as the solution of the operator equation

$$A(E)w = f \text{ in } Y_0^*, \tag{2.6}$$

where $A(E) : Y_0 \to Y_0^*$ is defined via the symmetric bilinear form

$$\langle A(E)w, v\rangle = \int_\Omega \frac{Et^3}{12(1-\nu^2)}\left[(w_{xx}+w_{yy})(v_{xx}+v_{yy}) - (1-\nu)(w_{xx}v_{yy}+w_{yy}v_{xx}-2w_{xy}v_{xy})\right] dxdy \tag{2.7}$$

on $Y_0 \times Y_0$. The space Y_0 is determined by the boundary conditions on w under consideration, where

$$Y_0 = \{w \in Y \,|\, w|_{\partial\Omega} = 0\} \tag{2.8}$$

corresponds to (2.3) and

$$Y_0 = \left\{ w \in Y \,|\, w|_{\partial\Omega} = 0 \,\wedge\, \frac{\partial w}{\partial n} = 0 \right\} \tag{2.9}$$

represents (2.2). The next theorem shows that, given an appropriate parameter E, problem (2.6) is uniquely solvable.

THEOREM 2.1. *For any Young's modulus E belonging to the set*

$$\tilde{Q} = \left\{ E \in H^1(\Omega) \,|\, \underline{\gamma} \le E \le \bar{\gamma} \right\}, \tag{2.10}$$

where $\underline{\gamma}, \bar{\gamma}$ are positive constants, the direct problem (2.6) *admits a unique solution in Y_0.*

Proof. Simple manipulations of the bilinear form (2.7) yield that the operator $A(E)$ is continuous in the sense

$$|\langle A(E)w, v\rangle| \le \alpha_2 \|w\| \|v\|, \quad w, v \in Y,\ E \in \tilde{Q}, \tag{2.11}$$

with a positive constant $\alpha_2 = \alpha_2(\bar{\gamma})$. Furthermore, since $v \in Y_0$ (both for (2.8) and (2.9)), with v a polynomial of degree one, implies $v = 0$, we can apply the theorem on equivalent norms in order to obtain the ellipticity

$$\langle A(E)w, w\rangle \ge \alpha_1 \|w\|^2, \quad w, v \in Y_0,\ E \in \tilde{Q}, \tag{2.12}$$

with a positive constant $\alpha_1 = \alpha_1(\underline{\gamma})$. For details we refer to [10]. Hence, by virtue of the Lax–Milgram lemma (see, for instance, [17]) problem (2.6) admits a unique solution in Y_0 for any $E \in \tilde{Q}$. ∎

In the following we denote the unique solution of (2.6) by w_E in order to emphasize its dependence on the parameter E.

3. The inverse problem. Having introduced the direct problem (2.1), (2.6) as a first model for describing the bending of the glass sheet resulting from the sag bending process, we now discuss the associated inverse windshield problem. Given a target shape $\hat{w}$ that satisfies either the boundary condition (2.2) or (2.3), we want to find a positive Young's modulus $E = E(x, y)$ such that the corresponding direct problem admits $\hat{w}$ as its solution.

In this section, we first introduce and analyze a derivative free *iterative* regularization method, which then in fact allows us to numerically solve the inverse windshield problem in a stable way. This strategy is based on minimizing the deviation between a computed forward solution of the PDE and the desired target shape. We also focus on the *direct* approach, where the idea is to consider the state equation as a second-order (partial differential) equation for the unknown parameter. Since this equation then is of mixed type, numerical concepts for its solution are unavailable. Nevertheless, this approach demands special attention since the mixed type is also reflected in the iterative method.

3.1. The iterative approach. Introducing the set of admissible parameters

$$Q = \left\{ E \in X \,|\, \underline{\gamma} \leq E \leq \bar{\gamma} \right\}, \tag{3.1}$$

where X is a Hilbert space, and the parameter-to-output map

$$F : Q \to Y, E \to w_E,$$

where w_E denotes the solution of the direct problem (2.6), the inverse windshield problem can be formulated as the nonlinear operator equation

$$F(E) = \hat{w}. \tag{3.2}$$

In the following, we assume that *the exact data* $\hat{w} \in Y_0$ are attainable by a parameter $E_* \in Q$, i.e., that the windshield is manufacturable. Note that this does not imply that the solution E_* of (3.2) has to be unique. Already translated to the underlying real world problem, several solutions may even be of advantage since they give more freedom in choosing the strategy for heating the glass. Target shapes that would also be accepted as $\hat{w}$ by the car producer are taken into account as *perturbed data* $w^\delta \in Y_0$, where δ in

$$\|\hat{w} - w^\delta\| \leq \delta \tag{3.3}$$

has to be understood as a level of tolerance for the outcome of the bending process.

Parameter identification problems such as (3.2) are typically ill-posed in the sense that their solution does not depend continuously on the data. Hence, data but also round-off errors may be amplified by an arbitrarily large error if one applies methods to (3.2) that are only suited for well-posed problems; see [3]. In order to overcome these instabilities one has to use *regularization* methods. Iterative techniques—especially advantageous for nonlinear problems—are mostly based on a successive minimization of the output least-squares functional

$$E \to \frac{\lambda}{2} \|F(E) - w^\delta\|^2, \tag{3.4}$$

where λ is a scaling parameter; see the survey given in [5]. Though the initial guess E_0 is always supposed to lie in a neighborhood of E_*, i.e.,

$$E_* \in \mathcal{B}_{\rho/2}(E_0), \tag{3.5}$$

where ρ is chosen such that $\mathcal{B}_\rho(E_0) \subset Q$ is satisfied, stability can be enforced, i.e., a reliable approximation to the solution of (3.2) can be obtained, only if the iteration is stopped at the right time depending on δ. Denoting the iterates by E_k^δ, the discrepancy principle (see, for instance, [3] or [7]) suggests determining the stopping index $k_*(\delta)$ by

$$\|w^\delta - F(E_{k_*}^\delta)\| \leq \tau\delta < \|w^\delta - F(E_k^\delta)\|, \quad 0 \leq k < k_*, \tag{3.6}$$

for some sufficiently large $\tau > 0$. The (final) residual $w^\delta - F(E_{k_*}^\delta)$ then is of the order of the tolerance level, which is the best we should ask for.

All the classical iterative regularization methods for solving (3.2), (3.3) in a stable way, like the Landweber method,

$$E_{k+1}^\delta = E_k^\delta + \lambda F'(E_k^\delta)^*(w^\delta - F(E_k^\delta)) \tag{3.7}$$

(see [7]), which allow a comprehensive analysis of their convergence behavior, require the existence of the Fréchet derivative of F and further conditions on F'; see [5], which usually are hard to verify for parameter identification problems in higher dimensions. Instead, we apply the *derivative free* Landweber method,

$$E_{k+1}^{\delta} = E_k^{\delta} + \lambda L(E_k^{\delta})^*(w^{\delta} - w_k), \tag{3.8}$$

introduced in [9], where $L(E)^*$ denotes the Hilbert space adjoint of the linear operator

$$L(E) : X \to Y_0,\ h \to -JA(h)w_E \tag{3.9}$$

for $E \in Q$. Thereby, w_k is used as an abbreviation for $F(E_k^{\delta})$, $A(h)$ is defined by (2.7), and $J : Y_0^* \to Y_0$ represents the duality map. For the windshield problem it is of interest to approximate the given target shape also in terms of its second derivatives, since the related curvatures characterize the final optical quality of the windshield. Hence, it is appropriate to use the full Y-topology in building the adjoint operator of (3.9) (as well as in (3.3)).

In the following, we choose the Hilbert space $X = H^s(\Omega)$ with $s > d/2$ such that $X \subset L^{\infty}(\Omega)$ is satisfied. Obviously, we have $Q \subset \tilde{Q}$ such that (2.11) and (2.12) especially hold for Q and the forward operator F in fact is well defined. Furthermore, definition (2.7) yields

$$A(\cdot)u \in \mathcal{L}(X, Y_0^*) \tag{3.10}$$

because of

$$\langle A(h)v, w\rangle \le c\|h\|\|v\|\|w\|, \quad h \in X, v, w \in Y, \tag{3.11}$$

where c denotes the embedding constant. Together with (2.6) and (2.7), this also implies that the iteration operator is locally bounded, i.e.,

$$\|L(E)\| \le \hat{L}, \qquad E \in \mathcal{B}_{\rho}(E_0), \tag{3.12}$$

with $\hat{L} = c\|f\|/\alpha_1$.

In establishing convergence of the iterates of (3.8), we follow the basic concept of [7]. However, since we do not resort to strong conditions on the Fréchet derivative of F, we still have to proceed in a different manner. The first result shows that the error in the parameter is monotonically decreasing as long as the discrepancy principle is obeyed.

PROPOSITION 3.1. *Assume that E_* is a solution of* (3.2) *in $\mathcal{B}_{\rho/2}(E_0)$, and let λ and τ be chosen such that*

$$2\left(\alpha_1 - \frac{\alpha_2}{\tau}\right) - \lambda\hat{L}^2 \ge D \tag{3.13}$$

holds, where D is a fixed positive constant. In case of perturbed data w^{δ} satisfying (3.3)*, we denote by k_* the stopping index of the iteration according to the discrepancy principle* (3.6) *with τ satisfying* (3.13)*. Then we have*

$$\|E_* - E_{k+1}^{\delta}\| \le \|E_* - E_k^{\delta}\|, \ \ 0 \le k < k_*, \tag{3.14}$$

and

$$\sum_{k=0}^{k_*-1} \|w^{\delta} - w_k\|^2 \le \frac{\rho^2}{4\lambda D}. \tag{3.15}$$

For $\delta = 0$ (with $\tau = \infty$ in (3.13)), we have

$$\sum_{k=0}^{\infty} \|\hat{w} - w_k\|^2 \le \frac{\rho^2}{4\lambda D}. \tag{3.16}$$

Proof. Given $\|E_0 - E_*\| \le \rho/2$, we assume

$$\|E_k^\delta - E_*\| \le \rho/2$$

for $k < k_*(\delta)$ and argue by induction. Then, the iteration step (3.8) is well defined, yielding

$$\begin{aligned} &\|E_* - E_{k+1}^\delta\|^2 - \|E_* - E_k^\delta\|^2 \\ &\quad = -2\lambda(L(E_k^\delta)(E_* - E_k^\delta), w^\delta - w_k) + \lambda^2 \|L(E_k^\delta)^*(w^\delta - w_k)\|^2. \end{aligned} \tag{3.17}$$

The following considerations play the decisive role in our analysis and are only possible for the special iteration operator (3.9). Because of this operator's definition, (3.10), and

$$A(E_*)\hat{w} = A(E_k^\delta)w_k \ \text{ in } \ Y_0^*,$$

we get

$$\begin{aligned} &-(w^\delta - w_k, L(E_k^\delta)(E_* - E_k^\delta)) \\ &\quad = \langle w^\delta - w_k, A(E_* - E_k^\delta)w_k\rangle \\ &\quad = \langle w^\delta - w_k, A(E_*)w_k - A(E_*)\hat{w}\rangle \\ &\quad = -\langle w^\delta - w_k, A(E_*)w^\delta - A(E_*)w_k\rangle + \langle w^\delta - w_k, A(E_*)w^\delta - A(E_*)\hat{w}\rangle \\ &\quad \le -\alpha_1\|w^\delta - w_k\|^2 + \alpha_2\|w^\delta - w_k\|\|w^\delta - \hat{w}\|, \end{aligned} \tag{3.18}$$

where the inequality holds because of (2.12) and (2.11). Using (3.18) in (3.17), we obtain

$$\begin{aligned} &\|E_* - E_{k+1}^\delta\|^2 - \|E_* - E_k^\delta\|^2 \\ &\le \|w^\delta - w_k\|\lambda\left(2\alpha_2\delta - 2\alpha_1\|w^\delta - w_k\| + \lambda\hat{L}^2\|w^\delta - w_k\|\right). \end{aligned}$$

Following the discrepancy principle (3.6), we get from (3.13) that

$$\|E_* - E_{k+1}^\delta\|^2 + \lambda D\|w^\delta - w_k\|^2 \le \|E_* - E_k^\delta\|^2$$

for $k < k_* = k_*(\delta)$. This implies assertion (3.14) and $E_{k+1}^\delta \in \mathcal{B}_{\rho/2}(E_*) \subset \mathcal{B}_\rho(E_0)$. Furthermore, we can conclude that

$$\lambda D \sum_{k=0}^{k_*-1} \|w^\delta - w_k\|^2 \le \sum_{k=0}^{k_*-1} \left(\|E_k^\delta - E_*\|^2 - \|E_{k+1}^\delta - E_*\|^2\right)$$

holds, which leads to the inequality

$$k_*\tau^2\delta^2 \le \sum_{k=0}^{k_*-1} \|w^\delta - w_k\|^2 \le \frac{\rho^2}{4\lambda D}$$

and finally to assertion (3.16). ∎

Hence, the monotonicity of the iterates, which is the foundation for the forthcoming convergence results, can be guaranteed under natural assumptions already associated with the solvability of the direct problem. We see that condition (3.13) can always be satisfied by choosing the λ sufficiently small and τ sufficiently large. Note that, in case of perturbed data, the use of a "large" τ in the discrepancy principle (3.6) might cause a premature termination of the iteration, a problem that is also present when using other iterative methods, e.g., (3.7). However, our choice of τ in (3.13) no longer involves (in practical situations) unknown constants that are linked to conditions on F'; compare to [7].

The estimation (3.16) shows that, in the absence of data noise, the residual norms of the iterates tend to zero for $k \to \infty$; hence, if the iteration converges, the limit certainly is a solution of the inverse windshield problem. In the case of perturbed data, (3.15) yields the existence of a unique stopping index k_* such that $\|w^\delta - w_k\| > \tau\delta$ holds for all $k < k_*$ but is violated at $k = k_*$.

The next theorem shows that, for precise data, the iterates E_k in fact converge to a solution of the inverse windshield problem. Furthermore, in the presence of data perturbations, the discrepancy principle (3.6) renders the derivative free Landweber iteration (3.8) a regularization method; i.e., we have $E^\delta_{k_*(\delta)} \to E_*$ as $\delta \to 0$.

THEOREM 3.2 (convergence). *Let $\delta = 0$ in (3.3). If (3.2) is solvable in $\mathcal{B}_{\rho/2}(E_0)$, then E_k converges to a solution $E_* \in \mathcal{B}_{\rho/2}(E_0)$ of (3.2), i.e.,*

$$E_k \to E_*, \quad k \to \infty. \tag{3.19}$$

In the case of perturbed data w^δ satisfying (3.3), let the iteration (3.8) be stopped at $k_*(\delta)$, according to the discrepancy principle (3.6), (3.13). Then

$$E^\delta_{k_*(\delta)} \to E_*, \quad \delta \to 0. \tag{3.20}$$

Proof. Again we can follow [7], but once more we require only the properties of the PDE-operator $A(E)$. For exact data, the basic idea is to verify that E_k is a Cauchy sequence. If $\tilde{E}$ denotes any solution of (3.2) in $\mathcal{B}_{\rho/2}(E_0)$, i.e., $w_{\tilde{E}} = \hat{w}$, the crucial ingredient for the proof is

$$\begin{aligned}(\hat{w} - w_r, L(E_r)(\tilde{E} - E_l)) &= -\langle A(\tilde{E} - E_l)w_r, \hat{w} - w_r\rangle \\ &= -\langle A(\tilde{E} - E_r)w_r, \hat{w} - w_r\rangle - \langle A(E_r - E_l)w_r, \hat{w} - w_r\rangle \\ &= \langle A(\tilde{E})\hat{w} - A(\tilde{E})w_r, \hat{w} - w_r\rangle \\ &\quad - \langle A(E_l)w_l - A(E_l)w_r, \hat{w} - w_r\rangle, \end{aligned} \tag{3.21}$$

which holds because of (3.9), (3.10), and

$$\begin{aligned} A(E_r)w_r &= A(\tilde{E})\hat{w} \ \text{ in } \ Y_0^*, \\ A(E_r)w_r &= A(E_l)w_l \ \text{ in } \ Y_0^*. \end{aligned}$$

Given (3.21), one can show as in [7] that $E_k - \tilde{E}$ and hence E_k are Cauchy sequences. Denoting the limit of E_k by E_*, we obtain that E_* is a solution of (3.2) since the residues $\hat{w} - w_k$ converge to zero for $k \to \infty$; see Proposition 3.1.

In the case of perturbed data, the proof given in [7] is independent of the iteration operator and therefore also applies to (3.8). ∎

Hence, the derivative free iteration (3.8) in combination with (3.6) provides a numerically stable algorithm for solving the inverse windshield problem. In order to

make it more transparent, we build the inner product in X of both sides with a test function $h \in X$. Using (3.9) and rearranging the terms then yields

$$(E_{k+1}^\delta - E_k^\delta, h) = -\lambda \frac{t^3}{12(1-\nu^2)} \left\{ \int_\Omega h(w_{kxx}(w^\delta - w_k)_{xx} + w_{kyy}(w^\delta - w_k)_{yy}) dxdy \right.$$
$$+ \int_\Omega \nu h(w_{kxx}(w^\delta - w_k)_{yy} + w_{kyy}(w^\delta - w_k)_{xx}) dxdy$$
$$\left. + \int_\Omega 2(1-\nu) h w_{kxy}(w^\delta - w_k)_{xy} dxdy \right\}. \tag{3.22}$$

As opposed to (3.7), where $F'(E_k^\delta)^*$ also requires us to solve (2.6) with f replaced by the current residual $w^\delta - F(E_k^\delta)$, (3.8) calls only for the computation of w_k. In that sense, the total number of "direct problems" to be solved is cut in half by (3.8).

3.2. The direct approach. Given a target shape $\hat{w}$, one also might look for a solution of the inverse problem by considering (2.1) as a PDE for E, i.e.,

(3.23)
$$((\hat{w}_{xx} + \nu\hat{w}_{yy})E)_{xx} + 2(1-\nu)(\hat{w}_{xy}E)_{xy} + ((\hat{w}_{yy} + \nu\hat{w}_{xx})E)_{yy} = \frac{12(1-\nu^2)}{t^3} f \text{ in } \Omega.$$

Usually, parameter identification problems are, when regarded as equations for the unknown parameter, of first order; i.e., the parameter appears at most up to its first derivatives. However, we now face an inverse problem that is of second order in the parameter. The type of (3.23) depends on the sign of

$$\bar{\Delta} = (\hat{w}_{xx} + \nu\hat{w}_{yy}) \cdot (\hat{w}_{yy} + \nu\hat{w}_{xx}) - (1-\nu)^2 \hat{w}_{xy}^2 : \tag{3.24}$$

(3.23) is elliptic where $\bar{\Delta} > 0$ and hyperbolic where $\bar{\Delta} < 0$. Note that the type depends in fact on the given target shape $\hat{w}$. The discriminant $\bar{\Delta}$ can also be written as

$$\bar{\Delta} = 4\nu \left(\frac{\hat{w}_{xx} + \hat{w}_{yy}}{2} \right)^2 + (1-\nu)^2 (\hat{w}_{xx}\hat{w}_{yy} - \hat{w}_{xy}^2), \tag{3.25}$$

where

$$C_G = \hat{w}_{xx}\hat{w}_{yy} - \hat{w}_{xy}^2$$

is the Gaussian curvature and

$$C_m = \frac{1}{2}(\hat{w}_{xx} + \hat{w}_{yy})$$

is the mean curvature of the shape.

Concentrating on rectangular frames—experiences identified them as the most problematic ones for the sag bending process—we next follow [16] in order to demonstrate that the direct approach for the inverse windshield problem leads to PDEs that are always of mixed type. Furthermore, we will see that there is a significant difference in the type between shapes satisfying (2.2) and those fulfilling (2.3).

In the simply supported case, (2.3), (2.4), (2.5), and the positivity of E imply that the product of $(\hat{w}_{xx} + \nu\hat{w}_{yy})$ and $(\hat{w}_{yy} + \nu\hat{w}_{xx})$ in (3.24) is zero on any simply

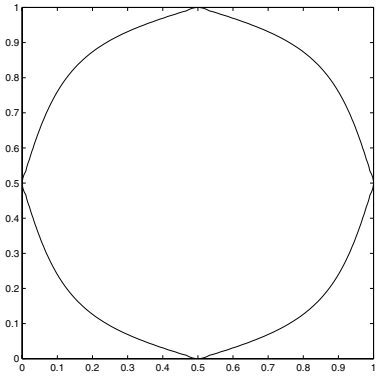


FIG. 1. *The simply supported case: an elliptic center area with adjacent hyperbolic corners.*

supported edge. Hence, $\bar{\Delta} \leq 0$ holds on the edges and (3.23) gets hyperbolic or parabolic there.

On the other hand, windshields usually have a positive Gaussian curvature C_G in their interior; i.e., choosing an interior point, its neighborhood lies on only one side of the tangential plane. Then, (3.25) shows that $\bar{\Delta} > 0$ such that the parameter (3.23) is elliptic in these regions. As a consequence, the equation type changes from hyperbolic near the edges to elliptic near the center of the region. Furthermore, it is shown in [16] that there is only one parabolic curve, i.e., a line defined by the points satisfying $\bar{\Delta} = 0$, and that it intersects each of the four sides of the squared frame at a single point. Hence, (3.23) is elliptic in the center and hyperbolic next to the corners of the frame. This typical behavior is illustrated in Figure 1.

In the clamped case (2.2), the equation for E will be elliptic in the center region according to the positive Gaussian curvature C_G of the target shape. However, along the edges of the frame, the situation is significantly different. For a rectangular frame, the zero gradient condition on the boundary becomes $w_x = 0$ and $w_y = 0$ on the edges $x = const.$ and $y = const.$, respectively. Concentrating on a single edge, e.g., $x = const.$, and differentiating $w_x = 0$ as well as the zero deflection condition $w = 0$ with respect to y, we obtain that w_y, w_{yy}, and w_{xy} also vanish along that edge. However, the discriminant $\bar{\Delta}$ (see (3.25)) there reduces to

$$\bar{\Delta}_{x=const.} = \nu w_{xx}^2,$$

such that (3.23) cannot be hyperbolic along that edge. In fact, since w_{xx} vanishes only at the ends of the edge $x = const.$, the equation is elliptic along the edge and parabolic only at the very corners. Nevertheless, it is shown in [16] that the elliptic regions near the frame and in the center are divided by a hyperbolic ring. In contrast to the simply supported case, there now exist two distinct parabolic lines, where the outer one does not touch the edges of the domain Ω at all. A typical formation of the elliptic and hyperbolic regions is shown in Figure 2.

Facing a second-order PDE of mixed type, questions concerning existence, uniqueness, and stability of a solution to (3.23) arise. Naturally, one would call for boundary conditions for E on $\partial\Omega$ in case of a purely elliptic equation ($\bar{\Delta} > 0$ on Ω) and for Cauchy data on a suitable (noncharacteristic) part $\Gamma \subset \partial\Omega$ in a purely hyperbolic case ($\bar{\Delta} < 0$ on Ω). But for the present problem (3.23) with $\hat{w}$ satisfying (2.2) or

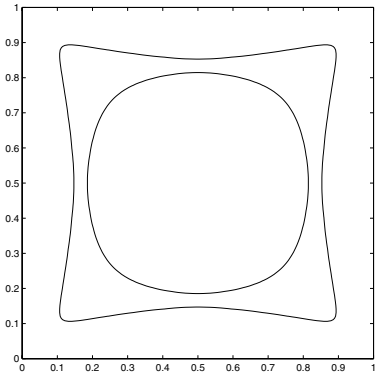


FIG. 2. *The clamped case: a hyperbolic ring between two elliptic areas.*

(2.3), it is not at all obvious how to proceed. The study in [16] of the characteristics of (3.23), where the two characteristic directions are given by

$$\frac{dy}{dx} = \frac{(1-\nu)\hat{w}_{xy} \pm \sqrt{(1-\nu)^2\hat{w}_{xy}^2 - (\hat{w}_{xx} + \nu\hat{w}_{yy}) \cdot (\hat{w}_{yy} + \nu\hat{w}_{xx})}}{\hat{w}_{xx} + \nu\hat{w}_{yy}}$$

(see [13]), gave some additional insight—for instance, although (3.23) is hyperbolic near the corners of the frame in the simply supported case, one cannot prescribe Cauchy data due to the fact that the edges are characteristics. However, it is not even clear whether side conditions on E should be prescribed at all.

So far, results about existence, uniqueness, or stability with respect to the data $\hat{w}$ of a solution to problem (3.23) are available only for special symmetric cases; see [14]. Since, in particular, numerical techniques for solving the equation of mixed type are unavailable, the direct approach (at the moment) is not suited to solving the inverse windshield problem. Nevertheless, returning to our iterative approach (3.8), we shall see that the features of the direct approach must not be neglected.

3.3. A first theoretical link. In general, the convergence in (3.19), (3.20) for iterative regularization methods may be arbitrarily slow; see [15]. Rate estimates can be obtained only under additional assumptions on the quality of the initial guess E_0 that are often difficult to comprehend; see [5]. Enhancing (3.8) by an additional stabilizing term, i.e., considering

$$E_{k+1}^\delta = E_k^\delta + L(E_k^\delta)^*(w^\delta - w_k) - \beta_k(E_k^\delta - E_0)$$

with a certain nonnegative sequence of decaying parameters β_k, the convergence rates

$$\|E_k - E_*\| = \mathcal{O}(\sqrt{\beta_k}) \qquad \text{(for exact data) and}$$
$$\|E_{k_*(\delta)}^\delta - E_*\| = \mathcal{O}(\sqrt{\beta_{k_*(\delta)}})$$

could be proven in [9] under the so-called *weak source condition*

$$\exists u \in Y_0, \quad E_* - E_0 = L(E_*)^* u. \tag{3.26}$$

In order to gain more insight into (3.26), we use the definition of $L(E_*)$ and multiply both sides by an arbitrary element $h \in X$. Then condition (3.26) assumes the

existence of a source function $u \in Y_0$ such that

$$\begin{aligned}\frac{12(1-\nu^2)}{t^3}(E_* - E_0, h) = &- \int_\Omega h(\hat{w}_{xx}u_{xx} + \hat{w}_{yy}u_{yy})dxdy \\ &- \int_\Omega \nu h(\hat{w}_{xx}u_{yy} + \hat{w}_{yy}u_{xx})dxdy \\ &- 2\int_\Omega (1-\nu)h\hat{w}_{xy}u_{xy}d\vec{x}\end{aligned}$$

holds. If $E_* - E_0$ is sufficiently smooth and if the boundary values of the initial guess E_0 coincide with those of E_*, we obtain

$$(E_* - E_0, h) = \int_\Omega N_X^* N_X (E_* - E_0) \cdot h \, dx, \tag{3.27}$$

where N_X denotes the linear operator that generates the norm in X, e.g., $N_X^* N_X = (I - \Delta - \Delta^2)$ for $X = H^2(\Omega)$. Hence, the weak source condition can be understood as a solvability condition for the second-order differential equation

$$\begin{aligned}-\frac{12(1-\nu^2)}{t^3}N_X^* N_X(E_* - E_0) = &(\hat{w}_{xx}u_{xx} + \hat{w}_{yy}u_{yy}) \\ &+ \nu(\hat{w}_{xx}u_{yy} + \hat{w}_{yy}u_{xx}) \\ &+ 2(1-\nu)\hat{w}_{xy}u_{xy}\end{aligned}$$

for the unknown function $u \in Y_0$. Rearranging the terms on the right-hand side, we end up with

$$\begin{aligned}-\frac{12(1-\nu^2)}{t^3}N_X^* N_X(E_* - E_0) = &(\hat{w}_{xx} + \nu\hat{w}_{yy})u_{xx} + (\hat{w}_{yy} + \nu\hat{w}_{xx})u_{yy} \\ &+ 2(1-\nu)\hat{w}_{xy}u_{xy}.\end{aligned} \tag{3.28}$$

Now, building the discriminant of (3.28) shows that the type of (3.28) is identical to that of the second-order PDE (3.23) for E.

Although we assume the attainability of the target shape $\hat{w}$, which can in fact be understood as a solvability assumption for the parameter equation (3.23), this does not automatically imply the solvability of (3.28), since lower-order terms in the unknown function are missing in the latter. We also mention that the boundary conditions for a possible solution of (3.28) are already determined by the space Y_0. Since both (2.2) and (2.3) are natural boundary conditions for a fourth-order equation, they might be inappropriate for (3.28).

Nevertheless, (3.28) gives a first theoretical coupling between the direct and the iterative approaches to solving the inverse windshield problem. The next section shows that the inverse problem can be practically solved by our iterative method (3.8), but it also numerically confirms the influence of the mixed type of (3.23) on the course of the iteration. We emphasize that the relation between the parameter PDE and the iterative regularization method is specific neither to the windshield problem nor to method (3.8). We refer to [4], where the classical Landweber iteration (3.7) was applied to a second-order parameter identification problem with a type ranging from purely hyperbolic to purely elliptic in dependence on the given target.

4. Numerical experiments.

4.1. Preliminaries. Though the windshield problem yields only strictly mixed type equations (3.23), still the significant difference between the simply supported and the hyperbolic situation allows us both to test the iterative method and to numerically investigate the influence of the equation type on its outcome. For that purpose, neither the thickness t of the plate, the right-hand side f, the Poisson ratio ν, nor the scaling of the parameter E are of relevance. With $\nu = 0.5$, the direct problem (2.6) then turns into

$$\int_\Omega E\Big[(w_{xx}+w_{yy})(v_{xx}+v_{yy}) - \frac{1}{2}(w_{xx}v_{yy}+w_{yy}v_{xx}-2w_{xy}v_{xy})\Big]\,dxdy = \int_\Omega \tilde{f}v\,dxdy, \qquad v\in Y_0, \tag{4.1}$$

with the solution space given by either (2.8) or (2.9). The use of a nonphysical right-hand side $\tilde{f}$, i.e., not including gravity force, facilitates the construction of test examples for which the solution of the inverse problem is analytically known.

Although the convergence analysis of (3.8), and of the methods discussed in [5], applied to the windshield problem would require a parameter space satisfying $X \subset L^\infty(\Omega)$ (see (3.1)), we choose $X = H^1(\Omega)$ for the numerics. On the one hand, this allows us to keep the numerical effort low (since the use of higher-order elements for the parameter is avoided), and on the other hand, it responds to the natural wish to keep regularity sufficient for the direct problem; compare to (2.10). All our tests have shown that the iterates remain in the domain $\tilde{Q}$ of the parameter-to-output map F without the use of a projection operator.

As a last small deviation from our theoretical foundations, we shall use a line search algorithm (see [6]) in order to accelerate (3.8). This results in an iteration index dependent "scaling" parameter λ_k. (Compared to a constant λ as required by the theory, this has no other influence on the course of the iteration than speeding it up; see [9].) In other words, the iteration finally reads as

$$E_{k+1}^\delta = E_k^\delta + \lambda_k \bar{E}_k, \tag{4.2}$$

where *the update* $\bar{E}_k$ satisfies

$$(\bar{E}_k,h) = -\int_\Omega h\Big\{(w_{kxx}(w^\delta-w_k)_{xx}+w_{kyy}(w^\delta-w_k)_{yy}) + \frac{1}{2}(w_{kxx}(w^\delta-w_k)_{yy}+w_{kyy}(w^\delta-w_k)_{xx}) + w_{kxy}(w^\delta-w_k)_{xy}\Big\}\,dxdy; \tag{4.3}$$

compare to (3.22). Due to the choice $X = H^1(\Omega)$, equation (4.3) can be considered as the weak formulation of

$$\bar{E}_k - \Delta\bar{E}_k = -\Big\{(w_{kxx}(w^\delta-w_k)_{xx}+w_{kyy}(w^\delta-w_k)_{yy}) + \frac{1}{2}(w_{kxx}(w^\delta-w_k)_{yy}+w_{kyy}(w^\delta-w_k)_{xx}) + w_{kxy}(w^\delta-w_k)_{xy}\Big\} \ \text{ in } \ \Omega, \tag{4.4}$$

$$\frac{\partial\bar{E}_k}{\partial n} = 0 \ \text{ on } \ \partial\Omega. \tag{4.5}$$

Hence, one iteration step for solving the inverse windshield problem consists of the following:

1. Given E_k^δ, calculate the solution w_k of the direct problem.
2. Build the residual $w^\delta - w_k$, and solve problem (4.4), (4.5) for the update $\bar{E}_k$.
3. Then, the new iterate E_{k+1}^δ is given by (4.2).

We emphasize that the second-order PDE (4.4) for $\bar{E}_k$ is purely elliptic, and hence the iterative algorithm never requires us to solve the second-order equation (3.23) for the parameter; thus there is no obvious connection to the mixed type resulting from the direct approach. The boundary condition (4.5) shows that the boundary flux $\frac{\partial E_0}{\partial n}$ of the initial guess E_0 is maintained during the whole iteration.

All computations to be presented in the following are based on the PDE Toolbox of MATLAB, using the finite element method. For the parameter we chose the built-in linear ansatz functions, while the solutions of the direct problem were represented by means of the discrete Kirchhoff triangle; see [1]. Furthermore, a regular and uniform triangular mesh with 665 nodes was used for $\Omega = [0,1] \times [0,1]$.

4.2. Simply supported vs. clamped target shape. For the numerical test we consider a clamped target shape

$$\hat{w}_C = -25(x^2 - 2x^3 + x^4)(y^2 - 2y^3 + y^4), \tag{4.6}$$

a simply supported target shape

$$\hat{w}_S = -(x - 2x^3 + x^4)(y - 2y^3 + y^4), \tag{4.7}$$

and a true parameter

$$E_* = 1 + x + 2y \tag{4.8}$$

on the unit square. The right-hand side $\tilde{f}$ in (4.1) is chosen such that $F(E_*) = \hat{w}_C$ or $F(E_*) = \hat{w}_S$ holds, respectively. Though in fact we treat two different direct problems, one for the simply supported and one for the clamped plate, this is no barrier for testing our algorithm and for comparing the respective inverse problems with respect to the parameter PDE structure. The elliptic and hyperbolic regions of (3.23) corresponding to $\hat{w}_S$ and $\hat{w}_C$ are those shown in Figures 1 and 2.

Ignoring data perturbations for the moment, we choose

$$E_0 = 4 \tag{4.9}$$

as an initial guess, meaning a relative deviation from E_* of approximately 80% measured with respect to the norm in X. The course of the iterations is documented in Figures 3 and 4, where the relative error

$$\frac{\|E_* - E_k\|}{\|E_*\|} \tag{4.10}$$

in the parameter, but also the relative error in the output, i.e.,

$$\frac{\|\hat{w} - w_k\|}{\|\hat{w}\|}, \tag{4.11}$$

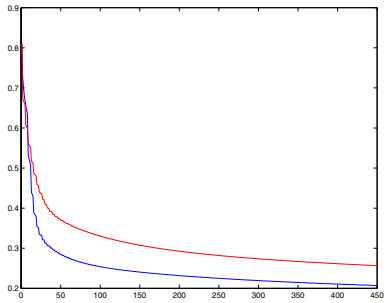


FIG. 3. (4.10) *vs.* k.

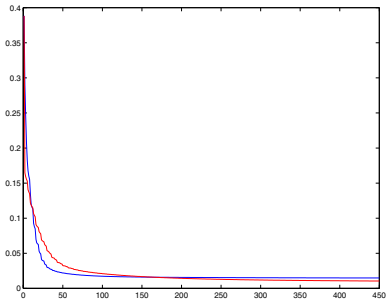


FIG. 4. (4.11) *vs.* k.

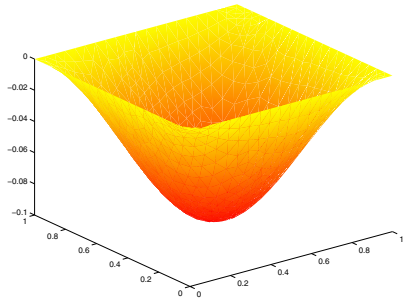


FIG. 5. *Computed output* w_{450}*, simply supported.*

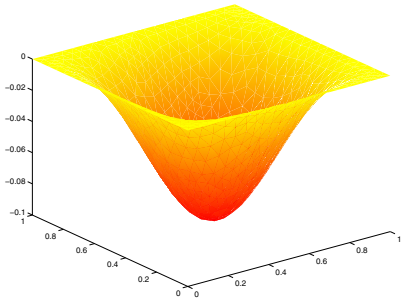


FIG. 6. *Computed output* w_{450}*, clamped.*

is plotted against the iteration index. (The relative error in the $L^2(\Omega)$-norm is approximately tenth part of that shown.) The simply supported case is represented by the upper line, the clamped case by the lower one. Figure 4 shows that the simply supported and the clamped target shapes are approximated with nearly the same quality by our method. The relative error is smaller than 2%, which is remarkable since the deviation in (4.11) is measured with respect to $H^2(\Omega)$. Regarding the relative error in the parameter, we observe a first difference between the simply supported and the clamped situation. Starting both computations from (4.9), it is an open question why the "clamped" error is significantly lower than the "simply supported" one.

Figures 5 and 6 confirm the quality of the computed outputs, while the corresponding parameters E_{450} are shown in Figures 7 and 8. In fact, the total $L^2(\Omega)$-deviations between E_* and E_{450} are nearly identical, while the total deviations in the gradient are higher in the simply supported case than in the clamped one. The oscillations along the boundary $\partial\Omega$ in both cases are caused by the attempt to satisfy the boundary condition

$$\frac{\partial E_k}{\partial n} = 0$$

due to the initial guess (4.9); compare to (4.5).

Not knowing the boundary values of the solution E_* (as in the previous example), one might think of an iteration procedure that ignores them entirely. For that purpose

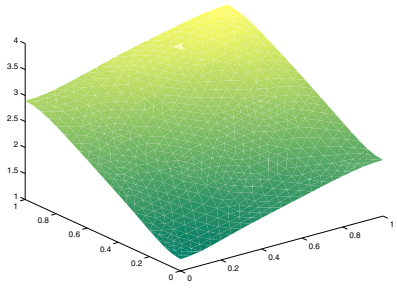


FIG. 7. E_{450}, *simply supported example.*

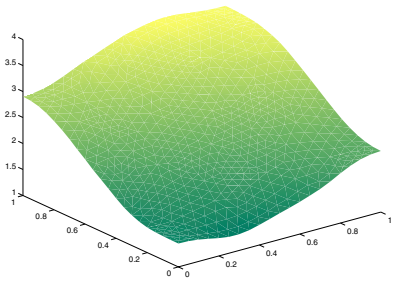


FIG. 8. E_{450}, *clamped example.*

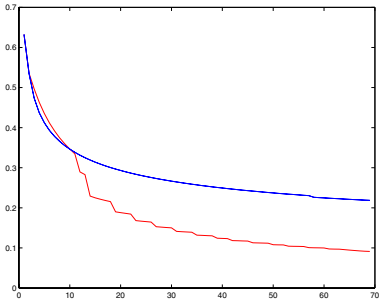


FIG. 9. (4.13) *vs.* k, *simply supported vs. clamped.*

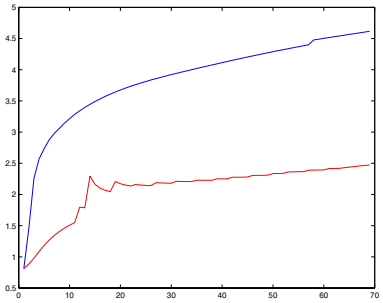


FIG. 10. (4.10) *vs.* k, *simply supported vs. clamped.*

we could use the $L^2(\Omega)$-inner product in the left-hand side of (4.3), resulting in the update rule

$$\begin{aligned}\bar{E}_k = -\Big\{ & (w_{k\,xx}(w^\delta - w_k)_{xx} + w_{k\,yy}(w^\delta - w_k)_{yy}) \\ & + \frac{1}{2}(w_{k\,xx}(w^\delta - w_k)_{yy} + w_{k\,yy}(w^\delta - w_k)_{xx}) \\ & + w_{k\,xy}(w^\delta - w_k)_{xy}\Big\} \quad \text{in } \Omega.\end{aligned} \tag{4.12}$$

Algorithm (4.12) can be related to the abstract formulation (3.8) by building the adjoint of the iteration operator $L(E_k^\delta)$ with respect to only the rougher space $L^2(\Omega)$. As opposed to (4.4), equation (4.12) does not describe a boundary value problem for the update $\bar{E}_k$. Though the parameters are still considered as elements belonging to $H^1(\Omega)$, boundary traces of the initial guess are not maintained during the iteration (at least not in an obvious way). Figures 9 and 10 show the performance of the iteration when using only the update rule (4.12), where the relative error

$$\frac{\|E_* - E_k\|_{L^2(\Omega)}}{\|E_*\|_{L^2(\Omega)}} \tag{4.13}$$

is plotted versus k in Figure 9, while the error (4.10) with respect to $H^1(\Omega)$ is recorded in Figure 10. The error behavior of the outputs w_k is similar to that shown in

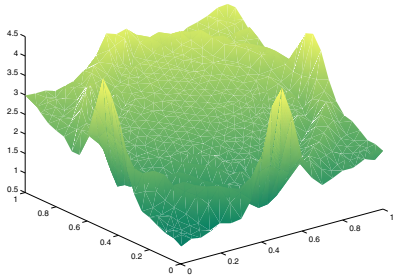


FIG. 11. E_{69}, *simply supported example.*

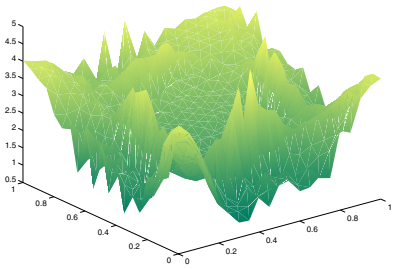


FIG. 12. E_{69}, *clamped example.*

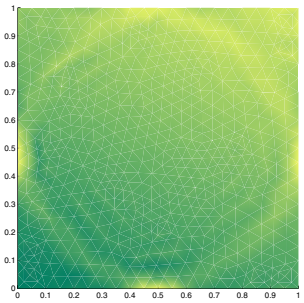


FIG. 13. E_{69}, *top view, simply supported.*

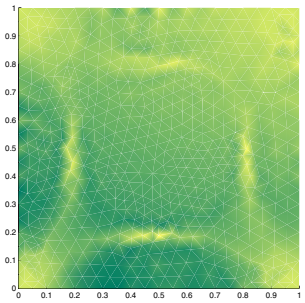


FIG. 14. E_{69}, *top view, clamped.*

Figure 4; hence we concentrate on only the parameters E_k. The $L^2(\Omega)$-error in the simply supported situation (lower line) now lies below its clamped counterpart (upper line), a ranking opposed to that shown in Figure 3. Regarding the $H^1(\Omega)$-norm, the iteration shows a divergent behavior; i.e., the error increases from the very beginning, leading to highly oscillating parameters, as illustrated in Figures 11 and 12. In these figures, dark shading means small errors, while light tone represents large deviations from the true parameter. Views of the computed solutions from above are given in Figures 13 and 14; a comparison to the elliptic and hyperbolic regions (see Figures 1 and 2) in the respective parameter PDE shows that its mixed type is reflected in the error structure of the iterative solutions. The parabolic line in Figure 1 is clearly observable in Figure 13, but the two parabolic lines bordering the hyperbolic ring in Figure 2 are also reflected in Figure 14. In particular the parabolic points lying on $\partial\Omega$ are highlighted: While in the simply supported example the iterates stay with the initial value of E_0 at the parabolic midpoints of each side of the frame and exactly reach the solution at the very corners, the parameters are left completely unchanged at the parabolic corners in the clamped case.

From the discussion of the direct approach in section 3 we know that the second derivatives of any solution of the direct problem, i.e., especially w_{kxx}, w_{kyy}, and w_{kxy}, vanish at the parabolic boundary points, both in the simply supported and the clamped situation. Hence, the right-hand side in (4.12) is zero at these points such that the initial guess E_0 cannot change there, explaining the behavior shown in Figures 11 and 12 along $\partial\Omega$. For that reason, these results cannot be improved by choosing a finer grid when staying with (4.12); on the contrary, the peaks would get even sharper. Concerning the interior of Ω, the parabolic lines for w_k are not fixed

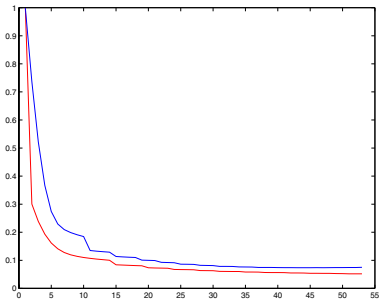


FIG. 15. (4.10) *vs.* k.

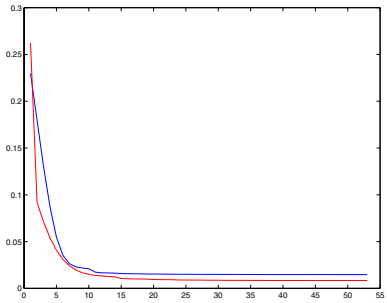


FIG. 16. (4.11) *vs.* k.

but tend during the iteration towards those of the target $\hat{w}$, as shown in Figures 1 and 2. Hence, their influence on (4.12) is not as strong as that of the nonchanging parabolic boundary points, at which the peaks in Figures 11 and 12 are highest.

We summarize our observations by

$$\bar{E}_k \approx 0 \Leftrightarrow \bar{\Delta} = 0,$$

where $\bar{\Delta}$ is the discriminant of the parameter PDE (3.23), finally suggesting that the parameter cannot be identified along the parabolic lines determined by $\bar{\Delta} = 0$. This lack of identifiability also exists during the iteration based on update (4.4), (4.5) but is then blurred out due to the smoothing effect of the PDE for $\bar{E}_k$.

Considering the direct approach for solving the inverse problem via the second-order PDE (3.23), the question of whether or what kind of boundary conditions on E should be prescribed is unanswered. This automatically translates into an uncertainty about the "right inner product" for the left-hand side in (4.3). However, as opposed to the direct approach, the iterative process at least allows us to test several choices. So far, we have considered two possibilities, namely, the neglect of boundary conditions via (4.12) and the prescription of Neumann data via (4.5). Aiming at a smooth approximation of the parameter—motivated by the motivating sag bending process and only then suitable for translation into a heating procedure—the latter variant is certainly preferable. If the boundary values of the solution (or the desired) E_* are given, we could use this information by a further manipulation of the iterative process (4.2). Restricting the test functions h in (4.3) from the space $H^1(\Omega)$ to $H_0^1(\Omega)$, we again can interpret (4.3) as the weak formulation of the elliptic PDE (4.4) for the update $\bar{E}_k$ but now with

$$\bar{E}_k = 0 \ \text{ on } \ \partial\Omega \tag{4.14}$$

as its boundary condition. Then the Dirichlet traces of the initial guess E_0 are maintained during the iteration. In terms of formulation (3.8), equation (4.4) in combination with (4.14) can be understood as building the adjoint of the iteration operator with respect to $H_0^1(\Omega)$.

The course of the iteration using (4.14) with an initial guess

$$E_0 = E_* + \frac{3}{2}\sin(\pi x)\sin(\pi y)$$

is recorded in Figures 15 and 16. The relative errors in the parameter and the output now are nearly identical in the simply supported (lower line) and the clamped situation

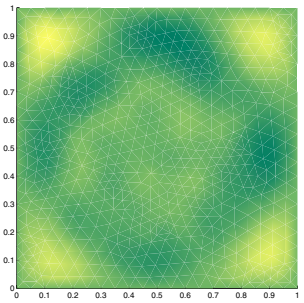


FIG. 17. $E_* - E_{53}$, *simply supported, top view.*

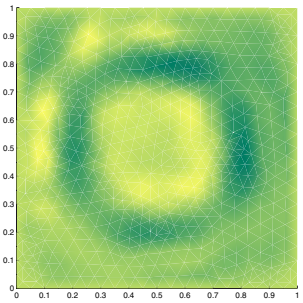


FIG. 18. $E_* - E_{53}$, *clamped, top view.*

TABLE 4.1
The data error.

	δ	$\frac{\|w^\delta - \hat{w}\|}{\|\hat{w}\|}$	$\frac{\|w^\delta - \hat{w}\|_{H^1(\Omega)}}{\|\hat{w}\|_{H^1(\Omega)}}$	$\frac{\|w^\delta - \hat{w}\|_{L^2(\Omega)}}{\|\hat{w}\|_{L^2(\Omega)}}$
$w = w_S$	0.292	0.293	0.012	0.001
$w = w_C$	0.417	0.29	0.02	0.002

(upper line); furthermore, they are (of course) lower than their counterparts from Figures 3 and 4. Nevertheless, the difference between the simply supported and the clamped examples becomes apparent when viewing the absolute error between E_* and the respectively computed parameters E_{53} from the top, as illustrated in Figures 17 and 18. Once again, the structure of the mixed type parameter PDE (3.23) is reflected in the results obtained by the iterative parameter identification method.

Finally, we briefly comment on the influence of data perturbations on the inverse windshield problem. Staying with the update rule (4.14), which led to the best results in the noise free situation, we now consider random perturbations w_C^δ and w_S^δ of the exact data (4.6) and (4.7). The respective relative data errors are given in Table 4.1. Though the errors are about 29% when measured with respect to the full $H^2(\Omega)$-norm, they are less than 0.5% if the perturbations are considered only in $L^2(\Omega)$. The numbers in the table also show that for approximating the given target shield (whether exact or perturbed) the full $H^2(\Omega)$-norm is indeed the appropriate one. Only then can errors in the second-order derivatives and the related curvature terms be minimized, which is essential for the optical quality of the windshield. Figures 19 and 20 now show the behavior that is typical for any iterative parameter identification method in the presence of data noise. While the relative error (4.11) (with $\hat{w}$ replaced by w^δ) in the output is monotonically decreasing, the error in the parameter shows a semiconvergent behavior (even though the true boundary values were fixed). Hence, a reliable approximation of E_* can be obtained only by stopping the iteration at the right time, for instance, according to the discrepancy principle (3.6). Furthermore, Figure 19 indicates that the iteration for the clamped case (upper line) is more sensitive to data perturbations than for the simply supported one (lower line).

The theoretical and numerical results presented in this paper clearly demonstrate that the inverse windshield problem (3.2) can be solved in a stable way by the derivative free iteration method (3.8) under natural assumptions and with minimal effort. Furthermore, we have seen that the direct approach, though methodologically differ-

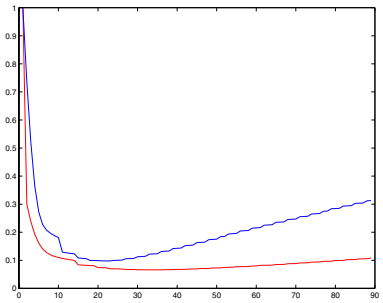


FIG. 19. (4.10) *vs.* k.

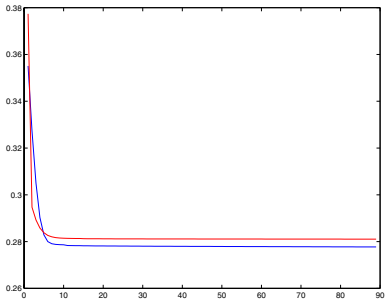


FIG. 20. (4.11) *vs.* k.

ent and (so far) not admitting a numerical implementation, is coupled to the iteration. A better understanding of its mixed type structure is of mathematical interest in its own right, but might also help to further improve the performance of the iterative algorithm.

Acknowledgment. I'd like to express my gratitude to Prof. Heinz W. Engl and the group of J. R. Ockendon for bringing the windshield problem to my attention and for fruitful discussions about the two approaches.

REFERENCES

[1] D. BRAESS, *Finite Elements*, Cambridge University Press, Cambridge, UK, 2001.
[2] P. G. CIARLET, *Mathematical Elasticity. Volume* I: *Three-Dimensional Elasticity*, Stud. Math. Appl. 20, North–Holland, Amsterdam, 1988.
[3] H. W. ENGL, M. HANKE, AND A. NEUBAUER, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, 1996.
[4] H. W. ENGL AND P. KÜGLER, *The influence of the equation type on iterative parameter identification problems which are elliptic or hyperbolic in the parameter*, European J. Appl. Math., 14 (2003), pp. 129–163.
[5] H. W. ENGL AND O. SCHERZER, *Convergence rate results for iterative methods for solving nonlinear ill-posed problems*, in Surveys on Solution Methods for Inverse Problems, D. Colton, H. W. Engl, A. K. Louis, J. McLaughlin, and W. F. Rundell, eds., Springer, Vienna, New York, 2000, pp. 7–34.
[6] R. FLETCHER, *Practical Methods of Optimization*, Vol. 1, John Wiley and Sons, New York, 1980.
[7] M. HANKE, A. NEUBAUER, AND O. SCHERZER, *A convergence analysis of the Landweber iteration for nonlinear ill-posed problems*, Numer. Math., 72 (1995), pp. 21–37.
[8] D. KRAUSE AND H. LOCH, EDS., *Mathematical Simulation in Glass Technology*, Springer-Verlag Berlin, Heidelberg, New York, 2002.
[9] P. KÜGLER, *A Derivative-Free Landweber Iteration for Parameter Identification in Elliptic Partial Differential Equations with Application to the Manufacture of Car Windshields*, Ph.D. thesis, Johannes Kepler Universität Linz, Austria, 2002.
[10] W. LITVINOV, *Optimization in Elliptic Problems with Applications to Mechanics of Deformable Bodies and Fluid Mechanics*, Birkhäuser-Verlag, Basel, Boston, Berlin, 2001.
[11] S. MANSERVISI, *Control and optimization of the sag bending process in glass windscreen design*, in Progress in Industrial Mathematics at ECMI98, L. Arkeryd, J. Bergh, P. Brenner, and R. Pettersson, eds., Teubner, Stuttgart, 1999, pp. 97–105.
[12] O. S. NARAYANASWAMY, *Stress and structural relaxation in tempering glass*, J. Amer. Ceramic Soc., 61 (1978), pp. 146–152.
[13] J. OCKENDON, S. HOWISON, A. LACEY, AND A. MOVCHAN, *Applied Partial Differential Equations*, Oxford University Press, London, 1999.
[14] D. SALAZAR AND R. WESTBROOK, *Inverse problems of mixed type in linear plate theory*, preprint.

[15] E. Schock, *Approximate solution of ill-posed equations: Arbitrarily slow convergence vs. superconvergence*, in Constructive Methods for the Practical Treatment of Integral Equations, G. Hämmerlin and K. Hoffmann, eds., Birkhäuser, Basel, 1985, pp. 234–243.

[16] D. Temple, *An Inverse System—An Analysis Arising from Windscreen Manufacture*, M.Sc. thesis, University of Oxford, England, 2002.

[17] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, Vol. II/a, Springer-Verlag, New York, Berlin, Heidelberg, 1980.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 878–901

ON THE SOLUTION TO THE RIEMANN PROBLEM FOR THE COMPRESSIBLE DUCT FLOW*

NIKOLAI ANDRIANOV† AND GERALD WARNECKE†

Abstract. The quasi-one-dimensional Euler equations in a duct of variable cross section form probably one of the simplest nonconservative systems. We consider the Riemann problem for it and discuss its properties. In particular, for some initial conditions, the solution to the Riemann problem appears to be nonunique. In order to rule out the nonphysical solutions, we provide two-dimensional computations of the Euler equations in a duct of corresponding geometry and compare them with the one-dimensional (1D) results. Then, the physically relevant 1D solutions satisfy a kind of entropy rate admissibility criterion.

Key words. nozzle flow, nonstrictly hyperbolic, resonance

AMS subject classifications. 35L65, 35L67, 76N99

DOI. 10.1137/S0036139903424230

1. Introduction.

In recent decades, considerable attention has been paid to both theoretical and numerical investigations of systems of conservation laws. This interest is caused by a wide range of applications for conservation laws, like fluid mechanics, astrophysics, meteorology, etc. However, there is an important class of problems which is described by the *nonconservative* systems, i.e., systems which cannot be written in divergence form. The example which we have in mind is two-phase flows.

The theory of nonconservative systems is still under development. The same can be said about the numerical methods for such systems. Here, we consider probably one of the simplest nonconservative systems, the system of the Euler equations in a duct of variable cross section. In [4] we have shown that this system can be formally obtained from the Baer–Nunziato model of two-phase flows [5], which describes the flame spread and the deflagration-to-detonation transition (DDT) in gas-permeable, reactive granular materials. Therefore, the results of [4] to a certain extent will also hold for the Euler equations in a duct of variable cross section. Conversely, a number of results which we present here are also valid for the Baer–Nunziato model in [4].

One can distinguish two parts in the system of Euler equations in a duct. These are a strictly hyperbolic system and an additional equation, which states that the cross section does not change in time. The resulting system appears to be only nonstrictly hyperbolic, with the *resonant behavior* when one of the nonlinear wave families has a zero wave speed. Such resonant systems have been studied before; see Isaacson and Temple [14, 15] and the references therein. We refer also to a recent work of Goatin and LeFloch [12] for a study of the Riemann problem for nonconservative resonant systems.

There are several difficulties concerning the solution of nonconservative systems. Due to the presence of nonconservative terms, one cannot use the definition of a weak solution from the theory of conservation laws. A general definition based on the theory

*Received by the editors March 6, 2003; accepted for publication (in revised form) September 25, 2003; published electronically March 30, 2004.
http://www.siam.org/journals/siap/64-3/42423.html

†Institut für Analysis und Numerik, Otto-von-Guericke Universität Magdeburg, PSF 4120, D-39016 Magdeburg, Germany (nikolai.andrianov@mathematik.uni-magdeburg.de, gerald.warnecke@mathematik.uni-magdeburg.de).

of nonconservative products was given in Dal Maso, LeFloch, and Murat [10]. In the particular case of the Riemann problem for the Euler equations in a duct, it appears that the system of governing equations is locally equivalent to some conservative system. This allows us to give a corresponding definition of the weak solution; see section 3.

It appears that, for certain initial conditions, the solution to the Riemann problem is not unique, despite the fact that all shocks locally satisfy a usual entropy criterion. We discuss the conditions which lead to nonuniqueness. For the selection of a physically relevant solution we compare some examples with two-dimensional (2D) duct flow computations. It appears that the 1D solutions picked out by 2D computations satisfy a kind of entropy rate admissibility criterion in analogy to that of Dafermos [8].

The paper is organized as follows. In section 2 we carry out the characteristic analysis of the system of governing equations. In section 3 we introduce the notion of a weak solution for the Riemann problem. In section 4 we discuss the structure of the Riemann solution and point out the conditions which lead to uniqueness and nonuniqueness. Finally, in section 5 we propose a criterion in order to select the physically relevant solution to the Riemann problem. We justify this criterion by comparing the 1D exact solutions with the computations of the Euler equations in a duct of corresponding geometry, averaged over the cross section.

2. Mathematical analysis. The system of Euler equations in a duct of variable cross section can be written in the following form:

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{h}(\mathbf{u})A_x, \tag{2.1}$$

where

$$\mathbf{u} = \begin{bmatrix} A \\ A\rho \\ A\rho v \\ A\rho E \end{bmatrix}, \qquad \mathbf{f}(\mathbf{u}) = \begin{bmatrix} 0 \\ A\rho v \\ A(\rho v^2 + p) \\ Av(\rho E + p) \end{bmatrix}, \qquad \mathbf{h}(\mathbf{u}) = \begin{bmatrix} 0 \\ 0 \\ p \\ 0 \end{bmatrix}. \tag{2.2}$$

In (2.1), $A = A(x)$ is the variable cross section, ρ is the density, v the velocity, p the pressure, and $E = e + v^2/2$ the specific total energy. We assume that the gas obeys the stiffened gas equation of state (EOS)

$$e = \frac{p + \gamma\pi}{\rho(\gamma - 1)}, \tag{2.3}$$

where γ and π are thermodynamic constants. When $\pi = 0$, we recover the usual ideal gas EOS.

Usually, the cross section $A = A(x)$ is assumed to be given a priori; see, e.g., Zucrow and Hoffman [18]. In (2.1), we consider it as an additional unknown, and we add the trivial equation $A_t = 0$ to determine it. The advantages of this approach are twofold. First, the system (2.1) belongs to the class of *resonant* systems; see Isaacson and Temple [14, 15]. Thus, one can use the results of [14, 15] for the system (2.1). Secondly, as we have noted in [4], system (2.1) can be formally obtained from the governing equations for the Baer–Nunziato model of two-phase flows [5]. Since system (2.1) is much simpler than the two-phase flow model [5], one can gain deeper insight into the structure of the model [5] by studying (2.1).

A particular issue about system (2.1) is the presence of the *nonconservative term* $p\,\partial A/\partial x$. Due to this term, one cannot write (2.1) in divergence form. Consequently,

one cannot define a weak solution and find the Rankine–Hugoniot conditions in the usual way, as it is done for systems of conservation laws.

In what follows, we will consider the Riemann problem for (2.1), i.e., equip it with piecewise constant initial data

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}_L, & x \le x_0, \\ \mathbf{u}_R, & x > x_0. \end{cases} \tag{2.4}$$

As usual, we will assume that the Riemann problem (2.1), (2.4) admits self-similar solutions, i.e.,

$$\mathbf{u}(x,t) = \mathbf{u}(\xi), \quad \xi = \frac{x - x_0}{t}. \tag{2.5}$$

To carry out the characteristic analysis of (2.1), it is convenient to use the primitive variables

$$\mathbf{v} = (A, \rho, v, \eta)^T, \tag{2.6}$$

where $\eta = (p + \pi)/\rho^\gamma$ is the isentrope. Then, for smooth solutions, the system (2.1) is equivalent to

$$\mathbf{v}_t + \mathbf{A}(\mathbf{v})\mathbf{v}_x = 0, \tag{2.7}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \rho v/A & v & \rho & 0 \\ 0 & c^2/\rho & v & \dfrac{1}{\rho}\dfrac{\partial p}{\partial \eta} \\ 0 & 0 & 0 & v \end{bmatrix}, \tag{2.8}$$

and $c = \sqrt{\gamma(p+\pi)/\rho}$ is the sound speed. The eigenvalues of $\mathbf{A}$ are

$$\lambda_0 = 0, \qquad \lambda_1 = v - c, \qquad \lambda_2 = v, \qquad \lambda_3 = v + c, \tag{2.9}$$

and the corresponding eigenvectors are

$$\mathbf{r}_0 = \begin{bmatrix} A(v^2 - c^2)/(vc^2) \\ -v\rho/c^2 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{r}_1 = \begin{bmatrix} 0 \\ 1 \\ -c/\rho \\ 0 \end{bmatrix}, \quad \mathbf{r}_2 = \begin{bmatrix} 0 \\ -\dfrac{\partial p}{\partial \eta} \\ 0 \\ c^2 \end{bmatrix}, \quad \mathbf{r}_3 = \begin{bmatrix} 0 \\ 1 \\ c/\rho \\ 0 \end{bmatrix}. \tag{2.10}$$

Note that situations are possible in which one of λ_1, λ_2, or λ_3 coincides with λ_0. Moreover, when λ_1 or λ_3 coincide with λ_0, the corresponding eigenvectors become linearly dependent. In this case a parabolic degeneracy occurs. To summarize, the system of governing equations (2.1) is hyperbolic away from the points where either $\lambda_1 = \lambda_0$ or $\lambda_3 = \lambda_0$. Note that system (2.1) is nonstrictly hyperbolic when $\lambda_2 = \lambda_0$.

We can easily see that the 1-, 2-, and 3-characteristic fields are exactly the same as for the usual one-dimensional Euler equations. The 1- and 3-characteristic fields are genuinely nonlinear, and the 2-field is linearly degenerate. The Riemann invariants for these fields coincide with those of the Euler equations.

It is obvious that the 0-characteristic field is linearly degenerate. In the solution to the Riemann problem (2.1), (2.4), this field corresponds to a stationary contact discontinuity. The condition $A_t = 0$ in (2.1) implies that A can have a jump only across the stationary 0-contact and is constant to the left and to the right of it. Therefore, away from the 0-contact the term $p\,\partial A/\partial x$ disappears, and we are left with a conservation law there. Then, for a shock wave with nonzero speed, we can use the usual Rankine–Hugoniot conditions.

Using the expression for the eigenvector $\mathbf{r}_0$, we can find the following two 0-Riemann invariants, which are constant across the 0-contact,

$$\begin{aligned} \eta &= \text{const}, \\ \frac{v^2}{2} + \frac{c^2}{\gamma - 1} &= \text{const}\,. \end{aligned} \tag{2.11}$$

In order to find the third one, we can, e.g., rewrite the system (2.1) in matrix form,

$$\mathbf{u}_t + \mathbf{B}(\mathbf{u})\mathbf{u}_x = 0,$$

where $\mathbf{u}$ is given by (2.2), and calculate the eigenvectors of $\mathbf{B}$. Then the third 0-Riemann invariant is

$$A\rho v = \text{const}\,. \tag{2.12}$$

The three relations (2.11), (2.12) express the constancy of entropy, Bernoulli's law, and conservation of mass, respectively.

There exist at most two solutions to the system (2.11), (2.12); see [2]. We determine which solution will be admissible using the following criterion.

DEFINITION 2.1 (evolutionarity criterion). *Consider a discontinuity Σ in a physical flow, which is governed by a $d \times d$ hyperbolic system. Denote the number of characteristics incoming to Σ by n and coinciding with Σ by c. Further, denote the number of unknown variables on both sides of Σ together with the speed of Σ by $N = 2d + 1$, and the number of relations across Σ by m. Then Σ is called* evolutionary *if*

$$N = n + c + m.$$

For the evolutionary discontinuity Σ, all N variables on it can be found using $n+c$ relations along the incoming and coinciding characteristics, and m relations across Σ. Therefore, Σ is well determined in the flow; i.e., it *evolves* in time.

The notion of *evolutionarity* goes back to at least Landau and Lipschitz [16, paragraph 88], who studied the stability of shock waves in gas dynamics. Evolutionary discontinuities are discussed in the context of magnetohydrodynamics [11] and two-phase flow [4].

In order for a contact discontinuity to be evolutionary, the number of characteristics impinging on it from the one side must be equal to the number of characteristics leaving it from the other side. For a 0-contact in the solution to the Riemann problem (2.1), (2.4), this is equivalent to the fact that the eigenvalues (2.9) on both sides of the 0-contact do not change their sign.

For a strictly hyperbolic system, the evolutionarity criterion is equivalent to the Lax shock condition. For resonant hyperbolic systems, i.e., for systems of type (2.1), it is equivalent to the criterion of Isaacson and Temple [14, 15]; see also Goatin and LeFloch [12]. For proofs of the above statements, see [2].

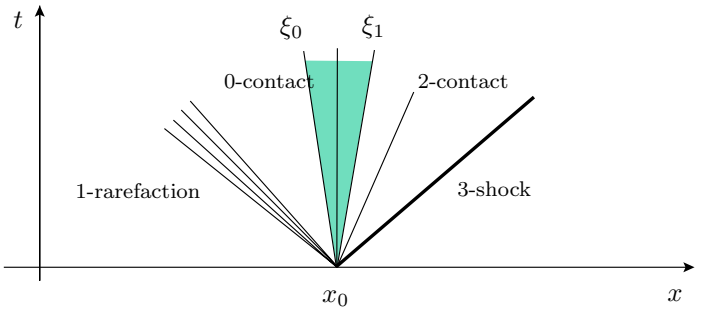


FIG. 1. *A typical Riemann problem for the* 1*D Euler equations in a duct.*

3. Weak solution to the Riemann problem. As we have mentioned above, for the Riemann problem (2.1), (2.4) the nonconservative term $p\,\partial A/\partial x$ plays a role only across one wave in the solution, the stationary 0-contact. In the rest of domain, A is constant and equal to its left or right value. Therefore, everywhere away from the stationary contact the system (2.1) reduces to the usual 1D Euler equations

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0, \tag{3.1}$$

where

$$\mathbf{u} = \begin{bmatrix} \rho \\ \rho v \\ \rho E \end{bmatrix}, \qquad \mathbf{f}(\mathbf{u}) = \begin{bmatrix} \rho v \\ \rho v^2 + p \\ v(\rho E + p) \end{bmatrix}. \tag{3.2}$$

For the system (3.1), we can define a weak solution in the usual manner.

On the other hand, across the stationary contact, the relations (2.11), (2.12) hold. Let us look for a conservative system of equations such that the Rankine–Hugoniot conditions for this system are exactly (2.11), (2.12). Remember that across the stationary contact the flow is isentropic; see (2.11). As long as some other waves do not coincide with the stationary contact, we can choose a small sector around the stationary contact where the flow is isentropic, too. Consider Figure 1 for a typical Riemann problem. The solution in the sector, bounded by the rays ξ_0 and ξ_1 around the stationary contact, is governed by the relations (2.11), (2.12). In the left and right parts of this sector, system (2.1) is equivalent to the system

$$\begin{aligned} A_t &= 0, \\ (A\rho)_t + (A\rho v)_x &= 0, \\ (A\rho E)_t + (Av(\rho E + p))_x &= 0, \\ \eta_t + v\eta_x &= 0. \end{aligned} \tag{3.3}$$

Note that the last equation in (3.3) is trivially satisfied everywhere in the sector, since the flow is isentropic there. Therefore, system (3.3) may be rewritten as

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x = 0, \tag{3.4}$$

where

$$\mathbf{U} = \begin{bmatrix} A \\ A\rho \\ A\rho E \end{bmatrix}, \qquad \mathbf{F}(\mathbf{U}) = \begin{bmatrix} 0 \\ A\rho v \\ Av(\rho E + p) \end{bmatrix}. \tag{3.5}$$

The Rankine–Hugoniot conditions for this system across a zero-speed discontinuity, augmented with the condition $\eta = \text{const}$, are exactly the relations (2.11), (2.12). Since the system (3.4) is in divergence form, we can use the usual definition of a weak solution for it. Note that the approach we have used here is exactly the same as that of [4].

Consider a system of conservation laws

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0, \tag{3.6}$$

where the pair $(\mathbf{u}, \mathbf{f}(\mathbf{u}))$ are either $(\mathbf{u}, \mathbf{f}(\mathbf{u}))$ in (3.2) or $(\mathbf{U}, \mathbf{F}(\mathbf{U}))$ in (3.5), and let us restrict ourselves to self-similar solutions $\mathbf{u}(x,t) = \mathbf{u}(\xi)$, $\xi = (x - x_0)/t$ of that system. Then, for the smooth solutions, system (3.6) is equivalent to

$$-\mathbf{u}_\xi \xi + \mathbf{f}(\mathbf{u})_\xi = 0. \tag{3.7}$$

Consider the Riemann problem for (3.7); i.e., augment (3.7) with constant initial data

$$\mathbf{u}(-\infty) = \mathbf{u}_L, \qquad \mathbf{u}(\infty) = \mathbf{u}_R. \tag{3.8}$$

If we multiply (3.7) by a test function $\phi \in C_0^1(]\xi_0, \xi_1[)$ and integrate over all ξ, we get

$$\int_{\xi_0}^{\xi_1} (\mathbf{u}(\phi\,\xi)_\xi - \mathbf{f}(\mathbf{u})\phi_\xi)\, d\xi = 0. \tag{3.9}$$

Now $\mathbf{u}$ does not need to be differentiable anymore, and we can use it to define a weak solution to the Riemann problem (3.7), (3.8). Remember, that *locally* the *non-conservative* system (2.1) can be reduced to a *conservative* one, either (3.1) or (3.4). Thus, we can give a definition of a *global* weak solution to the Riemann problem (2.1), (2.4) as a composition of the weak solutions to the conservative systems (3.1) and (3.4).

DEFINITION 3.1. *Consider a sector, bounded by the rays ξ_0 and ξ_1, such that the stationary contact lies in it, and assume that it is the only ray of discontinuity there. Then, a function $\mathbf{u} = \mathbf{u}(\xi) \in L^\infty_{\text{loc}}(\mathbb{R})$ is called a* weak solution of the Riemann problem (2.1), (2.4) *if for any small $\epsilon > 0$ the following hold:*

1. *To the left of ξ_0, i.e., $\xi \in]-\infty, \xi_0]$,*

$$\int_{-\infty}^{\xi_0} (\mathbf{u}(\phi\,\xi)_\xi - \mathbf{f}(\mathbf{u})\phi_\xi)\, d\xi = 0 \quad \textit{for all } \phi \in C_0^1(]-\infty, \xi_0 + \epsilon[),$$

$\mathbf{u}(\xi) = \mathbf{u}(x,t)$, $\mathbf{f}(\mathbf{u}) = \mathbf{f}(\mathbf{u})$, and $\mathbf{u}$, $\mathbf{f}(\mathbf{u})$ are given by (3.2).

2. *To the right of ξ_1, i.e., $\xi \in [\xi_1, \infty[$,*

$$\int_{\xi_1}^{\infty} (\mathbf{u}(\phi\xi)_\xi - \mathbf{f}(\mathbf{u})\phi_\xi)\, d\xi = 0 \quad \textit{for all } \phi \in C_0^1(]\xi_1 - \epsilon, \infty[),$$

$\mathbf{u}(\xi) = \mathbf{u}(x,t)$, $\mathbf{f}(\mathbf{u}) = \mathbf{f}(\mathbf{u})$, and $\mathbf{u}$, $\mathbf{f}(\mathbf{u})$ are given by (3.2).

3. *Inside of the sector, bounded by ξ_0 and ξ_1, i.e., $\xi \in [\xi_0, \xi_1]$,*

$$\int_{\xi_0}^{\xi_1} (\mathbf{u}(\phi\,\xi)_\xi - \mathbf{f}(\mathbf{u})\phi_\xi)\, d\xi = 0 \quad \textit{for all } \phi \in C_0^1(]\xi_0 - \epsilon, \xi_1 + \epsilon[),$$

$\mathbf{u}(\xi) = \mathbf{U}(x,t)$, $\mathbf{f}(\mathbf{u}) = \mathbf{F}(\mathbf{U})$, and $\mathbf{U}$, $\mathbf{F}(\mathbf{U})$ are given by (3.5).

Remark 1. Note that, in the sector bounded by $\xi_0 - \epsilon$ and $\xi_0 + \epsilon$, statements 1 and 3 coincide, and in the sector bounded by $\xi_1 - \epsilon$ and $\xi_1 + \epsilon$, points 2 and 3 coincide.

Remark 2. In [10], Dal Maso, LeFloch, and Murat introduce a notion of the nonconservative product and give a definition of a weak solution to a general nonconservative system on its basis. In particular, this applies also to the system (2.1). In contrast to the definition of [10], we have used some physical observations in Definition 3.1, which are valid only for systems of a certain structure like (2.1) or for the Baer–Nunziato model we considered in [4]. Therefore, Definition 3.1 might be helpful in choosing a criterion for a physically admissible solution to the Riemann problem for such systems.

4. Nonuniqueness of the Riemann solution. It appears that the solution to the Riemann problem (2.1), (2.4) is in general nonunique; for the same left and right states $\mathbf{u}_L$, $\mathbf{u}_R$, one can get completely different Riemann solutions. As was pointed out by Isaacson and Temple [15], the reason for this behavior is that system (2.1) is nonstrictly hyperbolic and nonconservative. For system (2.1) without the term $p\,\partial A/\partial x$, i.e., for a nonstrictly hyperbolic *conservative* system, they showed that the Riemann solution is unique in a neighborhood of a state where one of the nonlinear wave speeds vanishes. On the other hand, there are examples of nonstrictly hyperbolic systems that are conservative but still have a nonunique Riemann solution; see, e.g., Dafermos [9, Chapter IX].

The solution to the Riemann problem (2.1), (2.4) becomes nonunique when the mutual position of the waves, configuration of the Riemann problem, can change; see Figure 2 for the four possible cases. The problem of nonuniqueness arises from the fact that for certain sets of initial data more than one configuration is possible. In this light it makes sense to consider the conditions that lead to the different configurations of the Riemann problem.

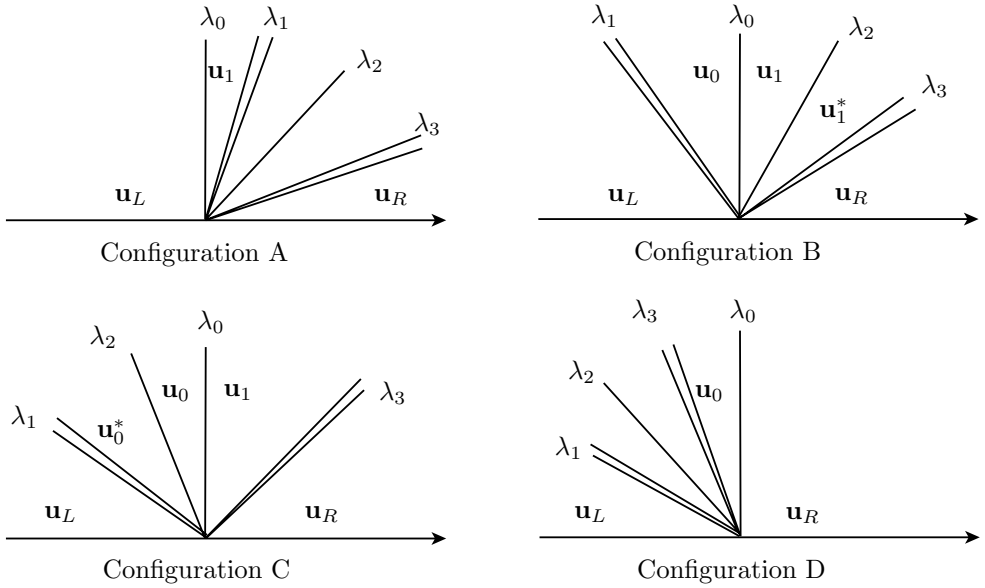


FIG. 2. *Possible configurations of the Riemann problem.*

In what follows, we will make extensive use of the wave curves for the Riemann problem (2.1), (2.4). These are the curves in the (v,p)-plane which represent the states which can be connected to $\mathbf{u}_L$, $\mathbf{u}_R$ by the admissible waves. It can be shown that the 1-curve, shock or rarefaction, is strictly increasing, and the 3-curve is strictly decreasing; see, e.g., Godlewski and Raviart [13]. Further, the left and right states of the 2-wave are projected to the same point in the (v,p)-plane. Therefore, the Riemann problem for the usual Euler equations (2.1), (2.4) with $A_L = A_R$ has a unique solution; see, e.g., [13, Theorem 3.1, p. 134].

In the solution of the Riemann problem with $A_L \neq A_R$, a jump appears across the stationary contact λ_0; i.e., there is a new wave in the solution of the Riemann problem. In [4], we have shown that the flow inside this wave is analogous to the stationary isentropic flow in the converging-diverging nozzle. For convenience, we repeat the main consequences of this fact here as follows:

- the stationary contact can be viewed as a porous film of infinitesimal thickness;
- each pore is a converging-diverging nozzle, and the cross sections on each side of it are A_L and A_R, respectively;
- the velocity inside a pore does not change its sign; moreover, if the flow is sub(super)sonic at the inlet, it is also sub(super)sonic at the outlet;
- in the direction of increasing cross section, the gas flow is accelerated and expanded when it is supersonic, and decelerated and compressed when it is subsonic.

For a given left state $\mathbf{u}_0$ of the stationary contact we can represent the 0-wave curve parametrically by A as follows:

$$\begin{cases} \rho = \rho(A;\mathbf{u}_0), \\ v = v(A;\mathbf{u}_0), \\ p = p(A;\mathbf{u}_0), \end{cases} \tag{4.1}$$

where the states must satisfy (2.11). The 3D curve (4.1) will be regular if the corresponding derivatives are continuous, therefore locally bounded, and nonzero simultaneously, i.e., if the tangent vector does not vanish,

$$\left(\frac{\partial\rho}{\partial A}, \frac{\partial v}{\partial A}, \frac{\partial p}{\partial A}\right) \neq 0. \tag{4.2}$$

We will formulate the following lemma under these conditions; later on, we will discuss situations when they are violated.

LEMMA 4.1. *Consider the Riemann problem* (2.1), (2.4) *with the stiffened gas EOS* (2.3), *and denote the states connected by the* 0*-wave by* $\mathbf{u}_0$ *and* $\mathbf{u}_1$. *Assume that the conditions* (4.2) *are fulfilled. The flow velocity* v *inside the* 0*-wave does not change its sign, and is either subsonic or supersonic everywhere in the flow. Denote its* signed Mach number *by* $\mathrm{M} = \frac{v}{c}$. *Then for the* 0*-wave curve* (4.1) *the following statements are true:*

1. *The* 0*-wave curve is strictly increasing (decreasing) in* p *if* $v < 0$ (> 0).
2. *The* 0*-wave curve is convex (concave) with respect to* p *if* $|\mathrm{M}| > 1$ (< 1).
3. *For increasing (decreasing) velocities and pressures in* $\mathbf{u}_0$, *the velocities and pressures in* $\mathbf{u}_1$ *also increase (decrease) and vice versa.*

4. *For the states* $\mathbf{u}_0$ *and* $\mathbf{u}_1$*,*

$$\begin{cases} \rho_0 \to \bar{\rho}, \\ v_0 \to 0, \\ p_0 \to \bar{p}, \end{cases} \iff \begin{cases} \rho_1 \to \bar{\rho}, \\ v_1 \to 0, \\ p_1 \to \bar{p} \end{cases}$$

for all $\bar{\rho}, \bar{p} > 0$.

Proof. We take (4.1) and fix $\mathbf{u}_0$ so that ρ, v, and p depend only on A. Since the flow inside the 0-wave is analogous to the converging-diverging flow (see [4]), we take the following relations from Courant and Friedrichs [7, (145.05) and (145.08)]:

$$\frac{dA}{A} + \frac{d\rho}{\rho} + \frac{dv}{v} = 0, \tag{4.3}$$

$$\frac{dA}{A} = \left(\frac{v^2}{c^2} - 1\right)\frac{dv}{v}, \tag{4.4}$$

where A is the variable cross section of a pore (see above) and ρ, c, and v are the corresponding parameters of the flow in the pore. Then (4.4) leads to

$$\frac{dv}{dA} = \frac{vc^2}{A(v^2 - c^2)}. \tag{4.5}$$

By the definition of the sound speed

$$\left.\frac{dp}{d\rho}\right|_{\eta} = c^2, \quad \text{i.e.,} \quad \frac{d\rho}{\rho} = \frac{dp}{\rho c^2}.$$

Substituting this into (4.3), we obtain

$$\frac{dp}{dA} = \rho c^2\left(-\frac{1}{A} - \frac{1}{v}\frac{dv}{dA}\right) = -\frac{\rho v^2 c^2}{A(v^2 - c^2)}. \tag{4.6}$$

Analogously,

$$\frac{d\rho}{dA} = -\frac{\rho v^2}{A(v^2 - c^2)}. \tag{4.7}$$

Now we see when the 0-wave curve will be regular, i.e., the conditions (4.2) will be fulfilled. From (4.5), (4.6), and (4.7) it follows that this will happen when either

$$v \neq 0 \qquad \text{or} \qquad |v| \neq c. \tag{4.8}$$

From now on, when discussing 0-wave curves, we will always assume that the conditions (4.8) are fulfilled unless stated otherwise.

Consider the system of ordinary differential equations (4.5), (4.6), (4.7):

$$\frac{d}{dA}\begin{pmatrix}\rho \\ v \\ p\end{pmatrix} = \frac{1}{A(v^2 - c^2)}\begin{pmatrix}-\rho v^2 \\ vc^2 \\ -\rho v^2 c^2\end{pmatrix} \tag{4.9}$$

with the initial data

$$\begin{cases} \rho(A_0) = \rho_0, \\ v(A_0) = v_0, \\ p(A_0) = p_0. \end{cases}$$

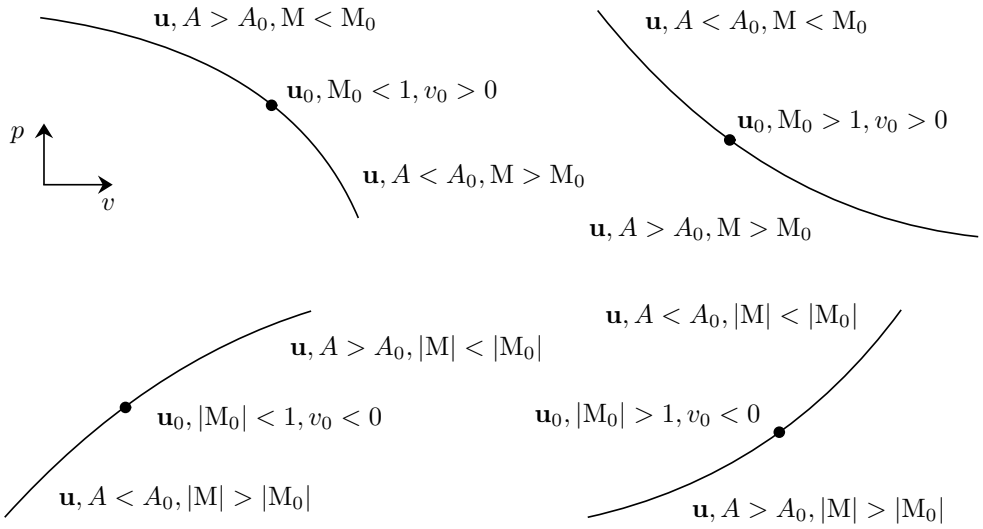


FIG. 3. *The* 0-curves *in the* (v, p)*-plane, parametrized with* A*, connect the states* $\mathbf{u}_0$ *and* $\mathbf{u}$*. Depending on* v_0 *and* M_0*, the curves exhibit different behavior.*

A straightforward calculation shows that, under conditions (4.8), the functions on the right-hand side of (4.9) are differentiable for $A > 0$. Then by the existence and uniqueness theorem there exists a unique integral curve of (4.9) as long as the right-hand side is Lipschitz continuous. This curve is nothing else but the 0-wave curve, passing through the point $\mathbf{u}_0$. Note that all states with $v = 0$ are stationary points of the system (4.9). Therefore, the solutions may approach such states only asymptotically.

From (4.5) and (4.6) we obtain

$$\frac{dp}{dv} = -\rho v,$$

and thus the statement 1 is proved.

For the proof of the statement 2 we calculate, using (4.5) and (4.6),

$$\frac{d^2p}{dv^2} = -\frac{d\rho}{dv}v - \rho = \rho\left(\frac{v^2}{c^2} - 1\right),$$

which gives the desired result. The possible 0-waves are presented in Figure 3.

Using $\eta = \frac{p+\pi}{\rho^\gamma}$ in the relations (2.11) across the 0-wave, we get

$$v_1 = v_0 \frac{A_0}{A_1}\left(\frac{p_0}{p_1}\right)^{1/\gamma},$$
$$p_1 = p_0\left(\frac{A_0 v_0}{A_1 v_1}\right)^{\gamma}.$$

Note that v_0 and v_1 always have the same sign; cf. (2.11). Differentiating the above

equations with respect to v_0 and p_0, respectively, we get

$$\frac{\partial v_1}{\partial v_0} = \frac{A_0}{A_1}\left(\frac{p_0}{p_1}\right)^{1/\gamma} > 0,$$
$$\frac{\partial p_1}{\partial p_0} = \left(\frac{A_0 v_0}{A_1 v_1}\right)^{\gamma} > 0,$$

thus proving statement 3.

For statement 4, it is enough to prove only "$\Longrightarrow$", since the statement is symmetric with respect to subscripts 0 and 1. Using $\eta = \frac{p+\pi}{\rho^\gamma}$ and $c^2 = \frac{\gamma(p+\pi)}{\rho}$ in the relations (2.11), we obtain

$$A_0(p_0+\pi)^{1/\gamma} v_0 = A_1(p_1+\pi)^{1/\gamma} v_1, \tag{4.10}$$

$$\frac{p_0+\pi}{\rho_0^\gamma} = \frac{p_1+\pi}{\rho_1^\gamma}, \tag{4.11}$$

$$\frac{v_0^2}{2} + \frac{\gamma\eta^{1/\gamma}(p_0+\pi)^{1-1/\gamma}}{\gamma-1} = \frac{v_1^2}{2} + \frac{\gamma\eta^{1/\gamma}(p_1+\pi)^{1-1/\gamma}}{\gamma-1}. \tag{4.12}$$

First, we wish to show that p_1 remains bounded, i.e., $p_1 \not\to \infty$. We prove this by contradiction, i.e., assuming that p_1 is unbounded. Then by statement 1 the pressure p_1 must exceed p_0, $p_1 > p_0$. Using that the left-hand side of (4.12) is bounded, the estimate

$$\frac{v_1^2}{2} + \frac{\gamma\eta^{1/\gamma}(p_1+\pi)^{1-1/\gamma}}{\gamma-1} > \frac{\gamma\eta^{1/\gamma}(p_1+\pi)^{1-1/\gamma}}{\gamma-1}$$

will give us the desired result that $p_1 < \text{const}$, since γ, $\pi > 0$ are constants, and η is constant along the 0-curve.

Now (4.10) implies that $v_1 \to 0$. Using this in (4.12), we get $p_1 \to \bar{p}$. Finally, by (4.11) we also have $\rho_1 \to \bar{\rho}$, which proves statement 4. ∎

Let us represent in the (v,p)-plane the possible scenarios of the solution of the Riemann problem (2.1), (2.4) which can lead to a solution in the form of Configuration B; see Figure 2. For this configuration, the 0-wave is next to the 1-wave. In the (v,p)-plane this means that the possible 0-curves necessarily start from the 1-wave curve. The correct 0-wave curve, i.e., the one which gives the solution to the Riemann problem (2.1), (2.4), connects the 1- and 3-curves in the (v,p)-plane. Figures 4 and 5 represent the wave curves for Configuration B.

Denote the states connected by the 0-curve by $\mathbf{u}_0$ and $\mathbf{u}_1$, and let $\mathbf{u}_0$ lie on the 1-curve. Observe that not all states $\mathbf{u}_0$ come into consideration. Indeed, by the definition of Configuration B, the velocities in the states $\mathbf{u}_0$ and $\mathbf{u}_1$ must be nonnegative. In Configuration B the 1-wave has a nonpositive velocity, and therefore the right-hand velocity $v_0 - c_0$ of its characteristic family must also be nonpositive, $v_0 - c_0 \le 0$. In terms of the Mach number, this means that $\mathrm{M}_0 \le 1$. Therefore, the candidates for $\mathbf{u}_0$ have necessarily $0 \le \mathrm{M}_0 \le 1$. This gives rise to the following definition.

DEFINITION 4.2. *Consider the Riemann problem* (2.1), (2.4). *Let us call the* 3*-curve crossing the* 1*-curve in the point with* $v = 0$ *the* left-bounding 3-curve. *Further, the following hold:*

1. *If* $A_L < A_R$, *consider the state* $\mathbf{u}_0$ *on the* 1*-curve with* $\mathrm{M}_0 = 1$, *connected with the state* $\mathbf{u}_1$ *with* $\mathrm{M}_1 < 1$ *by the* 0*-wave. Let us call the* 3*-curve passing through* $\mathbf{u}_1$ *the* right-bounding 3-curve*; see Figure* 4.

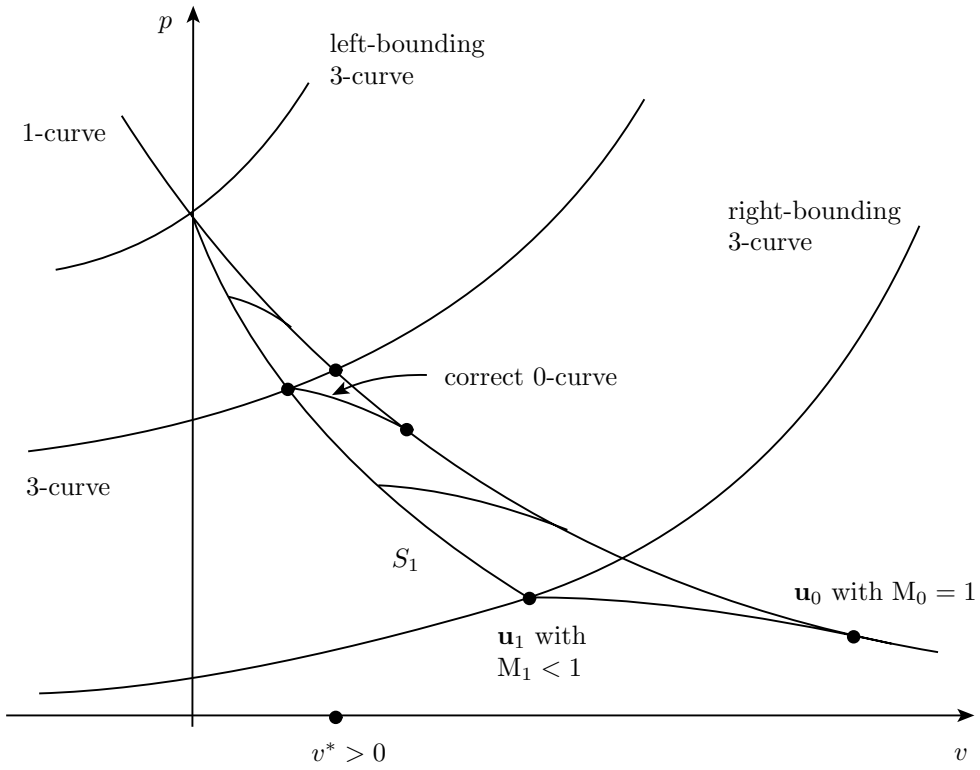


FIG. 4. *The wave curves for Configuration B with* $A_L < A_R$.

2. *If* $A_L > A_R$, *consider the state* $\mathbf{u}_0$ *on the* 1*-curve with* $\mathrm{M}_0 < 1$, *connected with the state* $\mathbf{u}_1$ *with* $\mathrm{M}_1 = 1$ *by the* 0*-wave. Let us call the* 3*-wave passing through this state* $\mathbf{u}_1$ *the* right-bounding 3-curve*; see Figure* 5.

Remark 3. We cannot use the results of Lemma 4.1 in Definition 4.2, since the parametrization (4.1) of the 0-wave will be singular for $\mathrm{M} = 1$; cf. (4.8). However, taking v as the 0-curve parameter in a neighborhood of sonic points, i.e., replacing (4.9) by the system

$$\frac{d}{dv}\begin{pmatrix} A \\ \rho \\ p \end{pmatrix} = \begin{pmatrix} A\,(v^2 - c^2)/(vc^2) \\ -\rho v/c^2 \\ -\rho v \end{pmatrix},$$

one can show that the length of the 0-wave curve remains finite.

Now we are ready to establish when the solution to the Riemann problem (2.1), (2.4) in form of Configuration B is possible.

THEOREM 4.3. *Consider the Riemann problem* (2.1), (2.4) *with the stiffened gas EOS* (2.3). *If the* 1- *and* 3*-curves intersect in the point* (v^*, p^*) *with* $v^* > 0$, *then the following scenarios are possible:*

1. *If the point* (v^*, p^*) *lies between the left- and right-bounding* 3*-curves of Definition* 4.2, *then Configuration B is realizable for all* $\mathbf{u}_L$ *on the* 1- *and for all* $\mathbf{u}_R$ *on the* 3*-curve. Moreover, the solution of this kind is unique, and Configuration C for the same Riemann problem is not realizable.*

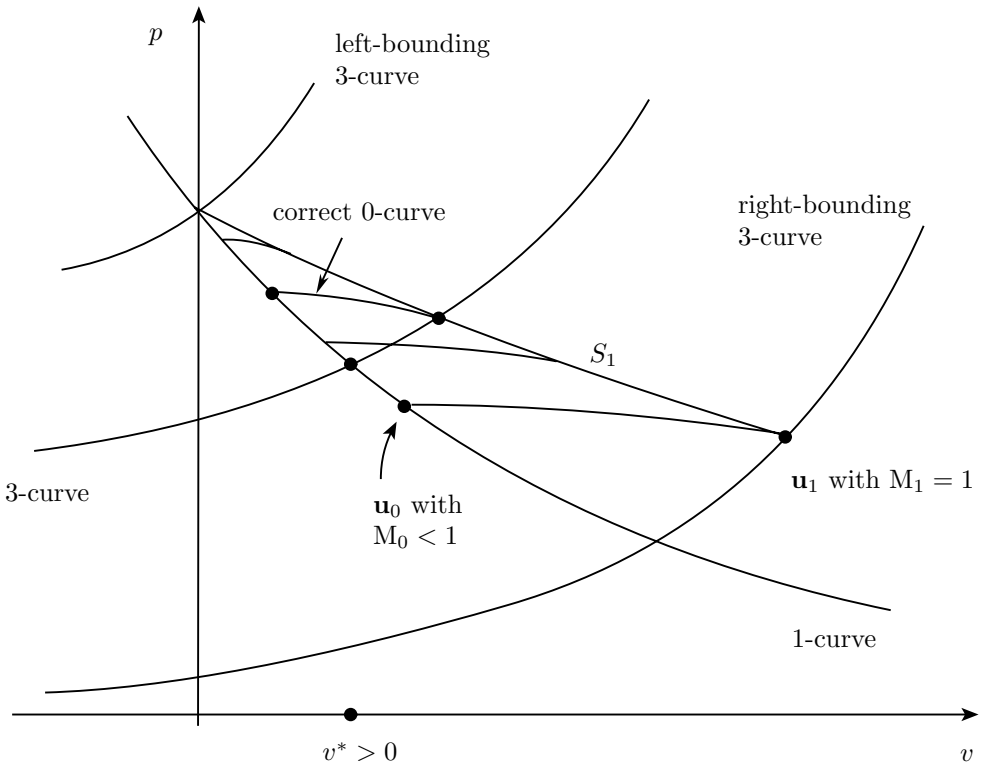


FIG. 5. *The wave curves for Configuration B with* $A_L > A_R$.

2. *If the point* (v^*, p^*) *lies to the right of the right-bounding* 3*-curve, then there exists no solution to the Riemann problem* (2.1), (2.4) *in form of Configuration B.*

3. *If Configuration A is realizable, then* $\mathrm{M}_L > 1$.

4. *If Configuration D is realizable, then* $\mathrm{M}_R < -1$.

Proof. 1. For Configuration B the states $\mathbf{u}_0$ must lie on the 1-curve. They are connected to the states $\mathbf{u}_1$ by 0-curves, such that these states $\mathbf{u}_1$ lie between the left- and right-bounding 3-curves; see Figures 4 and 5. Denote the projection of the set of all $\mathbf{u}_1$ to the (v, p)-plane by

$$S_1 = \{(v_1, p_1) : \mathbf{u}_1 = (A_1, \rho_1, v_1, p_1)^T\};$$

see Figures 4, 5. The set S_1 is defined pointwise, with each point (v_1, p_1) belonging to different integral curves of (4.9). Therefore, S_1 lies on the differentiable integral surface, obtained by taking all integral curves of (4.9) such that $\mathbf{u}_1$ are between the left- and right-bounding 3-curves.

Let us show that S_1 itself is a differentiable curve. All points $\mathbf{u}_1$ are given implicitly by the system (2.11), which can be rewritten as

$$\begin{aligned} F_1 &:= A_1\rho_1 v_1 - A_0\rho_0 v_0 = 0, \\ F_2 &:= \frac{p_1 + \pi}{\rho_1^\gamma} - \frac{p_0 + \pi}{\rho_0^\gamma} = 0, \\ F_3 &:= \frac{v_1^2}{2} + \frac{c_1^2}{\gamma - 1} - \frac{v_0^2}{2} - \frac{c_0^2}{\gamma - 1} = 0. \end{aligned} \tag{4.13}$$

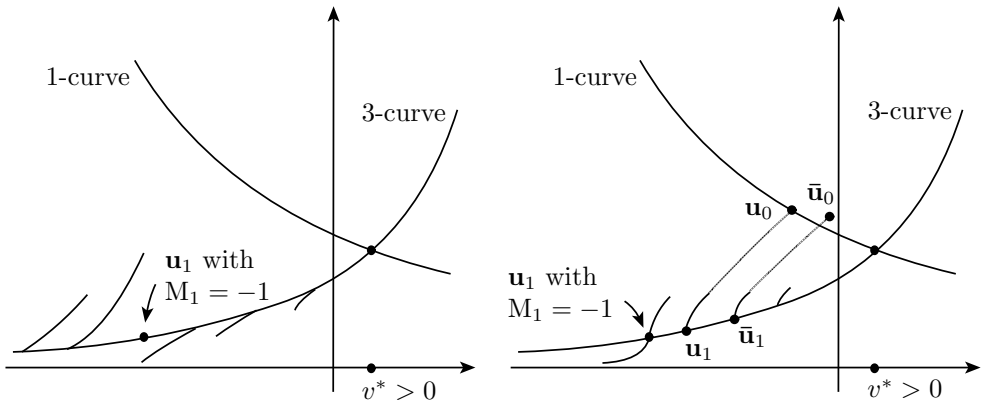


FIG. 6. *Configuration C is impossible if the* 1- *and* 3*-wave curves intersect in a state with* $v^* > 0$. *Left*: *case* $A_L < A_R$. *Right*: *case* $A_L > A_R$.

The functions F_1, F_2, F_3 are differentiable with respect to all their arguments for $\rho > 0$, and the points $\mathbf{u}_0$ lie on the *smooth* 1-wave curve. We can calculate the Jacobian determinant of the system (4.13):

$$J = \left| \frac{\partial(F_1, F_2, F_3)}{\partial(\rho_1, v_1, p_1)} \right| = \frac{A_1}{\rho_1^\gamma}(c_1^2 - v_1^2).$$

Note that $J \neq 0$ unless $\mathbf{u}_1$ lies on the right-bounding 3-curve, which is excluded by the assumptions of the theorem. Then, by the implicit function theorem, S_1 will be a differentiable curve locally at every point (v_1, p_1).

From Lemma 4.1 it follows that the mapping $(v_0, p_0) \mapsto (v_1, p_1)$ is one-to-one. Since the points $\mathbf{u}_0$ lie on the strictly decreasing 1-wave curve, S_1 will also be strictly decreasing; see statement 3 of Lemma 4.1. Also, S_1 will approach the point on the 1-curve with $v = 0$ asymptotically; see statement 4 of Lemma 4.1. Since the 3-wave is strictly increasing, there exists a unique intersection point with S_1; see Figures 4 and 5. This gives the solution to the Riemann problem (2.1), (2.4) in the framework of Configuration B.

Let us show that Configuration C for the same Riemann problem (2.1), (2.4) is impossible. Consider first the case $A_L < A_R$; see Figure 6 (left). If Configuration C were realizable, then the 0-wave would be next to the 3-wave; see Figure 2. In the (v, p)-plane, this means that the possible states $\mathbf{u}_1$ must lie on the 3-curve, and the velocities in $\mathbf{u}_1$ would be negative. This follows from the fact that the eigenvalue v_0 for the 2-wave is negative and $\operatorname{sign} v_0 = \operatorname{sign} v_1$. The states $\mathbf{u}_1$ with $\mathrm{M}_1 < -1$ are not admissible by the definition of Configuration C. Indeed, then we would have $v_1 + c_1 < 0$, so that the 3-wave is either a sonic rarefaction (i.e., there is a sign change in the characteristic speed $v + c$) or a shock with negative speed. Both these cases are excluded; see Figure 2. Note that the 0-wave, starting from the $\mathbf{u}_1$ on the 3-wave, can never intersect the 1-wave, since for the case $A_L < A_R$ it points in the opposite direction; see Figure 6 (left).

Now consider the case $A_L > A_R$; see Figure 6 (right). The only possibility for Configuration C to be realizable would be if the state $\mathbf{u}_1$ on the 3-curve with $v_1 < 0$, $|\mathrm{M}_1| < 1$ were connected with the 1-curve via the 0-curve. Assume that this is true, i.e., there exists a state $\mathbf{u}_1$ with $v_1 < 0$, $|\mathrm{M}_1| < 1$, connected to $\mathbf{u}_0$ with $v_0 < 0$ on

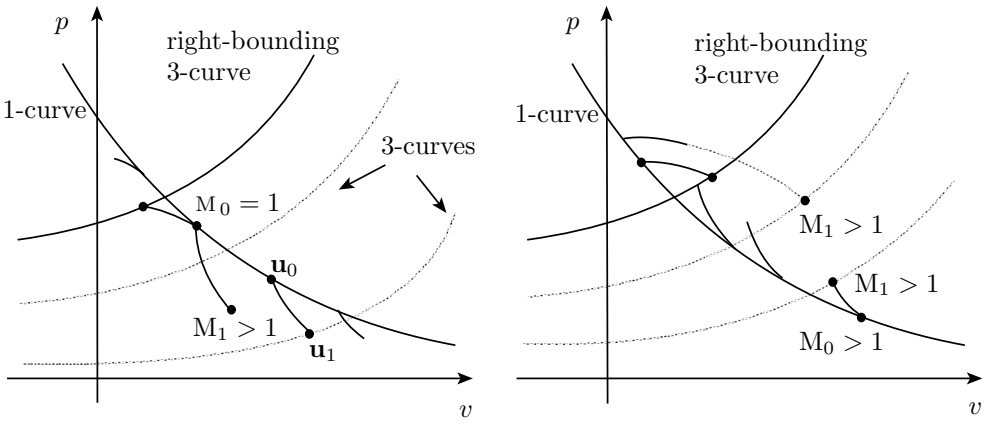


FIG. 7. *Configuration B is impossible if the* 1- *and* 3-*wave curves intersect in a state with* $v^* > 0$. *Left*: *case* $A_L < A_R$. *Right*: *case* $A_L > A_R$.

the 1-curve. Now let us move this state $\mathbf{u}_1$ towards the p-axis, so that it will become some state $\bar{\mathbf{u}}_1$, connected with $\bar{\mathbf{u}}_0$ by the 0-curve. As we move closer to the p-axis, the length L of the 0-curve, connecting $\bar{\mathbf{u}}_1$ with $\bar{\mathbf{u}}_0$, will remain positive by statement 3 of Lemma 4.1, since we have assumed that the 1- and 3-curves intersect in the point with $v^* > 0$; see Figure 6 (right). This contradicts statement 4 of Lemma 4.1, which states that L should shrink to zero. Thus, we have a unique way of connecting the 1- and 3-curves in the form of Configuration B, which is the intersection point of S_1 with the 3-curve.

2. This statement becomes obvious when we consider Figure 7. The 3-curve must lie to the right of the right-bounding 3-curve. For the case $A_L < A_R$, the state $\mathbf{u}_0$ on the 1-curve can be connected to a 3-curve only if $\mathrm{M}_0 > 1$; see Figure 7 (left). This is impossible by the definition of Configuration B. Indeed, in case $\mathrm{M}_L < 1$ we would have a sonic rarefaction in the solution of the Riemann problem. However, this is only possible if $A_L = A_R$; see [4]. In case $\mathrm{M}_L > 1$, we would have Configuration A.

For the case $A_L > A_R$, the 0-curve can connect the 1- and 3-curves if either

(i) the 0-curve crosses the right-bounding 3-curve or

(ii) the state $\mathbf{u}_L$ is supersonic with positive velocity, $\mathrm{M}_L > 1$.

For the first case, the 0-curve would connect the subsonic state $\mathbf{u}_0$ with a supersonic one, which is impossible; see the properties of the 0-wave above. For the second case we would have Configuration A.

3. We prove this statement by giving several examples, obtained with CONSTRUCT [3]. Consider the Riemann problem for (2.1) with the following initial data:

(4.14)

A_L	ρ_L	v_L	p_L	A_R	ρ_R	v_R	p_R
0.8	0.2069	3.991	0.07	0.3	0.1354	−3.1666	0.0833

closed with the EOS (2.3) with $\gamma = 1.4$ and $\pi = 0$. The wave curves for this Riemann problem are presented in Figure 8 (top). Observe that for these initial data both Configurations A and B are possible. Configuration A is realized when the left state $\mathbf{u}_L$ is connected first to the state $\bar{\mathbf{u}}_1$ with the 0-curve, and $\bar{\mathbf{u}}_1$ is then connected with $\bar{\mathbf{u}}^*$ via the 1-shock with speed $s = 0.948 > 0$. Note that if this speed were negative, then Configuration A with these initial data would be not realizable; cf. Figure 2. Since the intersection point of 1- and 3-curves lies between the left- and right-bounding waves

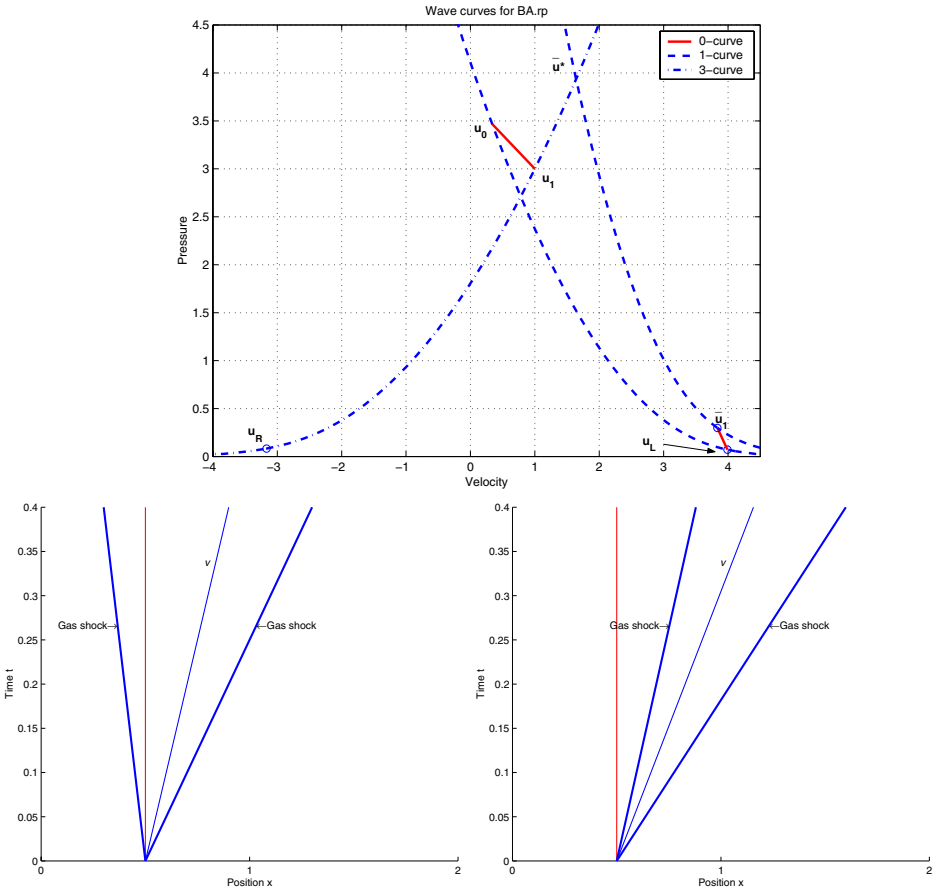


FIG. 8. *Top*: *the wave curves for the Riemann problem* (2.1), (4.14). *Bottom*: *the corresponding wave configurations B and A in the* (x,t)*-plane.*

(not shown in Figure 8), Configuration B is also possible. For this configuration, the left state $\mathbf{u}_L$ is connected with $\mathbf{u}_0$, and the latter is connected with $\mathbf{u}_1$. Both wave configurations in the (x,t)-plane are shown in Figure 8 (bottom).

However, if we slightly modify the initial data (4.14), we can easily obtain a Riemann problem with a unique solution. For instance, for the Riemann problem (2.1) with initial data

(4.15)

A_L	ρ_L	v_L	p_L	A_R	ρ_R	v_R	p_R
0.8	0.2069	3.0	0.2	0.3	0.1354	−3.1666	0.0833

,

also closed with the EOS (2.3) with $\gamma = 1.4$ and $\pi = 0$, only Configuration B is possible. Indeed, consider the wave curves for this Riemann problem in Figure 9. Again, the state $\mathbf{u}_L$ is first connected to the state $\bar{\mathbf{u}}_1$ with the 0-curve; from $\bar{\mathbf{u}}_1$, we draw the 1-curve until the intersection with the 3-curve, passing through $\mathbf{u}_R$. However, the corresponding wave will be a shock with negative speed $s = -0.198$. Therefore, Configuration A is now not possible. Configuration D is not possible for similar reasons. Since the 1- and 3-curves intersect between the corresponding left- and right-bounding 3-curves (not shown in Figure 9), Configuration B is possible for

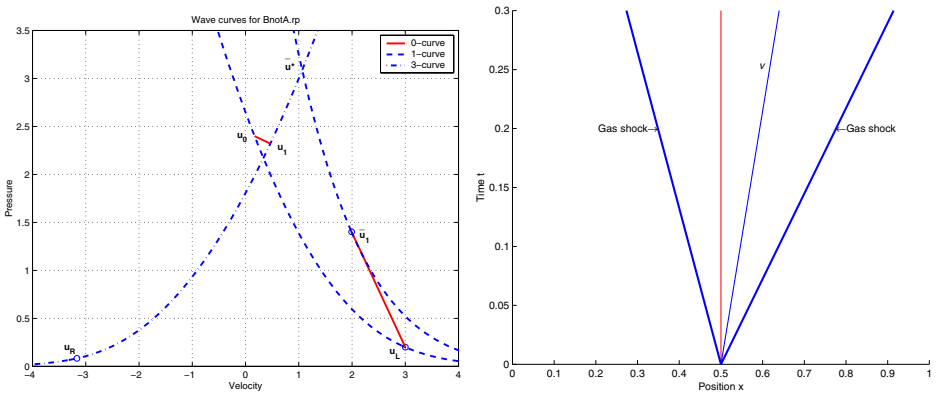


FIG. 9. *The wave curves the for the Riemann problem* (2.1), (4.15).

the Riemann problem (2.1), (4.15). The waves in the (x,t)-plane for this Riemann problem are shown in Figure 9 (right).

4. Consider the following Riemann initial data:

(4.16)

A_L	ρ_L	v_L	p_L	A_R	ρ_R	v_R	p_R
0.3	0.2	3.3	1	0.8	0.2	-4	0.07

.

Reasoning as above, we can show that the Riemann problem (2.1), (4.16), closed with the EOS (2.3) with $\gamma = 1.4$ and $\pi = 0$, has a nonunique solution in form of either Configuration B or Configuration D. The wave curves and the wave configurations are shown in Figure 10. □

For Configuration C, the results are completely analogous. The wave curves are presented in Figures 11 and 12. As for Configuration B, one can introduce the left- and right-bounding curves.

DEFINITION 4.4. *Consider the Riemann problem* (2.1), (2.4). *Let us call the* 1*-curve crossing the* 3*-curve in the point with* $v = 0$ *the* right-bounding 1-curve. *Further,*

1. *if* $A_L < A_R$, *consider the state* $\mathbf{u}_1$ *on the* 3*-curve with* $|\mathrm{M}_1| < 1$, *connected with the state* $\mathbf{u}_0$ *with* $\mathrm{M}_0 = -1$ *by the* 0*-wave. Let us call the* 1*-curve passing through this state* $\mathbf{u}_0$ *the* left-bounding 1-curve*; see Figure* 11.

2. *if* $A_L > A_R$, *consider the state* $\mathbf{u}_1$ *on the* 3*-curve with* $\mathrm{M}_1 = -1$, *connected with the state* $\mathbf{u}_0$ *with* $|\mathrm{M}_0| < 1$ *by the* 0*-wave. Let us call the* 1*-curve passing through* $\mathbf{u}_0$ *the* left-bounding 1-curve*; see* Figure 12.

Analogously to Theorem 4.3 we have the following result.

THEOREM 4.5. *Consider the Riemann problem* (2.1), (2.4) *with the stiffened gas EOS* (2.3). *If the* 1*- and* 3*-curves intersect in the point* (v^*, p^*) *with* $v^* < 0$, *then the following scenarios are possible:*

1. *If the point* (v^*, p^*) *lies between the left- and right-bounding* 1*-curves of Definition* 4.4, *then Configuration C is realizable, for all* $\mathbf{u}_L$ *on the* 1*- and all* $\mathbf{u}_R$ *on the* 3*-curve. Moreover, the solution of this kind is unique, and Configuration B for the same Riemann problem is not realizable.*

2. *If the point* (v^*, p^*) *lies to the left of the left-bounding* 1*-curve, then there exists no solution to the Riemann problem* (2.1), (2.4) *in form of Configuration C.*

3. *If Configuration A can be realized, then* $\mathrm{M}_L > 1$.

4. *If Configuration D can be realized, then* $\mathrm{M}_R < -1$.

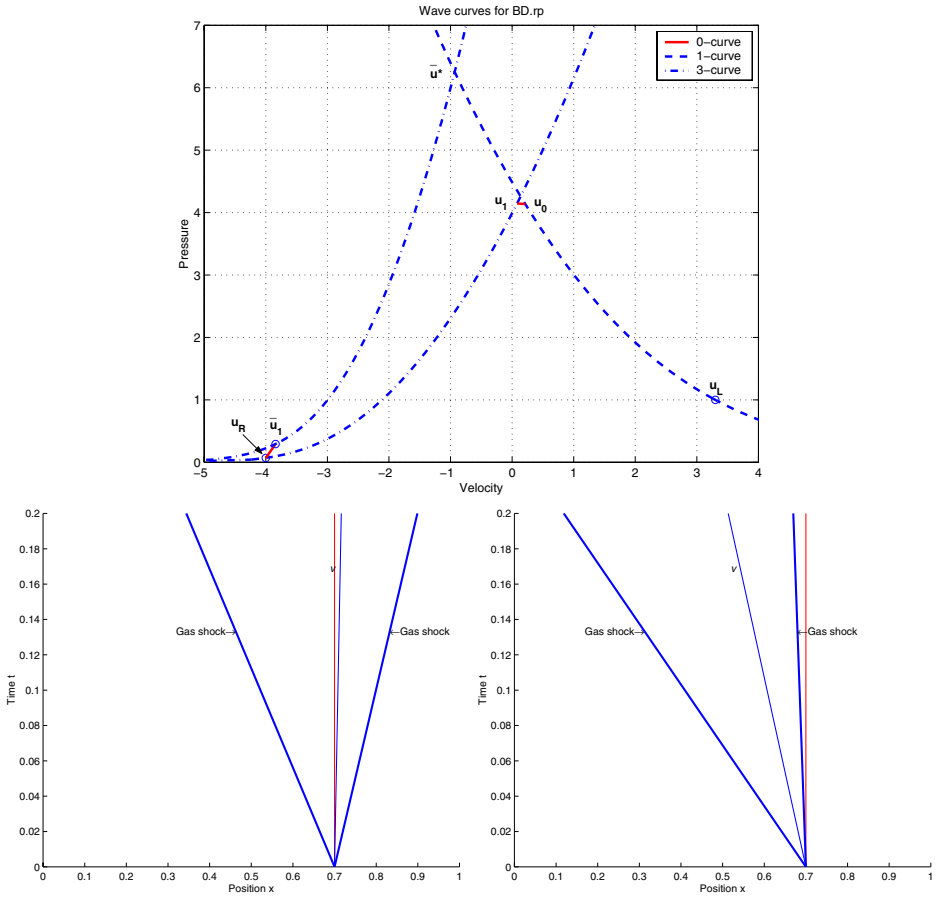


FIG. 10. *Top*: *the wave curves for the Riemann problems* (2.1), (4.16). *Bottom*: *the corresponding wave configurations B and D in the* (x,t)*-plane.*

Proof. The proof is analogous to that of Theorem 4.3. ∎

Theorem 4.3 states that the solution to the Riemann problem (2.1), (2.4) will be nonunique if for some initial data both Configurations B and A, or B and D, are realizable. Analogously, Theorem 4.5 says that for some initial data both Configurations C and A, or C and D, are possible. In all these cases the nonuniqueness appears when the mutual position of 0- and k-waves, $k = 1, 3$, changes. According to the analysis of Isaacson and Temple [15] (see also Goatin and LeFloch [12]), there can be a third wave configuration. It includes a "triple discontinuity," which consists of the 0-wave followed by a k-shock with zero speed, followed by another 0-wave. However, it is not obvious whether such a "triple discontinuity" will be evolutionary. It is not clear how to determine the number of relations across this discontinuity, as well as the number of characteristics impinging on it or leaving it. Therefore, we do not consider such "triple discontinuity" in the present work.

5. Which solution to take? To understand the origin of the nonuniqueness of the Riemann solution for the Euler equations in a duct of variable cross section, it is advantageous to consider analogous situations for other models. An immediate example is given by the usual Euler equations of gas dynamics. They are obtained

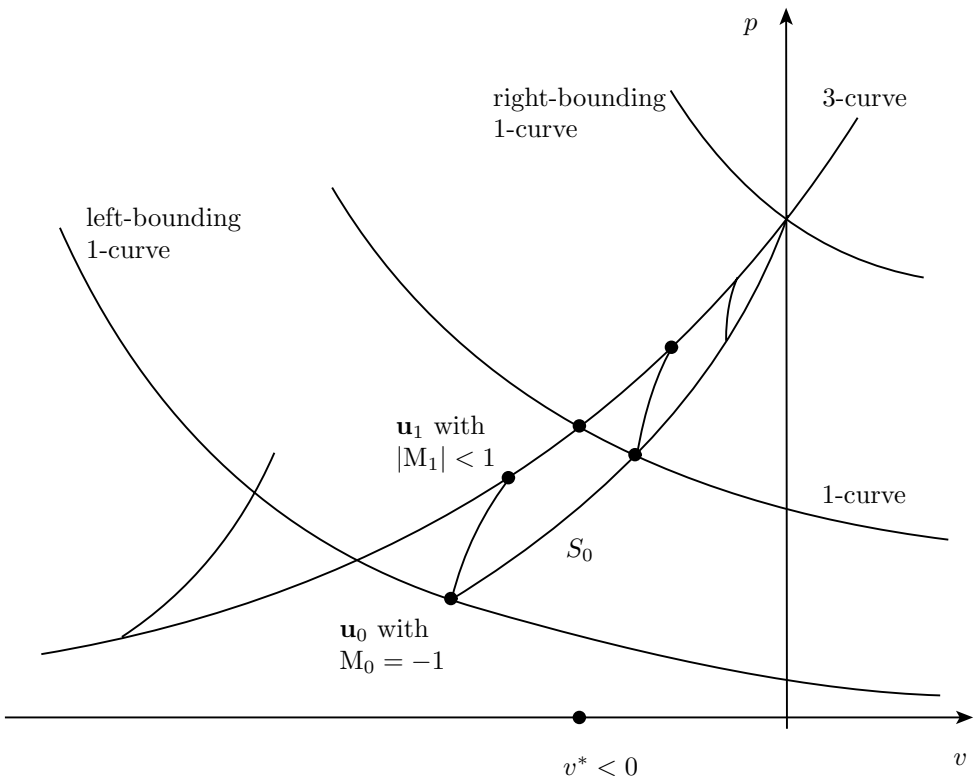


FIG. 11. *The wave curves for Configuration C with $A_L < A_R$. If the point (v^*, p^*) lies between the left- and right-bounding* 1*-curves, there exists a unique solution in the form of Configuration C.*

as the inviscid approximation to (in general viscous) fluid flows. It is well known that this approximation leads to nonunique discontinuous solutions, and therefore to nonunique solutions for Riemann problems. One needs to use an additional criterion, an entropy condition, in order to select the physically relevant solution. In fact, one possible way of obtaining an entropy condition is to add a viscous term to the Euler equations, i.e., to model the original viscous flow. Then the limit of solutions for vanishing viscosity will yield the physical entropy solution; see, e.g., Godlewski and Raviart [13].

For the Euler equations in a duct of variable cross section, the situation is somewhat analogous. In addition to neglecting viscosity, we have also neglected the multidimensional (2D or 3D) effects. In this light it is not surprising that we obtained nonuniqueness of the solutions to the Riemann problem. Roughly speaking, we have lost too much information on the truly multidimensional flow. However, one might hope to get a criterion for choosing the physically relevant solution by considering multidimensional effects, similarly to the limiting procedure for the usual Euler equations. One possible way would be to add to the system of governing equations some terms that would model these effects. For example, these terms might be obtained using the statistical ensemble averaging techniques in the spirit of recent work of Abgrall and Saurel [1]. This could be an interesting topic of future research.

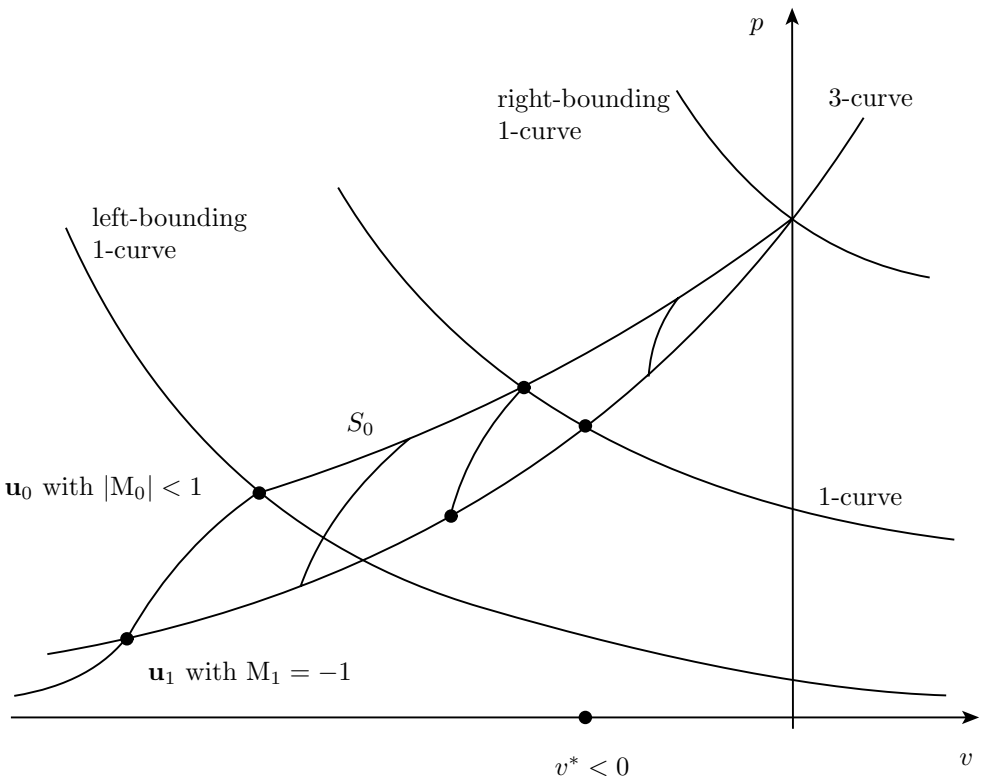


FIG. 12. *The wave curves for Configuration C with $A_L > A_R$.*

Here, we follow a more straightforward approach: We compare the results of the quasi-one-dimensional Euler equations in a duct of variable cross section with multidimensional computations of the usual Euler equations in a tube of corresponding geometry, averaged over the tube cross section. To this end, we employ the popular software package CLAWPACK provided by LeVeque [6, 17]. In the calculations below we have used CLAWPACK with a second order method and Roe's Riemann solver.

5.1. Diverse Riemann problems for the 1D model.

Nonuniqueness between Configurations A and B. Consider the 1D Riemann problem (2.1), (4.14), which has already been studied in section 4. This problem has a nonunique solution in the form of either Configuration A or Configuration B; see Figure 8. With CLAWPACK, we solve the 2D analogue of this 1D Riemann problem, i.e., the usual Euler equations in the corresponding 2D computational domain. The 2D solution on the 200×100 grid and the comparison of the averaged 2D solution with the exact solution to the 1D Riemann problem (2.1), (4.14) are shown in Figure 13 (top). We see that the 1D model slightly overestimates the 3-shock speed. Also, due to rich 2D motion in the left section of the domain, the 1D prediction of the position of 1-shock is quite approximate. The same can be said about the approximation of the flow near the jump in the cross section. However, the numerical solution clearly picks up Configuration B in the 1D solution of the Riemann problem (2.1), (4.14); see Figure 13.

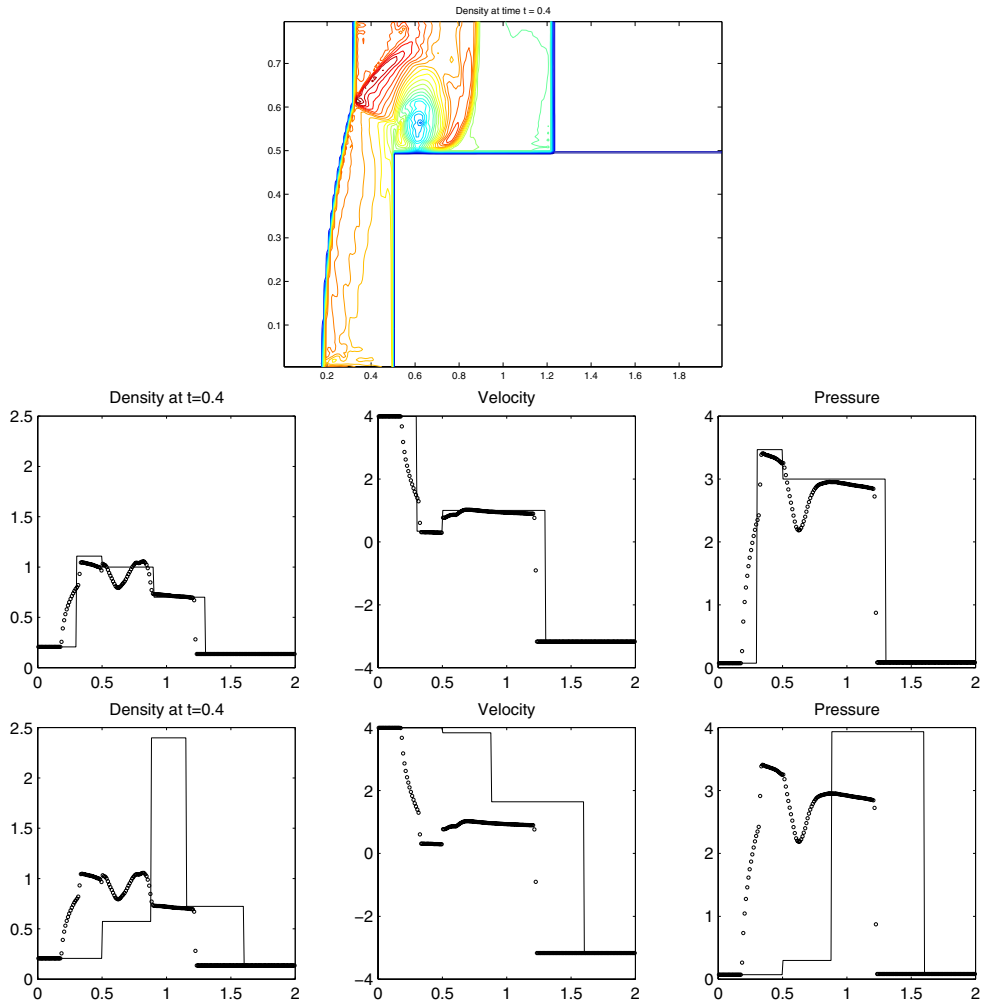


FIG. 13. *Top*: *the density contours of the* 2*D solution to the Riemann problem* (2.1), (4.14). *Middle*: *comparison of the averaged* 2*D solution (dots) with the* 1*D exact solution in the form of Configuration B (line). Bottom*: *comparison of the averaged* 2*D solution (dots) with the* 1*D exact solution in form of Configuration A (line).*

Unique solution in the form of Configuration B. As we established in section 4, the Riemann problem (2.1), (4.15) has a unique solution in form of Configuration B. The comparison of the averaged 2D solution on a 100×100 grid with the exact solution to the 1D Riemann problem is shown in Figure 14. Again, we observe that the shock speeds are slightly different; however, the main features of the 2D flow are correctly represented by the 1D model.

Nonuniqueness between Configurations B and D. The solution to the 1D Riemann problem (2.1), (4.16) is nonunique: It can be either Configuration B or Configuration D. Again, we calculate the corresponding 2D problem on a 100×100 grid and obtain the results shown in Figure 15. We see that the exact 1D solution for Configuration B perfectly fits the averaged 2D solution. On the other hand, the 1D solution for Configuration D has nothing in common with the averaged 2D solution.

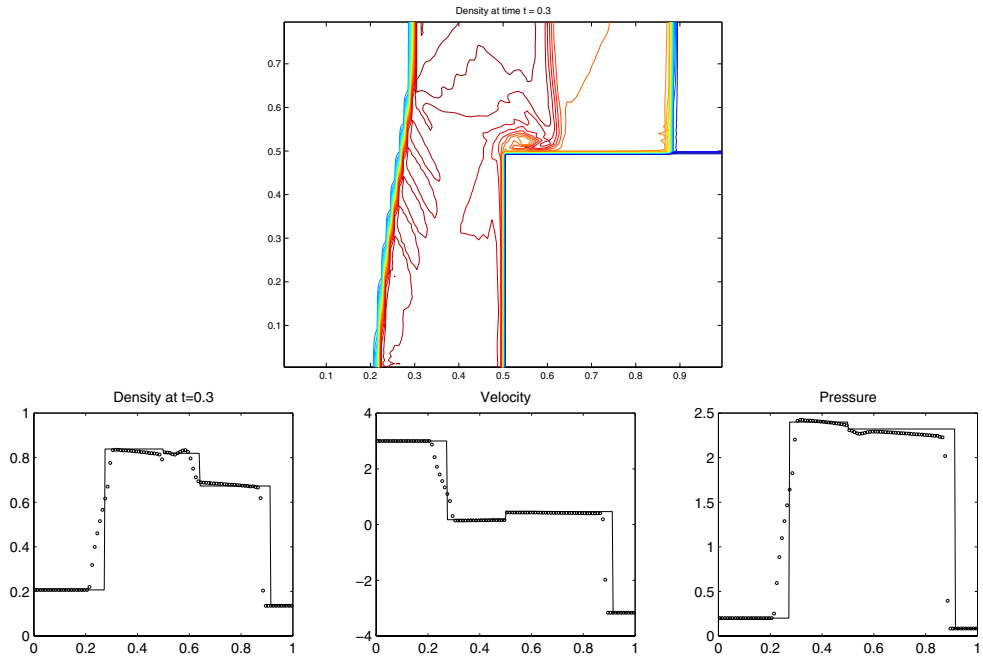


FIG. 14. *Top*: *the density contours of the* 2*D solution to the Riemann problem* (2.1), (4.15). *Bottom*: *the comparison of the averaged* 2*D solution (dots) with the* 1*D exact solution (line).*

5.2. A criterion for realizable solutions. The preceding computations show that for all configurations of Riemann problems the 1D solution can differ significantly from the 2D one. However, if the 1D solution to a Riemann problem is not unique, the 2D calculations clearly pick out one of the 1D solutions. In what follows, we will call the corresponding 1D Riemann solutions *physically relevant*. Let us investigate what distinguishes these solutions from the physically nonrelevant ones.

A classical way to exclude physically irrelevant solutions is to use a notion of entropy. However, one cannot use the entropy inequality used in the theory of conservation laws for the nonconservative system (2.1). In the particular case of the Riemann problem, one can use the approach of section 3 to define the entropy inequality. Note also that *locally* each discontinuity in the solution to the Riemann problems is entropy-satisfying; i.e., the entropy increases across shocks. However, it does not help to rule out the physically irrelevant solutions. This suggests an idea of using a *global* entropy condition.

A global entropy condition was proposed by Dafermos [8, 9], who called it the *entropy rate admissibility criterion*. It states that not only should the entropy increase but also it should be increasing at the maximum rate. The rigorous definition of this criterion can be applied only for the conservation laws. However, we can use the general idea for the nonconservative system (2.1) as well.

As a measure of the entropy increase rate, we use the jump in the isentrope $\eta = \frac{p+\pi}{\rho^\gamma}$ across shocks. These jumps are always positive. Then, our calculations show that physically relevant solutions, i.e., the ones picked out by 2D calculations, indeed have the maximal increase in entropy in comparison with the other solutions. This may be seen as an analogy of the entropy rate admissibility criterion of Dafermos [8].

Since the Euler equations in a duct can be formally obtained from the governing

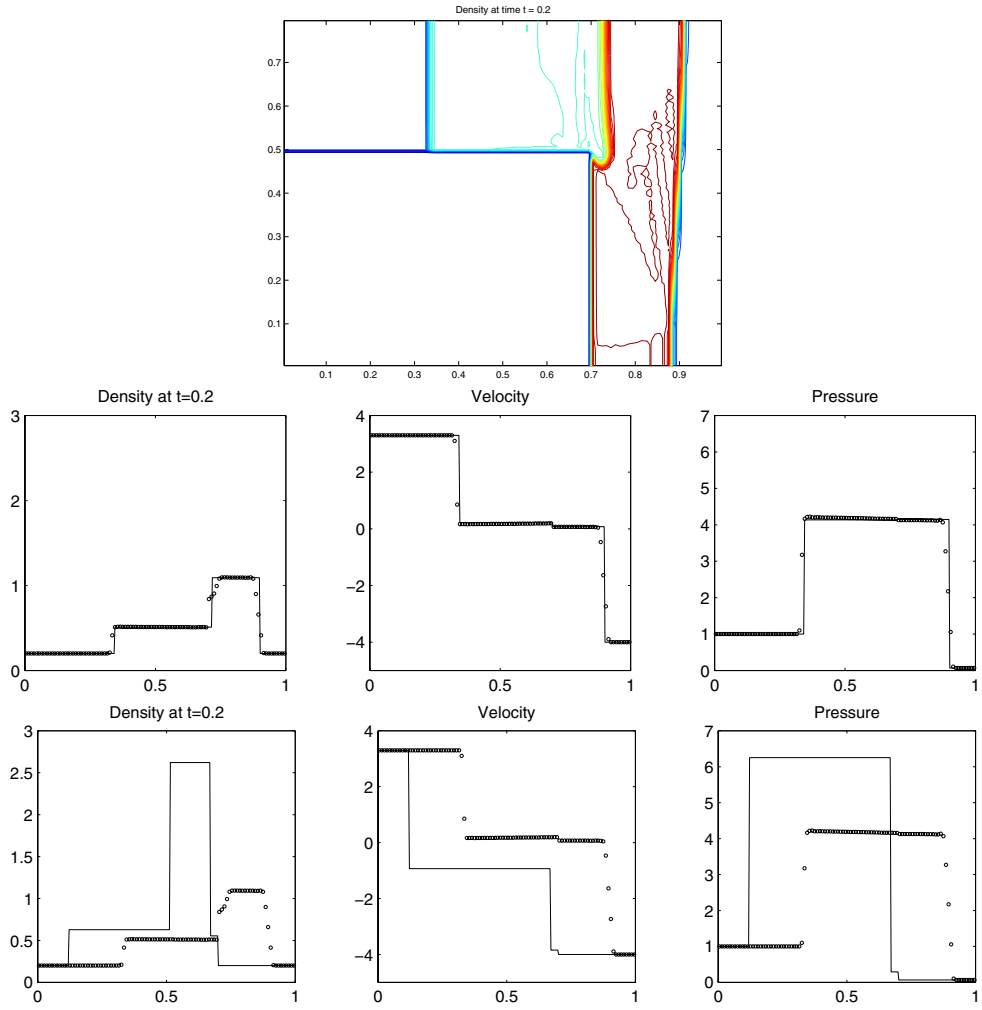


FIG. 15. *Top*: *the density contours of the* 2*D solution to the Riemann problem* (2.1), (4.16). *Middle*: *the comparison of the averaged* 2*D solution (dots) with the* 1*D exact solution in the form of Configuration B (line). Bottom*: *the comparison of the averaged* 2*D solution (dots) with the* 1*D exact solution in the form of Configuration D (line).

equations for the Baer–Nunziato model of two-phase flows [5], the solution to the Riemann problem for the Baer–Nunziato model will in general also not be unique. Therefore, the analysis of the Euler equations for a duct of variable cross section, and in particular the above criterion, can help in investigating the properties of the governing equations for the Baer–Nunziato model.

REFERENCES

[1] R. ABGRALL AND R. SAUREL, *Discrete equations for physical and numerical compressible multiphase mixtures*, J. Comput. Phys., 186 (2003), pp. 361–396.

[2] N. ANDRIANOV, *Analytical and numerical investigation of two-phase flows*, Ph.D. thesis, Institut für Analysis und Numerik, Universität Magdeburg, 2003; available athttp://www-ian.math.uni-magdeburg.de/home/andriano/diser.html.

[3] N. Andrianov, *CONSTRUCT: A collection of MATLAB routines for constructing the exact solution to the Riemann problem for the Baer–Nunziato model of two-phase flows*, available at http://www-ian.math.uni-magdeburg.de/home/andriano/CONSTRUCT.
[4] N. Andrianov and G. Warnecke, *The Riemann problem for the Baer–Nunziato model of two-phase flows*, J. Comput. Phys., 195 (2004), pp. 434–464.
[5] M. R. Baer and J. W. Nunziato, *A two-phase mixture theory for the deflagration-to-detonation transition (DDT) in reactive granular materials*, Int. J. Multiphase Flows, 12 (1986), pp. 861–889.
[6] clawpack software, available at http://www.amath.washington.edu/~claw.
[7] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Springer, New York, 1999.
[8] C. Dafermos, *The entropy rate admissibility criterion for solutions of hyperbolic conservation laws*, J. Differential Equations, 14 (1973), pp. 202–212.
[9] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer, Berlin, 2000.
[10] G. Dal Maso, P. G. LeFloch, and F. Murat, *Definition and weak stability of nonconservative products*, J. Math. Pures Appl., 74 (1995), pp. 483–548.
[11] S. A. E. G. Falle and S. S. Komissarov, *On the inadmissibility of non-evolutionary shocks*, J. Plasma Phys., 65 (2001), pp. 29–58.
[12] P. Goatin and P. G. LeFloch, *The Riemann problem for a class of resonant hyperbolic systems of balance laws*, preprint, Ecole Polytechnique, Palaiseau, France, 2003; available at http://www.math.ntnu.no/conservation/2003/008.html.
[13] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer, New York, 1996.
[14] E. Isaacson and B. Temple, *Nonlinear resonance in systems of conservation laws*, SIAM J. Appl. Math., 52 (1992), pp. 1260–1278.
[15] E. Isaacson and B. Temple, *Convergence of the* 2×2 *Godunov method for a general resonant nonlinear balance law*, SIAM J. Appl. Math., 55 (1995), pp. 625–640.
[16] L. D. Landau and E. M. Lipschitz, *Fluid Mechanics*, Pergamon Press, Oxford, 1987.
[17] R. J. LeVeque, *Wave propagation algorithms for multi-dimensional hyperbolic systems*, J. Comput. Phys., 131 (1997), pp. 327–353.
[18] M. J. Zucrow and J. D. Hoffman, *Gas Dynamics*, Vol. 2, John Wiley & Sons, New York, 1977.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 902–931

COMPLETE ELECTRODE MODEL OF ELECTRICAL IMPEDANCE TOMOGRAPHY: APPROXIMATION PROPERTIES AND CHARACTERIZATION OF INCLUSIONS*

NUUTTI HYVÖNEN†

Abstract. In electrical impedance tomography one tries to recover the spatial admittance distribution inside a body from boundary measurements. In theoretical considerations it is usually assumed that the boundary data consists of the Neumann-to-Dirichlet map; when conducting real-world measurements, the obtainable data is a linear finite-dimensional operator mapping electrode currents onto electrode potentials. In this paper it is shown that when using the complete electrode model to handle electrode measurements, the corresponding current-to-voltage map can be seen as a discrete approximation of the traditional Neumann-to-Dirichlet operator. This approximating link is utilized further in the special case of constant background conductivity with inhomogeneities: It is demonstrated how inclusions with strictly higher or lower conductivities can be characterized by the limit behavior of the range of a boundary operator, determined through electrode measurements, when the electrodes get infinitely small and cover all of the object boundary.

Key words. electrical impedance tomography, inverse boundary value problems, electrode models, variational principles

AMS subject classifications. 35R30, 35J25

DOI. 10.1137/S0036139903423303

1. Introduction.

The problem of electrical impedance tomography is as follows: Gather information about the admittance tensor σ in the elliptic equation

$$\nabla \cdot \sigma \nabla u = 0 \quad \text{in } \Omega$$

using measurements of current and potential on the boundary $\partial\Omega$. In mathematical analysis of this problem it is usually assumed that the obtainable data are all possible pairs of Neumann and Dirichlet boundary values, i.e., the linear Neumann-to-Dirichlet map. In particular, all uniqueness and reconstruction results have been formulated using this so-called *continuum model* (CM)—for more details we refer to the review paper [1]. However, when conducting real-life measurements with electrodes, one can control only the net currents through certain surface patches and measure the corresponding potentials on the electrodes, and so the real-life data consists, essentially, of a finite-dimensional linear electrode current-to-electrode voltage operator.

In this work we model the electrode measurements with the *complete electrode model* (CEM) [9], which has been shown to predict experimental data reasonably well [9] and also give fairly good numerical reconstructions for both experimental and simulated data [12], [11]. Our first goal is to show that the CEM forward problem can, actually, be seen as a Galerkin approximation of the CM forward problem, meaning that the forward solutions for both of these models can be obtained from the very same variational formulation using different function spaces. As a consequence, the forward solution of CEM, with correctly chosen electrode currents, may be considered an approximation for the forward solution of CM corresponding to a given current

*Received by the editors February 19, 2003; accepted for publication (in revised form) June 17, 2003; published electronically March 30, 2004.
http://www.siam.org/journals/siap/64-3/42330.html
†Institute of Mathematics, Helsinki University of Technology, P. O. Box 1100, FIN-02015 HUT, Finland (nuutti.hyvonen@hut.fi).

distribution—or the other way around—with the correspondence getting better as the electrodes get smaller and cover a larger portion of the object boundary. Section 2 considers these matters.

The second objective is to use the approximating link between the different forward models to modify the factorization method for characterizing inclusions, introduced and justified in [3] for electrical impedance tomography and, earlier, in [7] for inverse scattering, to the framework of the CEM. To be more precise, in section 3 the special case of constant background conductivity with inclusions of strictly higher or lower conductivities is considered. It is demonstrated how the inhomogeneities can be characterized by comparing the boundary values of a dipole-like singular solution and the range of a boundary operator, obtained through electrode measurements, as the electrodes grow in number, get infinitely small, and cover all of the object boundary.

2. Approximation properties of the CEM. In this section we aim to show that the CEM can be seen as a finite element approximation of the CM of impedance tomography. In the first subsection, we will introduce the different forward models and consider some of their basic properties. The second subsection will explain how one can approximate the forward solution of CM by the forward solution of CEM with correctly chosen input currents. In the final subsection, we will survey the resemblance between the current-to-potential boundary maps of CM and CEM.

2.1. Forward models. When performing mathematical analysis of the electrical impedance tomography problem, it is traditionally assumed that one is able to use any input current distribution from Sobolev space $H^{-1/2}$ resulting in boundary potentials of class $H^{1/2}$. On the other hand, when conducting real-world measurements, one can control only the net currents fed through a finite number of electrodes and measure the corresponding electrode potentials. In particular, one does not know the exact distribution of the current penetrating the object boundary.

2.1.1. Continuum forward model. Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, with a smooth boundary be our open bounded region of interest and let $\sigma : \Omega \to \mathbb{C}^{n\times n}$ be the corresponding admittance tensor. The forward problem of impedance tomography with continuous boundary measurements is as follows: For $f \in H_0^{-1/2}(\partial\Omega)$ find $u \in H^1(\Omega)/\mathbb{C}$ that satisfies weakly

$$\nabla \cdot \sigma\nabla u = 0 \quad \text{in } \Omega, \quad \nu \cdot \sigma\nabla u = f \quad \text{on } \partial\Omega, \tag{2.1}$$

where ν is the outer unit normal on $\partial\Omega$ and

$$H_0^{-1/2}(\partial\Omega) = \{v \in H^{-1/2}(\partial\Omega) \mid \langle v, \mathbf{1}\rangle_{L^2(\partial\Omega)} = 0\},$$

where $\langle \phi, \psi\rangle_{L^2(\partial\Omega)} = \int_{\partial\Omega} \phi\overline{\psi}dS$ denotes the dual pairing of the spaces $H^{-1/2}(\partial\Omega)$ and $H^{1/2}(\partial\Omega)$. In what follows, we also shall use this same notation for the L^2 inner product.

If it is assumed that the admittance tensor $\sigma \in \mathbb{C}^{n\times n}$ satisfies

$$\mathrm{Re}(\sigma x \cdot \overline{x}) \geq c|x|^2, \quad |\sigma x \cdot \overline{x}| \leq C|x|^2, \quad c, C > 0, \tag{2.2}$$

for all $x \in \mathbb{C}^n$ almost everywhere in Ω, then forward problem (2.1) has a unique solution that depends continuously on the boundary data.

THEOREM 2.1. *Let $f \in H_0^{-1/2}(\partial\Omega)$ and assume that inequalities* (2.2) *hold. Then forward problem* (2.1) *has a unique solution $u \in H^1(\Omega)/\mathbb{C}$, for which*

$$||u||_{H^1(\Omega)/\mathbb{C}} \leq C\,||f||_{H^{-1/2}(\partial\Omega)}.$$

Proof. For proof we refer to [10]. ∎

Before we can go any further with our analysis, we need to introduce the trace spaces on subsets of the boundary $\partial\Omega$. Below we will list only the basic definitions; for more information the reader should consult [4] and the references cited therein. For $\Gamma \subset \partial\Omega$ we define

$$H^{1/2}(\Gamma) = \{v|_\Gamma \mid v \in H^{1/2}(\partial\Omega)\}.$$

We denote the dual space of $H^{1/2}(\Gamma)$ by $\tilde{H}^{-1/2}(\Gamma)$, and note that $\tilde{H}^{-1/2}(\Gamma)$ can be identified with

$$H^{-1/2}_{\overline{\Gamma}}(\partial\Omega) = \{v \in H^{-1/2}(\partial\Omega) \mid \operatorname{supp} v \in \overline{\Gamma}\}.$$

In what follows, $\langle\cdot,\cdot\rangle_{L^2(\Gamma)}$ will denote either the dual pairing between $\tilde{H}^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$ or the $L^2(\Gamma)$ inner product. Finally, $\tilde{H}^{-1/2}_0(\Gamma)$ is defined to be the subspace of $\tilde{H}^{-1/2}(\Gamma)$ over which the dual evaluation with $\mathbf{1} \in H^{1/2}(\Gamma)$ vanishes.

Let us consider briefly the following question. If we are trying to use some given input current pattern $f \in H_0^{-1/2}(\partial\Omega)$ but we are only able to conduct current through a part of the boundary $\Gamma \subset \partial\Omega$ how much does this imperfection affect the forward solution? To begin with, we must choose how to restrict the current f onto the subset Γ; using $f|_\Gamma$ is not usually an option since all the current that goes into the object Ω must come out. Thus, in order to obtain reasonable currents on Γ, we define a L^2-orthogonal projection operator $P_1 : H_0^{-1/2}(\partial\Omega) \to \tilde{H}_0^{-1/2}(\Gamma) \subset H_0^{-1/2}(\partial\Omega)$, where the inclusion is achieved through zero continuation, by

$$P_1 f = f|_\Gamma + \frac{1}{|\Gamma|}\langle f, \mathbf{1}\rangle_{L^2(\partial\Omega\setminus\overline{\Gamma})}. \tag{2.3}$$

THEOREM 2.2. *Assume that $\sigma \in \mathbb{C}^{n\times n}$ satisfies* (2.2)*, and let u^0 be the solution of* (2.1) *corresponding to a given current pattern $f \in H_0^{-1/2}(\partial\Omega)$. Further, let u be the solution of problem* (2.1) *associated with the approximating input current $P_1 f \in H_0^{-1/2}(\partial\Omega)$. Then we have the estimate*

$$\left|\left|u^0 - u\right|\right|_{H^1(\Omega)/\mathbb{C}} \leq \frac{C}{|\Gamma|^{1/2}} \left|\left|f\right|\right|_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma})},$$

where $C > 0$ can be chosen independently of the geometry of Γ as a subset of $\partial\Omega$.

Proof. For $f \in H_0^{-1/2}(\partial\Omega)$ we have

$$\begin{aligned}
\left|\left|f - P_1 f\right|\right|_{H^{-1/2}(\partial\Omega)} &\leq \frac{1}{|\Gamma|}\left|\langle f, \mathbf{1}\rangle_{L^2(\partial\Omega\setminus\overline{\Gamma})}\right| \left|\left|\mathbf{1}\right|\right|_{\tilde{H}^{-1/2}(\Gamma)} + \left|\left|f\right|\right|_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma})} \\
&\leq \left(\frac{\left|\left|\mathbf{1}\right|\right|_{H^{1/2}(\partial\Omega\setminus\overline{\Gamma})} \left|\left|\mathbf{1}\right|\right|_{\tilde{H}^{-1/2}(\Gamma)}}{|\Gamma|} + 1\right) \left|\left|f\right|\right|_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma})} \\
&\leq \frac{C}{|\Gamma|^{1/2}} \left|\left|f\right|\right|_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma})},
\end{aligned}$$

where $C > 0$ can, clearly, be chosen independently of the geometry of Γ. Thus, the claim follows by applying Theorem 2.1 to the difference of the solutions $u^0 - u$. ∎

In a sense the result of Theorem 2.2 is quite natural: The discrepancy in the forward solution is bounded by the norm of the current that we were not able to use.

2.1.2. Complete electrode forward model. Next, we will introduce the CEM, which has been shown to model real-world electrode measurements reasonably well [9]. Assume that the boundary of the investigated object Ω is smooth and partially covered with electrodes $e_m \subset \partial\Omega$, $1 \le m \le M$, which are identified by the parts of the surface that they cover and assumed to be ideal conductors. The union of the electrode patches is denoted by $\Gamma_e = \cup_m e_m \subset \partial\Omega$. All electrodes are used for both current injection and voltage measurement, and the current and voltage patterns are denoted by $\{I_m\}, \{U_m\} \subset \mathbb{C}$, $1 \le m \le M$, respectively. To make the model even more flexible, we assume that on $\Gamma_n \subset \partial\Omega$, $\Gamma_n \cap \Gamma_e = \emptyset$, the current input is given in the continuous sense; i.e., on Γ_n the data belongs to $\tilde{H}^{-1/2}(\Gamma_n)$. Note that this kind of Neumann boundary is not usually included in the formulation of the CEM; here we introduce it to lighten our work load in section 3.

When conducting measurements with electrodes, a thin highly resistive layer is formed at the electrode-object interface [9]. It is characterized by the contact impedance $z : \partial\Omega \to \mathbb{C}$ that in our framework is assumed to be an integrable function satisfying

$$\mathrm{Re} z \ge z_0 > 0, \quad |z| \le z_1 < \infty, \tag{2.4}$$

almost everywhere on $\partial\Omega$. Note that the value of z between the electrodes indicates the fictitious value of the contact impedance, i.e., the value of the contact impedance if an electrode were present.

Traditionally, the electrode currents and potentials are handled as vectors of $\mathbb{C}^M$ [9]. However, encouraged by the fact that in CM the boundary potentials and currents are elements of L^2-based Sobolev spaces, in this work we interpret the electrode currents and potentials as elements of the subspace

$$T = \left\{ V \in L^2(\Gamma_e) \mid V = \sum_{m=1}^{M} \chi_{e_m} V_m,\ V_m \in \mathbb{C},\ 1 \le m \le M \right\} \subset L^2(\partial\Omega). \tag{2.5}$$

In what follows, we will also use the subspace

$$T_0 = \left\{ V \in T \,\middle|\, \int_{\partial\Omega} V dS = 0 \right\} \subset L^2_0(\partial\Omega), \tag{2.6}$$

to which the electrode currents belong if there is no Neumann boundary Γ_n.

With this convention the forward problem corresponding to the CEM is as follows. For input currents $I \in T$ and $g \in \tilde{H}^{-1/2}(\Gamma_n)$, with $I + g \in H_0^{-1/2}(\partial\Omega)$, find $(u^e, U^e) \in (H^1(\Omega) \oplus T)/\mathbb{C}$ that satisfies weakly

$$\begin{gathered} \nabla \cdot \sigma \nabla u^e = 0 \ \text{ in } \Omega, \quad \nu \cdot \sigma \nabla u^e = 0 \ \text{ on } \partial\Omega \setminus (\overline{\Gamma}_e \cup \overline{\Gamma}_n), \quad \nu \cdot \sigma \nabla u^e = g \ \text{ on } \Gamma_n, \\ u^e + z\nu \cdot \sigma \nabla u^e = U^e \ \text{ on } \Gamma_e, \quad \frac{1}{|e_m|} \int_{e_m} \nu \cdot \sigma \nabla u^e dS = I_m, \ \ 1 \le m \le M. \end{gathered} \tag{2.7}$$

Note that the above formulation of the complete electrode forward problem differs from the one in [9] by the scaling factor $1/|e_m|$ in the last equation of (2.7). However, the underlying physical interpretation stays the same: In [9] the net currents through electrodes were used; here we use the average currents. For more thorough physical justification of (2.7), the reader should consult [9].

THEOREM 2.3. *Assume that* (2.2) *and* (2.4) *hold and let* $I \in T$ *and* $g \in \tilde{H}^{-1/2}(\Gamma_n)$, *with* $I + g \in H_0^{-1/2}(\partial\Omega)$, *be given current patterns. Then problem* (2.7)

has a unique solution $(u^e, U^e) \in (H^1(\Omega) \oplus T)/\mathbb{C}$. *Further, this solution depends continuously on the data; i.e.,*

$$\inf_{c\in\mathbb{C}}\{||u^e - c||^2_{H^1(\Omega)} + ||U^e - c||^2_{L^2(\Gamma_e)}\}^{1/2} \leq C\{||I||_{L^2(\Gamma_e)} + ||g||_{\tilde{H}^{-1/2}(\Gamma_n)}\}, \tag{2.8}$$

where $C > 0$ *can be chosen independently of the geometry of* Γ_e *as a subset of* $\partial\Omega$.

By using material in [9], one could easily provide a proof for Theorem 2.3. However, since we have included the Neumann data on Γ_n in our model and, in addition, we are trying to build a connection between the complete electrode forward problem (2.7) and the continuum forward problem (2.1), we prefer a slightly different working order and postpone the proof until subsection 2.2.

2.2. Approximating with the CEM. In this subsection we aim to show that the CEM can be viewed as a real-world finite element approximation of the mathematically more tractable CM. To be more precise, with the help of the orthogonal projection $P_2 : L^2(\Gamma_e) \to T$ given by

$$P_2 f = \sum_{m=1}^{M} \chi_{e_m} \frac{1}{|e_m|} \int_{e_m} f dS, \quad f \in L^2(\Gamma_e), \tag{2.9}$$

we may write the main result of this subsection as follows in Theorem 2.4.

THEOREM 2.4. *Assume that* σ *and* z *satisfy* (2.2) *and* (2.4)*, respectively. Let* $f \in H_0^{-1/2}(\partial\Omega)$*, with* $f|_{\Gamma_e} \in L^2(\Gamma_e)$*, be a given input current and let* $u^0 \in H^1(\Omega)/\mathbb{C}$ *be the corresponding solution of* (2.1)*. Further, let* $(u^e, U^e) \in (H^1(\Omega) \oplus T)/\mathbb{C}$ *be the unique solution of* (2.7) *with the input currents* $P_2(P_1 f)|_{\Gamma_e} \in T$ *and* $(P_1 f)|_{\Gamma_n} \in \tilde{H}^{-1/2}(\Gamma_n)$*, where* P_2 *is given by* (2.9) *and* P_1 *by* (2.3) *with* $\Gamma = \Gamma_e \cup \Gamma_n$*. Then it holds that*

$$\left|\left|u^0 - u^e\right|\right|_{H^1(\Omega)/\mathbb{C}} \leq C\left\{\frac{1}{|\Gamma|^{1/2}}\,||f||_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma})} + \inf_{V\in T}\left|\left|U^0 - V\right|\right|_{L^2(\Gamma_e)/\mathbb{C}}\right\},$$

where $C > 0$ *can be chosen independently of the geometry of* Γ_e *as a subset of* $\partial\Omega$*, the subspace* $T \subset L^2(\Gamma_e)$ *is given in* (2.5)*, and* $U^0 = u^0|_{\Gamma_e} + zf|_{\Gamma_e}$.

Theorem 2.4 tells us, roughly speaking, that for a given current pattern the best correspondence between the solutions of the forward problems (2.1) and (2.7) is obtained when the electrodes are as small as possible and the gaps between the adjacent electrodes are as narrow as possible.

In order to prove Theorem 2.4, we need, first of all, a suitable variational problem: For $f \in L^2(\Gamma_e)$ and $g \in \tilde{H}^{-1/2}(\Gamma_n)$, with $f + g \in H_0^{-1/2}(\partial\Omega)$, find $(u, U) \in H = (H^1(\Omega) \oplus L^2(\Gamma_e))/\mathbb{C}$ such that

$$B((u, U), (v, V)) = F(v, V) \quad \text{for all } (v, V) \in H, \tag{2.10}$$

where

$$B((u, U), (v, V)) = \int_\Omega \sigma\nabla u \cdot \nabla\overline{v} dx + \int_{\Gamma_e} \frac{1}{z}(U - u)(\overline{V} - \overline{v}) dS, \tag{2.11}$$

and

$$F(v, V) = \int_{\Gamma_e} f\overline{V} dS + \int_{\Gamma_n} g\overline{v} dS, \tag{2.12}$$

where the latter term is to be interpreted in the sense of the dual pairing between $\tilde{H}^{-1/2}(\Gamma_n)$ and $H^{1/2}(\Gamma_n)$. To keep the motivation high, note that variational equation (2.10) is quite similar to the variational formulation of the complete electrode forward problem in [9], the only clear difference being the space from which the solution is sought. We claim that (2.10) has a unique solution with some interesting properties.

Before we can prove the unique solvability of (2.10), we still need to introduce an inner product on $H = (H^1(\Omega) \oplus L^2(\Gamma_e)/\mathbb{C}$, namely,

$$((u,U),(v,V))_{*H} = \int_\Omega \nabla u \cdot \nabla \overline{v} dx + \int_{\Gamma_e} (U-u)(\overline{V}-\overline{v}) dS, \tag{2.13}$$

with the corresponding norm

$$||(v,V)||^2_{*H} = ((v,V),(v,V))_{*H}. \tag{2.14}$$

The following lemma tells us that the above inner product and norm are well defined and concordant with the conventional quotient norm of H given by

$$||(v,V)||_H = \inf_{c\in\mathbb{C}} \{||v-c||^2_{H^1(\Omega)} + ||V-c||^2_{L^2(\Gamma_e)}\}^{1/2}.$$

LEMMA 2.5. *The sesquilinear map* $(\cdot,\cdot)_{*H} : H \times H \to \mathbb{C}$ *given by* (2.13) *defines an inner product which is concordant with the quotient topology of* $H = (H^1(\Omega) \oplus L^2(\Gamma_e))/\mathbb{C}$. *In consequence,* H *is a Hilbert space.*

Proof. Clearly, $(\cdot,\cdot)_{*H} : H \times H \to \mathbb{C}$ is well defined and satisfies all the inner product axioms. Hence, the only thing we need to show, in order to prove the claim, is that the usual quotient norm $||(\cdot,\cdot)||_H$ and the norm $||(\cdot,\cdot)||_{*H}$ defined in (2.14) are equivalent.

Let $(v,V) \in H$ be arbitrary. With the help of the trace theorem [5], we may estimate

$$\begin{aligned} ||(v,V)||_{*H} &\le ||\nabla v||_{L^2(\Omega)} + ||V-v||_{L^2(\Gamma_e)} \\ &\le ||v-c||_{H^1(\Omega)} + ||v-c||_{L^2(\Gamma_e)} + ||V-c||_{L^2(\Gamma_e)} \\ &\le C \left\{ ||v-c||^2_{H^1(\Omega)} + ||V-c||^2_{L^2(\Gamma_e)} \right\}^{1/2}. \end{aligned}$$

Since this holds for every $c \in \mathbb{C}$, we actually have

$$||(v,V)||_{*H} \le C\, ||(v,V)||_H\,. \tag{2.15}$$

On the other hand, by using the trace theorem and Poincaré's inequality [5], we get

$$\begin{aligned} ||(v,V)||^2_H &\le \inf_{c\in\mathbb{C}} \left\{ ||v-c||^2_{H^1(\Omega)} + 2\,||v-c||^2_{L^2(\Gamma_e)} \right\} + 2\,||V-v||^2_{L^2(\Gamma_e)} \\ &\le C \inf_{c\in\mathbb{C}} ||v-c||^2_{H^1(\Omega)} + 2\,||V-v||^2_{L^2(\Gamma_e)} \\ &\le C\, ||(v,V)||^2_{*H}\,. \end{aligned}$$

Combining this with (2.15) completes the proof. □

COROLLARY 2.6. *Assume that* (2.2) *and* (2.4) *hold. Then the sesquilinear form* $B : H \times H \to \mathbb{C}$ *given in* (2.11) *is continuous as follows:*

$$|B((u,U),(v,V))| \le C\, ||(u,U)||_H\, ||(v,V)||_H\,, \quad (u,U),(v,V) \in H,$$

and coercive as follows:

$$|B((v,V),(v,V))| \geq c\,||(v,V)||_H^2\,, \quad (v,V) \in H,$$

where the positive constants c and C can be chosen independently of the geometry of Γ_e as a subset of $\partial\Omega$.

Proof. The claim is a straightforward consequence of Lemma 2.5 together with inequalities (2.2) and (2.4). ∎

Next we aim at showing that the functional F on the right-hand side of (2.10) is continuous. To begin with, we extend the trace theorem for the quotient Sobolev spaces as follows in Lemma 2.7.

LEMMA 2.7. *The quotient trace map*

$$Tr : H^1(\Omega)/\mathbb{C} \to H^{1/2}(\partial\Omega)/\mathbb{C}, \quad v \mapsto v|_{\partial\Omega},$$

is bounded.

Proof. The claim is a straightforward consequence of the traditional trace theorem and Poincaré's inequality. ∎

LEMMA 2.8. *Let $f \in L^2(\Gamma_e)$ and $g \in \tilde{H}^{-1/2}(\Gamma_n)$ with $f+g \in H_0^{-1/2}(\partial\Omega)$. Then the linear functional $F : H \to \mathbb{C}$ given in* (2.12) *is well defined and continuous.*

Proof. Let us first show that $F : H \to \mathbb{C}$ is well defined. Consider two representatives (v,V) and $(v+c, V+c)$ of the same equivalence class in H. Since $f+g \in H_0^{-1/2}(\partial\Omega)$, we have

$$F(v+c,V+c) = \int_{\Gamma_e} f\overline{V}\,dS + \int_{\Gamma_n} g\overline{v}\,dS + \overline{c}\langle f+g, \mathbf{1}\rangle_{L^2(\partial\Omega)} = F(v,V). \tag{2.16}$$

Further, by the use of (2.16) and Lemma 2.7, we may estimate, for an arbitrary $(v,V) \in H$,

$$\begin{aligned} |F(v,V)| &= \inf_{c\in\mathbb{C}} \left| \int_{\Gamma_e} f(\overline{V}+\overline{c})dS + \int_{\Gamma_n} g(\overline{v}+\overline{c})dS \right| \\ &\leq \inf_{c\in\mathbb{C}} \{ ||f||_{L^2(\Gamma_e)}\, ||V+c||_{L^2(\Gamma_e)} + ||g||_{\tilde{H}^{-1/2}(\Gamma_n)}\, ||v+c||_{H^{1/2}(\Gamma_n)} \} \\ &\leq C\{ ||f||_{L^2(\Gamma_e)} + ||g||_{\tilde{H}^{-1/2}(\Gamma_n)} \}\, ||(v,V)||_H\,, \end{aligned} \tag{2.17}$$

where $C > 0$ can be chosen independently of the geometry of Γ_e as a subset of $\partial\Omega$. This completes the proof. ∎

Now we have introduced enough weaponry to consider the solvability of (2.10).

LEMMA 2.9. *Assume that* (2.2) *and* (2.4) *hold and let $f \in L^2(\Gamma_e)$ and $g \in \tilde{H}^{-1/2}(\Gamma_n)$, with $f+g \in H_0^{-1/2}(\partial\Omega)$, be given current patterns. Then variational equation* (2.10) *has a unique solution $(u,U) \in H = (H^1(\Omega) \oplus L^2(\Gamma_e))/\mathbb{C}$. Further, the first component of this solution, $u \in H^1(\Omega)/\mathbb{C}$, is the unique solution of the continuum forward problem* (2.1) *with the input current $f+g$, and the second component satisfies $U = u|_{\Gamma_e} + zf$.*

Proof. The existence of a unique solution for (2.10) is a straight consequence of the Lax–Milgram lemma [13], Corollary 2.6, and Lemma 2.8.

Let $u^0 \in H^1(\Omega)/\mathbb{C}$ be the unique solution of (2.1) corresponding to the input current $f+g \in H_0^{-1/2}(\partial\Omega)$ and define $(u^0, U^0) = (u^0, u^0|_{\Gamma_e} + zf) \in H$. For an

arbitrary $(v, V) \in H$, it holds that

$$\begin{aligned} B((u^0, U^0), (v, V)) &= \int_{\partial\Omega} \nu \cdot \sigma \nabla u^0 \overline{v} dS + \int_{\Gamma_e} f(\overline{V} - \overline{v}) dS \\ &= \int_{\Gamma_e} (\nu \cdot \sigma \nabla u^0 - f)\overline{v} dS + \int_{\Gamma_e} f\overline{V} dS + \int_{\Gamma_n} \nu \cdot \sigma \nabla u^0 \overline{v} dS \\ &= \int_{\Gamma_e} f\overline{V} dS + \int_{\Gamma_n} g\overline{v} dS = F(v, V), \end{aligned}$$

where we used Green's formula. Thus, $(u^0, U^0) \in H$ is a solution to (2.10), which completes the proof. ∎

Now, it is time to return to the complete electrode forward problem. Note that $H' = (H^1(\Omega) \oplus T)/\mathbb{C}$, where T is defined by (2.5), is a subspace of $H = (H^1(\Omega) \oplus L^2(\Gamma_e))/\mathbb{C}$. Thus, the variational problem: For $f \in L^2(\Gamma_e)$ and $g \in \tilde{H}^{-1/2}(\Gamma_n)$, with $f + g \in H_0^{-1/2}(\partial\Omega)$, find $(u, U) \in H'$ so that

$$B((u, U), (v, V)) = F(v, V) \quad \text{for all } (v, V) \in H', \tag{2.18}$$

where B and F are defined in (2.11) and (2.12), respectively, can be considered a Galerkin approximation for variational problem (2.10). We claim that the unique solution for this approximating variational problem is in fact the unique solution for the complete electrode forward problem (2.7) with a suitable electrode current.

LEMMA 2.10. *Assume that* (2.2) *and* (2.4) *hold and let* $f \in L^2(\Gamma_e)$ *and* $g \in \tilde{H}^{-1/2}(\Gamma_n)$, *with* $f + g \in H_0^{-1/2}(\partial\Omega)$, *be given current patterns. Then variational equation* (2.18) *has a unique solution* $(u^e, U^e) \in H' = (H^1(\Omega) \oplus T)/\mathbb{C}$ *which is also the unique solution of the complete electrode forward problem* (2.7) *corresponding to the input currents* $P_2 f \in T$, *where* P_2 *is given by* (2.9), *and* $g \in \tilde{H}^{-1/2}(\Gamma_n)$.

Proof. By Corollary 2.6 and Lemma 2.8, the sesquilinear form $B : H \times H \to \mathbb{C}$ is continuous and coercive, and the linear functional $F : H \to \mathbb{C}$ is continuous. In consequence, the restrictions $B : H' \times H' \to \mathbb{C}$ and $F : H' \to \mathbb{C}$ have these same properties. Further, since H' is a closed subspace of the Hilbert space H, it is also a Hilbert space, and so the unique existence of a solution to (2.18) follows from the Lax–Milgram lemma [13].

To prove that variational problem (2.18) is equivalent to the complete electrode forward problem (2.7) with the electrode current $P_2 f \in T$, we write the left-hand side of (2.18) componentwise; i.e., for $(u, U), (v, V) \in H'$ we have

$$B((u, U), (v, V)) = \int_\Omega \sigma \nabla u \cdot \nabla \overline{v} dS + \sum_{m=1}^{M} \int_{e_m} \frac{1}{z}(U_m - u)(\overline{V}_m - \overline{v}) dS.$$

With the same tactic, the right-hand side of (2.18) can be transformed into

$$F(v, V) = \sum_{m=1}^{M} \int_{e_m} f\overline{V}_m dS + \int_{\Gamma_n} g\overline{v} dS = \sum_{m=1}^{M} |e_m|(P_2 f)_m \overline{V}_m + \int_{\Gamma_n} g\overline{v} dS.$$

With this convention, the claimed equivalence between problems (2.7) and (2.18) follows by the same line of reasoning as in the proof of Proposition 3.1 in [9], with only slight alterations caused by the excess Neumann term on the right-hand side of (2.18). ∎

Now we have also the means to prove Theorem 2.3.

Proof of Theorem 2.3. For the given current patterns $I \in T \subset L^2(\Gamma_e)$ and $g \in \tilde{H}^{-1/2}(\Gamma_n)$, with $I + g \in H_0^{-1/2}(\partial\Omega)$, the unique existence of the solution $(u^e, U^e) \in H' = (H^1(\Omega) \oplus T)/\mathbb{C}$ for (2.7) follows from the equivalence between problems (2.18) and (2.7), considered in the proof of Lemma 2.10, by choosing $f = I$ in (2.18) and noting that $P_2 I = I$. Further, by using Corollary 2.6 and equation (2.17), we may estimate

$$\begin{aligned} ||(u^e, U^e)||_H^2 &\leq C|B((u^e, U^e), (u^e, U^e))| \\ &= C|F(u^e, U^e)| \\ &\leq C\{||I||_{L^2(\Gamma_e)} + ||g||_{\tilde{H}^{-1/2}(\Gamma_n)}\} \, ||(u^e, U^e)||_H \,, \end{aligned}$$

where the functional F, defined in (2.12), corresponds to currents $f = I$ and g. This completes the proof. □

There are a few things worth noticing. First, the solution $(u^e, U^e) \in H'$ of (2.7) satisfies

$$\nu \cdot \sigma \nabla u^e|_{\Gamma_e} = \frac{1}{z}(U^e - u^e|_{\Gamma_e}).$$

In particular, $\nu \cdot \sigma \nabla u^e|_{\Gamma_e} \in L^2(\Gamma_e)$. Second, the correspondence between problems (2.7) and (2.18) gives the complete electrode forward problem a variational formulation: For $I \in T$ and $g \in \tilde{H}^{-1/2}(\Gamma_n)$, with $I + g \in H_0^{-1/2}(\partial\Omega)$, find $(u^e, U^e) \in H'$ so that

$$B((u^e, U^e), (v, V)) = \int_{\Gamma_e} I\overline{V} dS + \int_{\Gamma_n} g\bar{v} dS \tag{2.19}$$

for all $(v, V) \in H'$.

Now we have derived the means to approximate the forward solution of the CM (2.1) by the forward solution of the complete electrode problem (2.7) with a correctly chosen electrode current pattern. First we will consider the case when no current is conducted through $\partial\Omega \setminus (\overline{\Gamma}_e \cup \overline{\Gamma}_n)$.

THEOREM 2.11. *Assume that σ and z satisfy* (2.2) *and* (2.4)*, respectively. Let $f \in H_0^{-1/2}(\partial\Omega)$, with $f|_{\Gamma_e} \in L^2(\Gamma_e)$ and $f|_{\partial\Omega\setminus(\overline{\Gamma}_e \cup \overline{\Gamma}_n)} = 0$, be a given input current and let $u \in H^1(\Omega)/\mathbb{C}$ be the corresponding solution of* (2.1)*. Further, let $(u^e, U^e) \in H' = (H^1(\Omega) \oplus T)/\mathbb{C}$ be the unique solution of* (2.7) *with the input currents $P_2(f|_{\Gamma_e}) \in T$ and $f|_{\Gamma_n} \in \tilde{H}^{-1/2}(\Gamma_n)$, where P_2 is given in* (2.9)*. Then it holds that*

$$||(u - u^e, U - U^e)||_H \leq C \inf_{V \in T} ||U - V||_{L^2(\Gamma_e)/\mathbb{C}} \,,$$

where $C > 0$ can be chosen independently of the geometry of Γ_e as a subset of $\partial\Omega$, the subspace $T \subset L^2(\Gamma_e)$ is given in (2.5)*, and $U = u|_{\Gamma_e} + zf|_{\Gamma_e}$.*

Proof. To begin with, note that, according to Lemma 2.9, the pair $(u, U) \in H = (H^1(\Omega) \oplus L^2(\Gamma_e))/\mathbb{C}$ satisfies the variational equation

$$B((u, U), (v, V)) = \int_{\Gamma_e} f\overline{V} dS + \int_{\Gamma_n} f\bar{v} dS \quad \text{for all } (v, V) \in H,$$

and, on the other hand, Lemma 2.10 tells us that (u^e, U^e) satisfies the very same equation with the space H replaced by the subspace H'. Since the sesquilinear form $B : H \times H \to \mathbb{C}$ is continuous and coercive, it follows from Cea's lemma [2] that

$$||(u - u^e, U - U^e)||_H \leq C \inf_{(v,V) \in H'} ||(u - v, U - V)||_H \,. \tag{2.20}$$

Choosing $v = u$ on the right-hand side of (2.20), we obtain

$$\begin{aligned}\inf_{(v,V)\in H'} ||(u-v, U-V)||_H &\leq \inf_{V\in T}\inf_{c\in\mathbb{C}} \left\{ ||c||^2_{H^1(\Omega)} + ||(U-V)-c||^2_{L^2(\Gamma_e)} \right\}^{1/2} \\ &= \inf_{V\in T} ||U-V||_{L^2(\Gamma_e)/\mathbb{C}}\,. \end{aligned} \tag{2.21}$$

Hence, combining (2.20) and (2.21), the claim follows. □

The following corollary also tells us that the normal derivative of the solution to (2.1) can be approximated with the normal derivative of the solution to (2.7).

COROLLARY 2.12. *Suppose that the assumptions of Theorem* 2.11 *are valid. Then, using the same notation as in Theorem* 2.11*, we have the estimate*

$$||f - \nu\cdot\sigma\nabla u^e||_{L^2(\Gamma_e)} = ||\nu\cdot\sigma\nabla u - \nu\cdot\sigma\nabla u^e||_{L^2(\Gamma_e)} \leq C \inf_{V\in T} ||U-V||_{L^2(\Gamma_e)/\mathbb{C}}\,.$$

Proof. Due to the boundary conditions of (2.7) and the way we have defined U in Theorem 2.11, we may estimate

$$\begin{aligned} ||\nu\cdot\sigma(\nabla u - \nabla u^e)||_{L^2(\Gamma_e)} &\leq C\,||z\nu\cdot\sigma(\nabla u - \nabla u^e)||_{L^2(\Gamma_e)} \\ &= C\,||(U-U^e)-(u-u^e)||_{L^2(\Gamma_e)} \\ &\leq C \inf_{c\in\mathbb{C}}\{||(U-U^e)-c||_{L^2(\Gamma_e)} + ||c-(u-u^e)||_{L^2(\Gamma_e)}\} \\ &\leq C\,||(u-u^e, U-U^e)||_H\,, \end{aligned}$$

where we took advantage of the trace theorem [5]. The claim follows by combining this with Theorem 2.11. □

Finally, it is time to provide a proof for Theorem 2.4 by combining Theorem 2.11 with Theorem 2.2.

Proof of Theorem 2.4. Let $u \in H^1(\Omega)/\mathbb{C}$ be the solution to the continuum forward problem (2.1) corresponding to the input current $P_1 f \in \tilde{H}_0^{-1/2}(\Gamma)$, where P_1 is defined by (2.3), and define $U = (u + zP_1 f)|_{\Gamma_e}$. According to Theorems 2.2 and 2.11, we can estimate

$$\begin{aligned} \big|\big|u^0 - u^e\big|\big|_{H^1(\Omega)/\mathbb{C}} &\leq \big|\big|u^0 - u\big|\big|_{H^1(\Omega)/\mathbb{C}} + ||u-u^e||_{H^1(\Omega)/\mathbb{C}} \\ &\leq C\left\{ \frac{1}{|\Gamma|^{1/2}}\,||f||_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma})} + \inf_{V\in T} ||U-V||_{L^2(\Gamma_e)/\mathbb{C}} \right\}, \end{aligned} \tag{2.22}$$

where the latter term may be divided into two parts by using the triangle inequality

$$\inf_{V\in T} ||U-V||_{L^2(\Gamma_e)/\mathbb{C}} \leq \inf_{V\in T} \big|\big|U^0 - V\big|\big|_{L^2(\Gamma_e)/\mathbb{C}} + \big|\big|U^0 - U\big|\big|_{L^2(\Gamma_e)/\mathbb{C}}\,. \tag{2.23}$$

Further, by using (2.3), (2.4), Lemma 2.7, Theorem 2.2, and the way U^0 and U are defined, we deduce that

$$\begin{aligned} \big|\big|U^0 - U\big|\big|_{L^2(\Gamma_e)/\mathbb{C}} &\leq \big|\big|u^0 - u\big|\big|_{L^2(\Gamma_e)/\mathbb{C}} + ||z(f-P_1 f)||_{L^2(\Gamma_e)/\mathbb{C}} \\ &\leq C\,\big|\big|u^0 - u\big|\big|_{H^1(\Omega)/\mathbb{C}} + \frac{1}{|\Gamma|}\left|\langle f, \mathbf{1}\rangle_{L^2(\partial\Omega\setminus\overline{\Gamma})}\right| ||z||_{L^2(\Gamma_e)/\mathbb{C}} \\ &\leq \left\{ \frac{C}{|\Gamma|^{1/2}} + \frac{z_1}{|\Gamma|}\,||\mathbf{1}||_{H^{1/2}(\partial\Omega\setminus\overline{\Gamma})}\,||\mathbf{1}||_{L^2(\Gamma_e)} \right\} ||f||_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma})} \\ &\leq \frac{C}{|\Gamma|^{1/2}}\,||f||_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma})}\,. \end{aligned} \tag{2.24}$$

The claim follows by combining (2.22), (2.23), and (2.24). □

2.3. Comparing current-to-voltage maps. It is time to move on to consider the current-to-potential maps corresponding to CM and CEM. In order to keep things simple, in this subsection we assume that there is no Neumann boundary; the introduction of Γ_n in (2.7) was just a technical detail that is useful for us in the next section—usually, it does not play any part in real-world measurements. As a further simplification, we assume that all used current patterns are square integrable.

When dealing with the inverse problem for the CM, it is usually assumed that the known data is the linear Neumann-to-Dirichlet map, i.e., the operator that maps the applied current pattern onto the boundary potential

$$\Lambda_\sigma : f \mapsto u_\sigma^0|_{\partial\Omega}, \tag{2.25}$$

which is isomorphic from $H_0^{-1/2}(\partial\Omega)$ onto $H^{1/2}(\partial\Omega)/\mathbb{C} \sim H_0^{1/2}(\partial\Omega)$ and depends nonlinearly on σ. On the other hand, when conducting real-life measurements with the CEM, the only information one is able to obtain is the linear relation between the applied average currents $I_m \in \mathbb{C}$, $1 \le m \le M$, and the electrode voltages $U_m^e \in \mathbb{C}$, $1 \le m \le M$, given by

$$R_\sigma I = U^e,$$

where $R_\sigma : T_0 \to T/\mathbb{C}$ can be expressed in matrix form since $T_0, T/\mathbb{C} \sim \mathbb{C}^{M-1}$. The next challenge is to build some kind of approximating link between the operators Λ_σ and R_σ.

Assume that there is no Neumann boundary; i.e., $\Gamma_n = \emptyset$ in (2.7). By combining R_σ with the projection

$$P = P_2 P_1 : L_0^2(\partial\Omega) \to T_0, \tag{2.26}$$

where T_0 is given by (2.6), and the projections P_1, with $\Gamma = \Gamma_e$, and P_2 are defined in (2.3) and (2.9), respectively, we get the map

$$R_\sigma P : L_0^2(\partial\Omega) \to T/\mathbb{C}, \quad f \mapsto U_\sigma^e = (u_\sigma^e + z\nu \cdot \sigma \nabla u_\sigma^e)|_{\Gamma_e},$$

where $(u_\sigma^e, U_\sigma^e) \in (H^1(\Omega) \oplus T)/\mathbb{C}$ is the solution of (2.7) corresponding to the electrode current Pf and the admittance σ.

The resemblance between the operators Λ_σ and $R_\sigma P$ is quite apparent. However, $R_\sigma P$ is not a pure current-to-voltage map, which prevents us from using Theorem 2.4 to investigate the situation further. Luckily, in many of the reconstruction algorithms for the CM, one does not use merely Λ_σ but the difference [1]

$$\Lambda_\sigma - \Lambda_{\mathbf{1}} : f \mapsto (u_\sigma^0 - u_{\mathbf{1}}^0)|_{\partial\Omega}, \tag{2.27}$$

where $\Lambda_{\mathbf{1}}$ is the Neumann-to-Dirichlet map corresponding to the unit admittance distribution, and $u_{\mathbf{1}}^0 \in H^1(\Omega)/\mathbb{C}$ is the associated forward solution for the input current $f \in L_0^2(\partial\Omega)$. For the complete electrode counterpart, we get the formula

$$(R_\sigma - R_{\mathbf{1}})P : f \mapsto (u_\sigma^e - u_{\mathbf{1}}^e + z\nu \cdot (\sigma \nabla u_\sigma^e - \nabla u_{\mathbf{1}}^e))|_{\Gamma_e}, \tag{2.28}$$

which is, actually, quite close to (2.27).

THEOREM 2.13. *Assume that σ and z satisfy* (2.2) *and* (2.4)*, respectively, and let $f \in L_0^2(\partial\Omega)$ be a given current pattern. It holds that*

$$||((\Lambda_\sigma - \Lambda_{\mathbf{1}}) - (R_\sigma - R_{\mathbf{1}})P)f||_{L^2(\Gamma_e)/\mathbb{C}} \le \tag{2.29}$$

$$C\left\{ \frac{1}{|\Gamma_e|^{1/2}} ||f||_{\tilde{H}^{-1/2}(\partial\Omega \setminus \overline{\Gamma}_e)} + \inf_{V \in T} ||U_\sigma^0 - V||_{L^2(\Gamma_e)/\mathbb{C}} + \inf_{V \in T} ||U_{\mathbf{1}}^0 - V||_{L^2(\Gamma_e)/\mathbb{C}} \right\},$$

where the boundary operators are defined by (2.27) *and* (2.28)*, and* $U_\sigma^0 = u_\sigma^0|_{\Gamma_e} + zf|_{\Gamma_e}$, $U_{\mathbf{1}}^0 = u_{\mathbf{1}}^0|_{\Gamma_e} + zf|_{\Gamma_e}$*, where* u_σ^0 *and* $u_{\mathbf{1}}^0$ *are the solutions of the continuum forward problem* (2.1) *corresponding to the input current* f *and the impedance tensors* σ *and* $\mathbf{1}$*, respectively.*

Proof. With the help of (2.27) and (2.28), the left-hand side of (2.29) can be divided into three parts as follows:

$$\begin{aligned}||((\Lambda_\sigma - \Lambda_{\mathbf{1}}) - (R_\sigma - R_{\mathbf{1}})P)f||_{L^2(\Gamma_e)/\mathbb{C}} \le \left|\left|u_\sigma^0 - u_\sigma^e\right|\right|_{L^2(\Gamma_e)/\mathbb{C}} + \left|\left|u_{\mathbf{1}}^0 - u_{\mathbf{1}}^e\right|\right|_{L^2(\Gamma_e)/\mathbb{C}} \\ + ||z\nu \cdot (\sigma\nabla u_\sigma^e - \nabla u_{\mathbf{1}}^e)||_{L^2(\Gamma_e)/\mathbb{C}}\,. \end{aligned} \tag{2.30}$$

For the first term on the right-hand side of (2.30), it follows from Lemma 2.7 and Theorem 2.4 that

$$\left|\left|u_\sigma^0 - u_\sigma^e\right|\right|_{L^2(\Gamma_e)/\mathbb{C}} \le C\left\{\frac{1}{|\Gamma_e|^{1/2}}\,||f||_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma}_e)} + \inf_{V\in T}\left|\left|U_\sigma^0 - V\right|\right|_{L^2(\Gamma_e)/\mathbb{C}}\right\}.$$

By the same means, we get an exactly similar estimate for the second term on the right-hand side of (2.30).

In order to handle the third term on the right-hand side of (2.30), let $u_\sigma, u_{\mathbf{1}} \in H^1(\Omega)/\mathbb{C}$ be the solutions of (2.1) with the input current $P_1 f \in L_0^2(\Gamma_e)$, where $P_1 : L_0^2(\partial\Omega) \to L_0^2(\Gamma_e)$ is defined by (2.3) with $\Gamma = \Gamma_e$, for the admittances σ and $\mathbf{1}$, respectively, and define $U_\sigma = u_\sigma|_{\Gamma_e} + zP_1 f$ and $U_{\mathbf{1}} = u_{\mathbf{1}}|_{\Gamma_e} + zP_1 f$. We use Corollary 2.12 to estimate

$$\begin{aligned}||z\nu \cdot (\sigma\nabla u_\sigma^e - \nabla u_{\mathbf{1}}^e)||_{L^2(\Gamma_e)/\mathbb{C}} &\le ||z(\nu\cdot\sigma\nabla u_\sigma^e - P_1 f)||_{L^2(\Gamma_e)} + ||z(P_1 f - \nu\cdot\nabla u_{\mathbf{1}}^e)||_{L^2(\Gamma_e)} \\ &\le C\left\{\inf_{V\in T}||U_\sigma - V||_{L^2(\Gamma_e)/\mathbb{C}} + \inf_{V\in T}||U_{\mathbf{1}} - V||_{L^2(\Gamma_e)/\mathbb{C}}\right\} \\ &\le C\left\{\frac{1}{|\Gamma_e|^{1/2}}\,||f||_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma}_e)} + \inf_{V\in T}\left|\left|U_\sigma^0 - V\right|\right|_{L^2(\Gamma_e)/\mathbb{C}}\right. \\ &\quad \left. + \inf_{V\in T}\left|\left|U_{\mathbf{1}}^0 - V\right|\right|_{L^2(\Gamma_e)/\mathbb{C}}\right\},\end{aligned}$$

where we also use assumption (2.4) and inequality (2.24) from the proof of Theorem 2.4. The claim follows by combining the estimates for the terms on the right-hand side of (2.30). ∎

Again it is advisable to note a couple of things. First, the above theorem could also have been formulated for currents $f \in H_0^{-1/2}(\partial\Omega)$ with $f|_{\Gamma_e} \in L^2(\Gamma_e)$; the notation would have been even more cumbersome, however. Second, the images of the boundary maps are compared only on $\Gamma_e \subset \partial\Omega$ since in a real-life measurement situation one is not measuring anything outside the electrodes and, thus, there is nothing to compare on $\partial\Omega\setminus\overline{\Gamma}_e$. Third, the correspondence between the maps $\Lambda_\sigma - \Lambda_{\mathbf{1}}$ and $R_\sigma - R_{\mathbf{1}}$ gets better when the area covered by the electrodes gets larger and the electrodes get smaller.

3. Characterizing inclusions. In this section we demonstrate how the boundary map $R_\sigma - R_{\mathbf{1}}$, considered in the previous subsection, can be used to characterize an inclusion $D \subset \Omega$ with conductivity significantly higher or lower than the constant background conductivity. The section is organized as follows. We begin by introducing our framework and listing some basic properties of R_σ. Section 3.1 presents a factorization of $R_\sigma - R_{\mathbf{1}}$ into three parts. In section 3.2 the operators needed in the

factorization are investigated further and, finally, in section 3.3 we provide the characterization for the inclusion. This work can be seen as a discrete version of article [3]; also, the mathematical methods used here resemble to some extent those in [3].

Our object $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, is assumed to be isotropic with a conductivity $0 < \sigma \leq C$ and a smooth boundary, and the used input current is assumed to be nonvariating in time. Further, we assume that the contact impedance $z : \partial\Omega \to \mathbb{R}$ is strictly positive and bounded. Let the subspaces $T \subset L^2(\partial\Omega)$ and $T_0 \subset L^2_0(\partial\Omega)$ be defined by (2.5) and (2.6), respectively, with the multiplier field $\mathbb{C}$ replaced by $\mathbb{R}$. Then, according to Theorem 2.3, for every $I \in T_0$ the complete electrode forward problem

$$\begin{aligned} &\nabla \cdot \sigma \nabla u = 0 \ \text{ in } \Omega, \quad \nu \cdot \sigma \nabla u = 0 \ \text{ on } \partial\Omega \setminus \overline{\Gamma}_e, \\ &u + z\nu \cdot \sigma \nabla u = U \ \text{ on } \Gamma_e, \quad \frac{1}{|e_m|} \int_{e_m} \nu \cdot \sigma \nabla u dS = I_m, \ \ 1 \leq m \leq M, \end{aligned} \tag{3.1}$$

has a unique solution $(u, U) \in H^1(\Omega) \oplus T_0$, where we have specified the ground level of the potential in an obvious way. The corresponding boundary map $R_\sigma : T_0 \to T_0$ is defined through $R_\sigma I = U$.

We emphasize the resemblance between R_σ and its continuous counterpart Λ_σ, given in (2.25), by showing that R_σ inherits some basic characteristics of Λ_σ.

LEMMA 3.1. *The operator $R_\sigma : T_0 \to T_0$ is self-adjoint and positive. Furthermore, R_σ is monotonically decreasing; i.e.,*

$$\langle I, R_\sigma I \rangle_{L^2(\partial\Omega)} > \langle I, R_{\tilde{\sigma}} I \rangle_{L^2(\partial\Omega)},$$

for $\sigma \leq \tilde{\sigma}$, $\sigma \neq \tilde{\sigma}$ on a set of nonzero measure, and $I \neq 0$.

Proof. The result follows by imitating the proof of Lemma 2.1 in [3] with the help of the weak formulation of (3.1) given by (2.19). □

Since T_0 is a finite-dimensional subspace of $L^2_0(\partial\Omega)$, the monotonicity property of Lemma 3.1 implies that $R_\sigma - R_{\tilde{\sigma}} : T_0 \to T_0$ has a bounded inverse if the assumptions of Lemma 3.1 are valid.

COROLLARY 3.2. *Let $\sigma \leq \tilde{\sigma}$, and $\sigma \neq \tilde{\sigma}$ on a set of nonzero measure. Then $R_\sigma - R_{\tilde{\sigma}} : T_0 \to T_0$ is strictly positive. In particular, $R_\sigma - R_{\tilde{\sigma}}$ is bijective and has a bounded inverse.*

Proof. From the monotonicity property of Lemma 3.1 we straight away obtain

$$\langle I, (R_\sigma - R_{\tilde{\sigma}}) I \rangle_{L^2(\partial\Omega)} > 0,$$

for every $I \in T_0$, $I \neq 0$. Since T_0 is finite-dimensional and $R_\sigma - R_{\tilde{\sigma}}$ is linear, this induces the estimate

$$\langle I, (R_\sigma - R_{\tilde{\sigma}}) I \rangle_{L^2(\partial\Omega)} \geq c \, ||I||^2_{L^2(\partial\Omega)}, \quad c > 0. \tag{3.2}$$

The injectivity, or, equivalently, the bijectivity, of $R_\sigma - R_{\tilde{\sigma}} : T_0 \to T_0$ follows trivially from (3.2), which completes the proof. □

3.1. Factorization of $R_\sigma - R_1$. From now on we assume that the conductivity inside Ω is of the form

$$\sigma = \begin{cases} \kappa & \text{in } D, \\ 1 & \text{in } \Omega \setminus \overline{D}, \end{cases} \tag{3.3}$$

where $\kappa \neq 1$ is a positive constant and D is an open connected subset of Ω with a smooth connected boundary and $\partial D \cap \partial\Omega = \emptyset$. Our aim is to prove the following theorem.

THEOREM 3.3. *Assume that the conductivity inside Ω is of the form given in (3.3). Then the difference of the boundary maps $R_\sigma, R_{\mathbf{1}} : T_0 \to T_0$ can be factorized as $R_\sigma - R_{\mathbf{1}} = LFL'$, where $L : H_0^{-1/2}(\partial D) \to T_0$ is continuous and surjective, its adjoint operator $L' : T_0 \to H_0^{1/2}(\partial D)$ is continuous and injective, and $F : H_0^{1/2}(\partial D) \to H_0^{-1/2}(\partial D)$ is self-adjoint, bijective, and either positive or negative definite.*

Before we can introduce the operators needed for the above factorization, we must consider some notational details. On the inner boundary ∂D we define

$$v^{\pm}(x) = \lim_{t\to 0^+} v(x \pm t\nu) \quad \text{and} \quad \frac{\partial v}{\partial \nu}^{\pm}(x) = \lim_{t\to 0^+} \nu \cdot \nabla v(x \pm t\nu),$$

for $x \in \partial D$ with $\nu(x)$ the unit normal pointing out of D, and further,

$$[v]_{\partial D} = v^+ - v^- \quad \text{and} \quad \left[\sigma \frac{\partial v}{\partial \nu}\right]_{\partial D} = \frac{\partial v}{\partial \nu}^{+} - \kappa \frac{\partial v}{\partial \nu}^{-}.$$

Let us now define L and L'. By replacing Ω with $\Omega \setminus \overline{D}$ and choosing $\Gamma_n = \partial D$ in (2.7) and Theorem 2.3, we note that for every $\phi \in H_0^{-1/2}(\partial D)$ the boundary value problem

$$\begin{aligned} &\Delta v = 0 \ \text{ in } \Omega \setminus \overline{D}, \quad \frac{\partial v}{\partial \nu} = 0 \ \text{ on } \partial\Omega \setminus \overline{\Gamma}_e, \quad \frac{\partial v}{\partial \nu}^{+} = \phi \ \text{ on } \partial D, \\ &v + z\frac{\partial v}{\partial \nu} = V \ \text{ on } \Gamma_e, \quad \frac{1}{|e_m|}\int_{e_m} \frac{\partial v}{\partial \nu} dS = 0, \quad 1 \le m \le M, \end{aligned} \tag{3.4}$$

has a unique solution $(v, V) \in H^1(\Omega \setminus \overline{D}) \oplus T_0$, where we have fixed the ground level of the potential. Thus, we may define the operator L by

$$L : H_0^{-1/2}(\partial D) \to T_0, \quad \phi \mapsto V. \tag{3.5}$$

With $I' \in T_0$, let us next consider the boundary value problem

$$\begin{aligned} &\Delta v' = 0 \ \text{ in } \Omega \setminus \overline{D}, \quad \frac{\partial v'}{\partial \nu} = 0 \ \text{ on } \partial\Omega \setminus \overline{\Gamma}_e, \quad \frac{\partial v'}{\partial \nu}^{+} = 0 \ \text{ on } \partial D, \\ &v' + z\frac{\partial v'}{\partial \nu} = V' \ \text{ on } \Gamma_e, \quad \frac{1}{|e_m|}\int_{e_m} \frac{\partial v'}{\partial \nu} dS = -I'_m, \quad 1 \le m \le M, \end{aligned} \tag{3.6}$$

which, according to Theorem 2.3, also has a unique solution $(v', V') \in H^1_{0,\partial D}(\Omega \setminus \overline{D}) \oplus T$, where

$$H^1_{0,\partial D}(\Omega \setminus \overline{D}) = \left\{ u \in H^1(\Omega \setminus \overline{D}) \;\middle|\; \int_{\partial D} u dS = 0 \right\}. \tag{3.7}$$

We define L' by

$$L' : T_0 \to H_0^{1/2}(\partial D), \quad I' \mapsto v'|_{\partial D}. \tag{3.8}$$

The following lemma shows that L and L' are bounded and adjoint, and have the mapping properties advertised above.

LEMMA 3.4. *The operators* $L : H_0^{-1/2}(\partial D) \to T_0$ *and* $L' : T_0 \to H_0^{1/2}(\partial D)$ *defined by* (3.5) *and* (3.8)*, respectively, are bounded* (*independently of the geometry of* Γ_e) *and adjoint. Further,* L *is surjective and* L' *is injective.*

Proof. We begin with the boundedness of L. For $\phi \in H_0^{-1/2}(\partial D)$ let $(v, V) \in H^1(\Omega \setminus \overline{D}) \oplus T_0$ be the unique solution of (3.4), suggested by Theorem 2.3. From the continuous dependence on the data (2.8) and since $V \in T_0 \subset L_0^2(\Gamma_e)$, it follows that

$$||V||_{L^2(\Gamma_e)} = ||V||_{L^2(\Gamma_e)/\mathbb{R}} \leq ||(v,V)||_{(H^1(\Omega\setminus\overline{D})\oplus L^2(\Gamma_e))/\mathbb{R}} \leq C\,||\phi||_{H^{-1/2}(\partial D)}\,,$$

which proves the continuity of $L : H_0^{-1/2}(\partial D) \to T_0$.

Next we shall prove that $L' : T_0 \to H_0^{1/2}(\partial D)$ is the adjoint of L. Let $(v, V) \in H^1(\Omega \setminus \overline{D}) \oplus T_0$ and $(v', V') \in H^1_{0,\partial D}(\Omega \setminus \overline{D}) \oplus T$ be the unique solutions of (3.4) and (3.6), respectively. Then it holds that

$$\begin{aligned}
\langle I', L\phi\rangle_{L^2(\partial\Omega)} &= \int_{\Gamma_e} \left(I' + \frac{\partial v'}{\partial \nu}\right) V dS - \int_{\Gamma_e} \frac{\partial v'}{\partial \nu} V dS \\
&= -\int_{\Gamma_e} \frac{\partial v'}{\partial \nu} V dS = -\int_{\Gamma_e} \frac{\partial v'}{\partial \nu}\left(v + z\frac{\partial v}{\partial \nu}\right) dS \\
&= -\int_{\Gamma_e} \frac{\partial v'}{\partial \nu} v dS - \int_{\Gamma_e} \left(z\frac{\partial v'}{\partial \nu} - V'\right)\frac{\partial v}{\partial \nu} dS \\
&= -\int_{\Gamma_e} \frac{\partial v'}{\partial \nu} v dS + \int_{\Gamma_e} v' \frac{\partial v}{\partial \nu} dS \\
&= -\int_{\partial D} \frac{\partial v'}{\partial \nu}^{+} v dS + \int_{\partial D} v' \frac{\partial v}{\partial \nu}^{+} dS = \langle L'I', \phi\rangle_{L^2(\partial D)},
\end{aligned}$$

where we used the boundary conditions that the pairs (v, V) and (v', V') satisfy together with Green's formula. Since L is bounded and L' is its adjoint operator, L' is also bounded.

The injectivity of L' is easy to obtain: Let $I' \in T_0$ be such that $L'I' = v'|_{\partial D} = 0$, which means, according to (3.6), that the Cauchy data of v' vanishes on ∂D. Since v' is harmonic in $\Omega \setminus \overline{D}$, this implies that $v' = 0$ from which it also follows that $I' = 0$. In addition, due to the finite-dimensionality of $\mathcal{R}(L)$, we have $T_0 = \mathcal{N}(L')^{\perp} = \overline{\mathcal{R}(L)} = \mathcal{R}(L)$, which proves the surjectivity of L. This completes the proof. ∎

Last but not least, let us introduce $F : H_0^{1/2}(\partial D) \to H_0^{-1/2}(\partial D)$. Let $\psi \in H_0^{1/2}(\partial D)$ and assume that $(w_\sigma, W_\sigma) \in (H^1(\Omega \setminus \partial D) \oplus T)/\mathbb{R}$ is the solution of the diffraction problem

$$\begin{aligned}
&\Delta w = 0 \ \text{ in } \Omega \setminus \partial D, \quad \frac{\partial w}{\partial \nu} = 0 \ \text{ on } \partial\Omega \setminus \overline{\Gamma}_e, \quad w + z\frac{\partial w}{\partial \nu} = W \ \text{ on } \Gamma_e, \\
&[w]_{\partial D} = \psi, \quad \left[\sigma \frac{\partial w}{\partial \nu}\right]_{\partial D} = 0, \quad \frac{1}{|e_m|}\int_{e_m} \frac{\partial w}{\partial \nu} dS = 0, \ \ 1 \leq m \leq M.
\end{aligned} \tag{3.9}$$

We define F by the mapping rule $\psi \mapsto \frac{\partial (w_\sigma - w_{\mathbf{1}})}{\partial \nu}^{+}|_{\partial D}$, where $w_{\mathbf{1}}$ is the solution of (3.9) with σ replaced by the unit conductivity $\mathbf{1}$.

Because the outer boundary condition of (3.9) is not of standard form, one must convince oneself that w_σ and $w_{\mathbf{1}}$, and thereby $F : H_0^{1/2}(\partial D) \to H_0^{-1/2}(\partial D)$, are actually well defined. The following two technical lemmas and a corollary answer all the necessary questions.

LEMMA 3.5. *For $\psi \in H_0^{1/2}(\partial D)$, diffraction problem* (3.9) *has a unique solution $(w_\sigma, W_\sigma) \in (H^1(\Omega \setminus \partial D) \oplus T)/\mathbb{R}$. Further,*

$$\left|\left| \frac{\partial w_\sigma}{\partial \nu}^+ \right|\right|_{H^{-1/2}(\partial D)} \leq C \, ||\psi||_{H^{1/2}(\partial D)}, \tag{3.10}$$

where $C > 0$ is independent of the geometry of Γ_e as a subset of $\partial\Omega$.

Proof. To start with, note that the corresponding traditional diffraction problem

$$\Delta w = 0 \ \text{ in } \Omega \setminus \partial D, \quad \frac{\partial w}{\partial \nu} = 0 \ \text{ on } \partial\Omega,$$
$$[w]_{\partial D} = \psi, \quad \left[\sigma \frac{\partial w}{\partial \nu}\right]_{\partial D} = 0$$

has a unique solution $w_0 \in H^1_{0,\partial\Omega}(\Omega \setminus \partial D)$ (cf. [8]), where the space is defined in equivalence with (3.7). Encouraged by this, we consider the following boundary value problem:

$$\begin{gathered} \nabla \cdot \sigma \nabla w = 0 \ \text{ in } \Omega, \quad \frac{\partial w}{\partial \nu} = 0 \ \text{ on } \partial\Omega \setminus \overline{\Gamma}_e, \\ w + z \frac{\partial w}{\partial \nu} = W - w_0 \ \text{ on } \Gamma_e, \quad \frac{1}{|e_m|} \int_{e_m} \frac{\partial w}{\partial \nu} dS = 0, \ \ 1 \leq m \leq M, \end{gathered} \tag{3.11}$$

and try to show that it has a unique solution $(w_e, W_e) \in (H^1(\Omega) \oplus T)/\mathbb{R}$.

From considerations in section 2.2, it follows in a straightforward manner that problem (3.11) is equivalent to the variational problem

$$B((w, W), (v, V)) = \int_{\Gamma_e} \frac{1}{z} w_0 (V - v) dS, \tag{3.12}$$

for all $(v, V) \in (H^1(\Omega) \oplus T)/\mathbb{R}$, where the bilinear form B is defined in (2.11). Since, according to Corollary 2.6, $B : (H^1(\Omega) \oplus T)/\mathbb{R} \times (H^1(\Omega) \oplus T)/\mathbb{R} \to \mathbb{R}$ is continuous and coercive, and the right-hand side of (3.12) clearly defines a continuous linear functional on $(H^1(\Omega) \oplus T)/\mathbb{R}$, equation (3.12) has a unique solution $(w_e, W_e) \in (H^1(\Omega) \oplus T)/\mathbb{R}$ due to the Lax–Milgram lemma [13]. Now, it is easy to see that $(w_0 + w_e, W_e) \in (H^1(\Omega \setminus \partial D) \oplus T)/\mathbb{R}$ satisfies the electrode diffraction problem given in (3.9) because of the continuity conditions that w_e and $\sigma \frac{\partial w_e}{\partial \nu}$ must satisfy on ∂D [8]. In particular, (3.9) has at least one solution.

Assume now that $(w_\sigma, W_\sigma) \in (H^1(\Omega \setminus \partial D) \oplus T)/\mathbb{R}$ is a solution of diffraction problem (3.9) corresponding to $\psi \in H_0^{1/2}(\partial D)$. Then, due to Green's formula and positivity of z, w_σ satisfies

$$\begin{aligned} \left|\left| \sigma^{1/2} \nabla w_\sigma \right|\right|^2_{L^2(\Omega)} &= \int_{\partial D} w_\sigma^- \kappa \frac{\partial w_\sigma}{\partial \nu}^- dS - \int_{\partial D} w_\sigma^+ \frac{\partial w_\sigma}{\partial \nu}^+ dS + \int_{\partial\Omega} w_\sigma \frac{\partial w_\sigma}{\partial \nu} dS \\ &= \int_{\partial D} (w_\sigma^- - w_\sigma^+) \frac{\partial w_\sigma}{\partial \nu}^+ dS + \int_{\Gamma_e} \left(W_\sigma - z \frac{\partial w_\sigma}{\partial \nu} \right) \frac{\partial w_\sigma}{\partial \nu} dS \\ &= - \int_{\partial D} \psi \frac{\partial w_\sigma}{\partial \nu}^+ dS - \int_{\Gamma_e} z \frac{\partial w_\sigma}{\partial \nu}^2 dS \\ &\leq ||\psi||_{H^{1/2}(\partial D)} \left|\left| \frac{\partial w_\sigma}{\partial \nu}^+ \right|\right|_{H^{-1/2}(\partial D)}, \end{aligned} \tag{3.13}$$

where we used the jump and boundary conditions of (3.9). Due to the boundedness of the mapping (see p. 381 in [5]),

$$H(\mathrm{div}, \Omega \setminus \overline{D}) \to H^{-1/2}(\partial D), \quad \mathbf{v} \mapsto (\nu \cdot \mathbf{v})^+|_{\partial D}, \tag{3.14}$$

where $H(\mathrm{div}, \Omega \setminus \overline{D}) = \{\mathbf{v} \in L^2(\Omega \setminus \overline{D})^n \,|\, \nabla \cdot \mathbf{v} \in L^2(\Omega \setminus \overline{D})\}$, it is true that

$$\left|\left| \frac{\partial w_\sigma}{\partial \nu}^+ \right|\right|_{H^{-1/2}(\partial D)} \leq C \, ||\nabla w_\sigma||_{L^2(\Omega)} \,, \tag{3.15}$$

where $C > 0$ has nothing to do with Γ_e. Using this once in (3.13), we get

$$||\nabla w_\sigma||_{L^2(\Omega)} \leq C \, ||\psi||_{H^{1/2}(\partial D)} \,, \tag{3.16}$$

from which it follows that the only solution of (3.9) corresponding to $\psi = 0$ is the zero element of $(H^1(\Omega \setminus \partial D) \oplus T)/\mathbb{R}$. Consequently, diffraction problem (3.9) has a unique solution. Together with (3.15), (3.16) also proves (3.10), which completes the proof. □

COROLLARY 3.6. *For $\psi \in H_0^{1/2}(\partial D)$, the solution of diffraction problem* (3.9)*, $(w_\sigma, W_\sigma) \in (H^1(\Omega \setminus \partial D) \oplus T)/\mathbb{R}$, is the unique minimizer of the energy functional*

$$E_\sigma(w, W) = \int_{\Omega \setminus \overline{D}} |\nabla w|^2 dx + \kappa \int_D |\nabla w|^2 dx + \int_{\Gamma_e} \frac{1}{z} |W - w|^2 dS$$

over the subset

$$H_\psi = \{(w, W) \in (H^1(\Omega \setminus \partial D) \oplus T)/\mathbb{R} \mid [w]_{\partial D} = \psi\}.$$

Proof. Let $(w, W) \in H_\psi$ be arbitrary and denote the difference $(w - w_\sigma, W - W_\sigma)$ by (v, V). In consequence, $(v, V) \in (H^1(\Omega \setminus \partial D) \oplus T)/\mathbb{R}$ with $[v]_{\partial D} = 0$ and we may write

$$\begin{aligned} E_\sigma(w, W) = E_\sigma(w_\sigma, W_\sigma) + E_\sigma(v, V) + 2 \Bigg\{ & \int_{\Omega \setminus \overline{D}} \nabla w_\sigma \cdot \nabla v dx + \kappa \int_D \nabla w_\sigma \cdot \nabla v dx \\ & + \int_{\Gamma_e} \frac{1}{z} (W_\sigma - w_\sigma)(V - v) dS \Bigg\}. \end{aligned} \tag{3.17}$$

We claim that the mixed terms on the right-hand side of (3.17) vanish.

Indeed, by Green's formula

$$\begin{aligned} \int_{\Omega \setminus \overline{D}} \nabla w_\sigma \cdot \nabla v dx + \kappa \int_D \nabla w_\sigma \cdot \nabla v dx &= \int_{\partial \Omega} \frac{\partial w_\sigma}{\partial \nu} v dS + \int_{\partial D} \left(\kappa \frac{\partial w_\sigma}{\partial \nu}^- - \frac{\partial w_\sigma}{\partial \nu}^+ \right) v dS \\ &= \int_{\Gamma_e} \frac{\partial w_\sigma}{\partial \nu} v dS, \end{aligned} \tag{3.18}$$

due to the jump condition of the normal derivative in (3.9). On the other hand,

$$\int_{\Gamma_e} \frac{1}{z} (W_\sigma - w_\sigma)(V - v) dS = \int_{\Gamma_e} \frac{\partial w_\sigma}{\partial \nu} (V - v) dS = - \int_{\Gamma_e} \frac{\partial w_\sigma}{\partial \nu} v dS, \tag{3.19}$$

which, together with (3.17), (3.18), and the positivity of E_σ, implies that

$$E_\sigma(w, W) \geq E_\sigma(w_\sigma, W_\sigma). \tag{3.20}$$

Since $E_\sigma(v,V)=0$ implicates that v, with $[v]_{\partial D}=0$, is constant on Ω and that $V=v|_{\Gamma_e}$, the inequality in (3.20) is strict for $(w,W)\neq(w_\sigma,W_\sigma)$ in H_ψ, which completes the proof. ∎

LEMMA 3.7. *The operator $F:H_0^{1/2}(\partial D)\to H_0^{-1/2}(\partial D)$ is well defined, continuous, self-adjoint, and bijective, with a continuous inverse operator $F^{-1}:H_0^{-1/2}(\partial D)\to H_0^{1/2}(\partial D)$. In addition, the operator $\operatorname{sgn}(1-\kappa)F:H_0^{1/2}(\partial D)\to H_0^{-1/2}(\partial D)$ is positive. Furthermore, there exist constants $C^+,C^->0$, independent of the geometry of Γ_e as a subset of $\partial\Omega$, such that*

$$||F||\le C^+,\quad \left|\left|F^{-1}\right|\right|\le C^-. \tag{3.21}$$

Proof. For $\psi\in H_0^{1/2}(\partial D)$, let $(w_\sigma,W_\sigma),(w_{\mathbf{1}},W_{\mathbf{1}})\in(H^1(\Omega\setminus\partial D)\oplus T)/\mathbb{R}$ be the solutions of (3.9) corresponding to the conductivities σ and $\mathbf{1}$, respectively. First of all, according to the divergence theorem,

$$\int_{\partial D}\frac{\partial(w_\sigma-w_{\mathbf{1}})}{\partial\nu}^{+}dS=\int_{\partial\Omega}\frac{\partial(w_\sigma-w_{\mathbf{1}})}{\partial\nu}dS=0,$$

from which it follows that $\frac{\partial(w_\sigma-w_{\mathbf{1}})}{\partial\nu}^{+}|_{\partial D}\in H_0^{-1/2}(\partial D)$. Together with Lemma 3.5, this proves that F is well defined and continuous and that the first part of (3.21) holds.

Next we want to establish the self-adjointness. To this end, for $\psi_1,\psi_2\in H_0^{1/2}(\partial D)$ let $(w_1,W_1),(w_2,W_2)$ be the corresponding solutions of diffraction problem (3.9) with the conductivity σ. By using Green's formula and the boundary conditions of (3.9), we may write

$$\begin{aligned}\int_{\partial D}\frac{\partial w_1}{\partial\nu}^{+}\psi_2dS&=\int_{\partial D}\frac{\partial w_1}{\partial\nu}^{+}w_2^+dS-\int_{\partial D}\kappa\frac{\partial w_1}{\partial\nu}^{-}w_2^-dS\\&=\int_{\partial\Omega}\left(\frac{\partial w_1}{\partial\nu}w_2-w_1\frac{\partial w_2}{\partial\nu}\right)dS+\int_{\partial D}\left(w_1^+\frac{\partial w_2}{\partial\nu}^{+}-\kappa w_1^-\frac{\partial w_2}{\partial\nu}^{-}\right)dS\\&=\int_{\Gamma_e}\frac{\partial w_1}{\partial\nu}\left(W_2-z\frac{\partial w_2}{\partial\nu}\right)dS-\int_{\Gamma_e}\frac{\partial w_2}{\partial\nu}\left(W_1-z\frac{\partial w_1}{\partial\nu}\right)dS\\&\quad+\int_{\partial D}(w_1^+-w_1^-)\frac{\partial w_2}{\partial\nu}^{+}dS=\int_{\partial D}\psi_1\frac{\partial w_2}{\partial\nu}^{+}dS.\end{aligned}$$

Since this holds also for $\mathbf{1}$ as conductivity, we actually have

$$\langle F\psi_1,\psi_2\rangle_{L^2(\partial D)}=\langle F\psi_2,\psi_1\rangle_{L^2(\partial D)};$$

i.e., F is self-adjoint.

Next we prove the positiveness of $\operatorname{sgn}(1-\kappa)F:H_0^{1/2}(\partial D)\to H_0^{-1/2}(\partial D)$. For $\psi\in H_0^{1/2}(\partial D)$, $\psi\neq0$, let (w_σ,W_σ) and $(w_{\mathbf{1}},W_{\mathbf{1}})$ be the solutions of (3.9) corresponding to the conductivities σ and $\mathbf{1}$, respectively. By careful use of Green's formula and the jump conditions of (3.9), we deduce

$$\begin{aligned}-\int_{\partial D}\frac{\partial w_\sigma}{\partial\nu}^{+}\psi dS&=\kappa\int_{\partial D}\frac{\partial w_\sigma}{\partial\nu}^{-}w_\sigma^-dS-\int_{\partial D}\frac{\partial w_\sigma}{\partial\nu}^{+}w_\sigma^+dS\\&=\int_{\Omega\setminus\overline{D}}|\nabla w_\sigma|^2dx+\kappa\int_D|\nabla w_\sigma|^2dx-\int_{\Gamma_e}\frac{\partial w_\sigma}{\partial\nu}w_\sigma dS\\&=E_\sigma(w_\sigma,W_\sigma),\end{aligned}$$

where the last equality follows from a slight modification of (3.19) in the proof of Corollary 3.6. Since similar reasoning also applies for $(w_{\mathbf{1}}, W_{\mathbf{1}})$, we have altogether

$$\langle F\psi, \psi\rangle_{L^2(\partial D)} = E_{\mathbf{1}}(w_{\mathbf{1}}, W_{\mathbf{1}}) - E_\sigma(w_\sigma, W_\sigma). \tag{3.22}$$

Assume first that $\kappa > 1$. Then, according to Corollary 3.6, it holds that

$$E_{\mathbf{1}}(w_{\mathbf{1}}, W_{\mathbf{1}}) < E_{\mathbf{1}}(w_\sigma, W_\sigma) \le E_\sigma(w_\sigma, W_\sigma).$$

Similarly, for $\kappa < 1$ we have

$$E_\sigma(w_\sigma, W_\sigma) < E_\sigma(w_{\mathbf{1}}, W_{\mathbf{1}}) \le E_{\mathbf{1}}(w_{\mathbf{1}}, W_{\mathbf{1}}).$$

Together with (3.22), these estimates prove the claim.

Then it is time to concentrate on the invertibility of F, beginning with the injectivity. Let $\psi \in H_0^{1/2}(\partial D)$ be such that $F\psi = 0$, meaning that the restricted difference $((w_\sigma - w_{\mathbf{1}})|_{\Omega\setminus\overline{D}}, W_\sigma - W_{\mathbf{1}}) \in (H^1(\Omega\setminus\overline{D}) \oplus T)/\mathbb{R}$ of the solutions to (3.9) satisfies boundary value problem (3.4) with $\phi = 0$. Thus, it follows from the unique solvability of (3.4) (see Theorem 2.3) that $w_\sigma = w_{\mathbf{1}} + c$, $c \in \mathbb{R}$, on $\Omega\setminus\overline{D}$, and as a consequence $w_\sigma^- = w_\sigma^+ - \psi = w_{\mathbf{1}}^+ - \psi + c = w_{\mathbf{1}}^- + c$ on ∂D. Hence, from the unique solvability of the Dirichlet problem

$$\Delta w = 0 \quad \text{in } D, \quad w = w_{\mathbf{1}}^- \quad \text{on } \partial D,$$

it follows that $w_\sigma = w_{\mathbf{1}} + c$ also in D. Combining these with the jump conditions of the normal derivatives in (3.9), on ∂D we have

$$\frac{\partial w_\sigma}{\partial \nu}^- = \frac{\partial w_{\mathbf{1}}}{\partial \nu}^- = \frac{\partial w_{\mathbf{1}}}{\partial \nu}^+ = \frac{\partial w_\sigma}{\partial \nu}^+ = \kappa \frac{\partial w_\sigma}{\partial \nu}^-;$$

i.e, all these normal derivatives must vanish. In consequence, (w_σ, W_σ) satisfies (3.4) with $\phi = 0$ in $\Omega\setminus\overline{D}$ and, in addition, w_σ satisfies Neumann problem with zero input current in D, meaning that $w_\sigma|_{\Omega\setminus\overline{D}}$ and $w_\sigma|_D$ equal constants. Hence, $\psi = w_\sigma^+ - w_\sigma^- \in H_0^{1/2}(\partial D)$ equals a constant which must be zero due to the normalization condition. Thus, $F : H_0^{1/2}(\partial D) \to H_0^{-1/2}(\partial D)$ is injective.

Next we move on to prove the surjectivity of F. For arbitrarily chosen $\phi \in H_0^{-1/2}(\partial D)$ we aim to construct $\psi \in H_0^{1/2}(\partial D)$ such that $F\psi = \phi$. First, we define an auxiliary pair $(v, V) \in (H^1(\Omega\setminus\overline{D}) \oplus T)/\mathbb{R}$ as the unique solution of (3.4) with the input current ϕ on ∂D, and we continue v to D as the unique H^1-solution of the Dirichlet problem

$$\Delta v = 0 \quad \text{in } D, \quad v^- = v^+ \quad \text{on } \partial D. \tag{3.23}$$

Hence, $(v, V) \in (H^1(\Omega\setminus\partial D) \oplus T)/\mathbb{R}$ with $[v]_{\partial D} = 0$. Further, we define $\varphi = \phi - \kappa \frac{\partial v}{\partial \nu}^-|_{\partial D}$ and note that $\varphi \in H_0^{-1/2}(\partial D)$ since $\int_{\partial D} \frac{\partial v}{\partial \nu}^- dS = 0$ due to the divergence theorem.

The next step is to define the diffraction solution corresponding to the unit conductivity. In the exterior domain $\Omega\setminus\overline{D}$ we choose $(w_{\mathbf{1}}, W_{\mathbf{1}}) \in (H^1(\Omega\setminus\overline{D}) \oplus T)/\mathbb{R}$ to be the unique solution of (3.4) with $\phi = \frac{1}{\kappa-1}\varphi$, whereas in the inner domain D we define $w_{\mathbf{1}}$ to be the unique H^1-solution of the Neumann problem

$$\Delta w_{\mathbf{1}} = 0 \quad \text{in } D, \quad \frac{\partial w_{\mathbf{1}}}{\partial \nu}^- = \frac{1}{\kappa - 1}\varphi \quad \text{on } \partial D, \quad \int_{\partial D} w_{\mathbf{1}}^- dS = \int_{\partial D} w_{\mathbf{1}}^+ dS. \tag{3.24}$$

Clearly, $(w_{\mathbf{1}}, W_{\mathbf{1}}) \in (H^1(\Omega \setminus \partial D) \oplus T)/\mathbb{R}$. As mentioned above, $(w_{\mathbf{1}}, W_{\mathbf{1}})$ plays here the role of the solution to diffraction problem (3.9) with conductivity $\mathbf{1}$ and, hence, we set $\psi = [w_{\mathbf{1}}]_{\partial D}$, which belongs to $H_0^{1/2}(\partial D)$ because of the normalization condition in (3.24). It is a straightforward task to check that the pairs $(w_{\mathbf{1}}, W_{\mathbf{1}}), (w_\sigma, W_\sigma) = (w_{\mathbf{1}} + v, W_{\mathbf{1}} + V) \in (H^1(\Omega \setminus \partial D) \oplus T)/\mathbb{R}$ satisfy diffraction problem (3.9) for the conductivities $\mathbf{1}$ and σ, respectively. Moreover, it holds that

$$F\psi = \frac{\partial (w_\sigma - w_{\mathbf{1}})^+}{\partial \nu}|_{\partial D} = \frac{\partial v^+}{\partial \nu}|_{\partial D} = \phi,$$

which proves that $F : H_0^{1/2}(\partial D) \to H_0^{-1/2}(\partial D)$ is surjective.

It is a consequence of the open mapping theorem that the inverse of the bijective bounded linear operator F is also bounded. Moreover, by walking the above constructional proof of the surjectivity in the opposite direction and using the continuous dependence on the boundary data of (3.4), (3.23), and (3.24), one easily sees that $F^{-1} : H_0^{-1/2}(\partial D) \to H_0^{1/2}(\partial D)$ is, actually, uniformly bounded with respect to the choice of the electrode configuration, i.e., with respect to the geometry of Γ_e as a subset of $\partial\Omega$. This completes the proof. ∎

Now we have gathered enough weaponry to prove the factorization of $R_\sigma - R_{\mathbf{1}}$.

Proof of Theorem 3.3. For a fixed electrode current $I \in T_0$ denote by (u_σ, U_σ), $(u_{\mathbf{1}}, U_{\mathbf{1}}) \in H^1(\Omega) \oplus T_0$ the solutions of the complete electrode forward problem, given in (3.1), with conductivities σ and $\mathbf{1}$, respectively. Since $u_\sigma - u_{\mathbf{1}}$ is harmonic in $\Omega \setminus \overline{D}$, it follows easily by using the divergence theorem and the complete electrode boundary conditions that

$$\int_{\partial D} \frac{\partial (u_\sigma - u_{\mathbf{1}})^+}{\partial \nu} dS = \int_{\partial \Omega} \frac{\partial (u_\sigma - u_{\mathbf{1}})}{\partial \nu} dS = 0.$$

Thus, $((u_\sigma - u_{\mathbf{1}})|_{\Omega \setminus \overline{D}}, U_\sigma - U_{\mathbf{1}})$ solves (3.4) for $\phi = \frac{\partial (u_\sigma - u_{\mathbf{1}})^+}{\partial \nu}|_{\partial D}$ and, in particular,

$$L\left(\frac{\partial (u_\sigma - u_{\mathbf{1}})^+}{\partial \nu}|_{\partial D} \right) = U_\sigma - U_{\mathbf{1}} = (R_\sigma - R_{\mathbf{1}})I.$$

By introducing the operator $G_\sigma : I \mapsto \frac{\partial u_\sigma}{\partial \nu}^+|_{\partial D}$ and setting $G = G_\sigma - G_{\mathbf{1}}$, we have so far derived the factorization

$$R_\sigma - R_{\mathbf{1}} = LG. \tag{3.25}$$

Note that G is a well-defined bounded operator from T_0 to $H_0^{-1/2}(\partial D)$ due to Theorem 2.3 and (3.14).

The next task is to calculate the dual operator $G'_\sigma : H_0^{1/2}(\partial D) \to T_0$ of G_σ. To this end, consider $(w_\sigma, W_\sigma) \in H^1(\Omega \setminus \partial D) \oplus T_0$ the solution of diffraction problem (3.9), with a fixed ground level of the potential, corresponding to $\psi \in H_0^{1/2}(\partial D)$. With the help of the jump conditions $[u_\sigma]_{\partial D} = \left[\sigma \frac{\partial u_\sigma}{\partial \nu}\right]_{\partial D} = 0$ (cf. [8]), $\left[\sigma \frac{\partial w_\sigma}{\partial \nu}\right]_{\partial D} = 0$, Green's formula, and the boundary conditions on u_σ and w_σ, we deduce

$$\begin{aligned}
\langle G_\sigma I, \psi \rangle_{L^2(\partial D)} &= \int_{\partial D} \frac{\partial u_\sigma}{\partial \nu}^+ w_\sigma^+ dS - \int_{\partial D} \kappa \frac{\partial u_\sigma}{\partial \nu}^- w_\sigma^- dS \\
&= \int_{\partial D} \left(\frac{\partial w_\sigma}{\partial \nu}^+ u_\sigma^+ - \kappa \frac{\partial w_\sigma}{\partial \nu}^- u_\sigma^- \right) dS + \int_{\partial \Omega} \left(\frac{\partial u_\sigma}{\partial \nu} w_\sigma - \frac{\partial w_\sigma}{\partial \nu} u_\sigma \right) dS
\end{aligned}$$

$$= \int_{\Gamma_e} \frac{\partial u_\sigma}{\partial \nu} w_\sigma dS + \int_{\Gamma_e} \frac{\partial w_\sigma}{\partial \nu}(U_\sigma - u_\sigma) dS = \int_{\Gamma_e} \frac{\partial u_\sigma}{\partial \nu} \left(w_\sigma + z \frac{\partial w_\sigma}{\partial \nu} \right) dS$$
$$= \int_{\Gamma_e} \left(\frac{\partial u_\sigma}{\partial \nu} - I \right) W_\sigma dS + \int_{\Gamma_e} I W_\sigma dS = \langle I, W_\sigma \rangle_{L^2(\partial\Omega)},$$

which shows that $G'_\sigma \psi = W_\sigma$. Hence, with $(w_{\mathbf{1}}, W_{\mathbf{1}}) \in H^1(\Omega \setminus \partial D) \oplus T_0$ the solution of diffraction problem (3.9) corresponding to ψ and the unit conductivity, we have

$$G'\psi = W_\sigma - W_{\mathbf{1}}.$$

The restriction $((w_\sigma - w_{\mathbf{1}})|_{\Omega \setminus \overline{D}}, W_\sigma - W_{\mathbf{1}}) \in H^1(\Omega \setminus \overline{D}) \oplus T_0$ solves (3.4) for $\phi = \frac{\partial (w_\sigma - w_{\mathbf{1}})}{\partial \nu}^+ |_{\partial D} \in H_0^{-1/2}(\partial D)$, which means that

$$L \left(\frac{\partial (w_\sigma - w_{\mathbf{1}})}{\partial \nu}^+ |_{\partial D} \right) = W_\sigma - W_{\mathbf{1}} = G'\psi. \tag{3.26}$$

Due to the way $F : H_0^{1/2}(\partial D) \to H_0^{-1/2}(\partial D)$ is defined and since $\psi \in H_0^{1/2}(\partial D)$ was chosen arbitrarily, relation (3.26) is equivalent to $LF = G'$. Taking the transpose of this and plugging it into (3.25), we thus obtain

$$R_\sigma - R_{\mathbf{1}} = LF'L' = LFL',$$

which is what we set out to prove. □

3.2. Some further properties of F, L, and L'. We define a new boundary operator by

$$|R_\sigma - R_{\mathbf{1}}| = \operatorname{sgn}(1 - \kappa)(R_\sigma - R_{\mathbf{1}}). \tag{3.27}$$

Due to the way we have defined our conductivity in (3.3), it follows trivially from Lemma 3.1 and Corollary 3.2 that $|R_\sigma - R_{\mathbf{1}}| : T_0 \to T_0$ is self-adjoint and strictly positive. Denoting the operator $\operatorname{sgn}(1 - \kappa)F : H_0^{1/2}(\partial D) \to H_0^{-1/2}(\partial D)$ by $|F|$, it follows from Theorem 3.3 that this new boundary operator can be factorized as $|R_\sigma - R_{\mathbf{1}}| = L|F|L'$. In the next subsection we will use the operator $|R_\sigma - R_{\mathbf{1}}|$ to characterize the inclusion D. However, to be successful in this task, we must devote the ongoing subsection to further investigations of $|F|$, L, and L'.

LEMMA 3.8. *The operator $|F| : H_0^{1/2}(\partial D) \to H_0^{-1/2}(\partial D)$ can be given as $|F| = F^{1/2}(F^{1/2})'$, where $F^{1/2} : L_0^2(\partial D) \to H_0^{-1/2}(\partial D)$ and $(F^{1/2})' : H_0^{1/2}(\partial D) \to L_0^2(\partial D)$ are bounded, bijective, and dual to each other. Further, it holds that*

$$\left|\left| F^{1/2} \right|\right| \leq C^+ \sqrt{C^-}, \quad \left|\left| F^{-1/2} \right|\right| \leq \sqrt{C^-},$$

where $C^+, C^- > 0$ are the constants introduced in Lemma 3.7.

Proof. Since $H_0^{1/2}(\partial D) \hookrightarrow L_0^2(\partial D) \hookrightarrow H_0^{-1/2}(\partial D)$ is a Gelfand triple and since $|F|^{-1} : H_0^{-1/2}(\partial D) \to H_0^{1/2}(\partial D)$ is isomorphic, self-adjoint, and positive, it follows from material in [3] that there exists a factorization

$$|F|^{-1} = (F^{-1/2})' F^{-1/2},$$

where $F^{-1/2} : H_0^{-1/2}(\partial D) \to L_0^2(\partial D)$ and $(F^{-1/2})' : L_0^2(\partial D) \to H_0^{1/2}(\partial D)$ are bounded, bijective, and dual to each other, with

$$\left|\left|F^{-1/2}\right|\right| = \left|\left|(F^{-1/2})'\right|\right| \leq \left|\left|F^{-1}\right|\right|^{1/2} \leq \sqrt{C^-}. \tag{3.28}$$

Further, for $\eta \in L_0^2(\partial D)$ we may estimate

$$\left|\left|F^{1/2}\eta\right|\right|_{H^{-1/2}(\partial D)} = \left|\left||F|(F^{-1/2})'\eta\right|\right|_{H^{-1/2}(\partial\Omega)} \leq C^+\sqrt{C^-}\,||\eta||_{L^2(\partial D)},$$

where we used Lemma 3.7 and (3.28). ∎

In what follows we denote by $\mathcal{N}(L)^\perp \subset H_0^{-1/2}(\partial D)$ the orthogonal complement of $\mathcal{N}(L) \subset H_0^{-1/2}(\partial D)$ with respect to the inner product of the Hilbert space $H^{-1/2}(\partial D)$. Let $Q : \mathcal{R}(FL') \to \mathcal{N}(L)^\perp$ be an orthogonal projection; i.e., for $\phi \in \mathcal{R}(FL') \subset H_0^{-1/2}(\partial D)$,

$$Q\phi = \phi_\perp \in \mathcal{N}(L)^\perp \quad \text{with } L\phi_\perp = L\phi. \tag{3.29}$$

Note that Q is well defined due to the projection theorem [6]. In addition, we claim that Q is a bijection.

COROLLARY 3.9. *The orthogonal projection $Q : \mathcal{R}(FL') \to \mathcal{N}(L)^\perp$ defined by* (3.29) *is bijective with the norm estimates*

$$||Q|| \leq 1, \quad \left|\left|Q^{-1}\right|\right| \leq (C^+C^-)^2, \tag{3.30}$$

where $C^+, C^- > 0$ are the constants introduced in Lemma 3.7.

Proof. To begin with, note that the left-hand inequality of (3.30) is obvious. In order to obtain the right-hand inequality, let $\phi \in \mathcal{R}(FL') = \mathcal{R}(|F|L')$ and $\phi_\perp = Q\phi \in \mathcal{N}(L)^\perp$, and note that $\phi - \phi_\perp \in \mathcal{N}(L)$. In consequence, we may write

$$\begin{aligned} ||\phi_\perp||_{H^{-1/2}(\partial D)} &\geq \sup_{||\psi||_{H^{1/2}}=1, \psi\in\mathcal{R}(L')} \langle \phi_\perp, \psi\rangle_{L^2(\partial D)} \\ &= \sup_{||\psi||_{H^{1/2}}=1, \psi\in\mathcal{R}(L')} \langle \phi, \psi\rangle_{L^2(\partial D)}. \end{aligned}$$

Further, since $|F|^{-1}\phi \in \mathcal{R}(L')$, we have

$$||\phi_\perp||_{H^{-1/2}(\partial D)} \geq \frac{1}{||\,|F|^{-1}\phi||_{H^{1/2}(\partial D)}} \langle \phi, |F|^{-1}\phi\rangle_{L^2(\partial D)} = \frac{\left|\left|F^{-1/2}\phi\right|\right|^2_{L^2(\partial D)}}{||\,|F|^{-1}\phi||_{H^{1/2}(\partial D)}},$$

and so we finally obtain

$$||\phi_\perp||_{H^{-1/2}(\partial D)} \geq \frac{||\phi||_{H^{-1/2}(\partial D)}}{\left|\left|F^{1/2}\right|\right|^2 ||\,|F|^{-1}||} \geq \frac{1}{(C^+C^-)^2}\,||\phi||_{H^{-1/2}(\partial D)} \tag{3.31}$$

by Lemmas 3.7 and 3.8.

According to Lemmas 3.4 and 3.7, $L : H_0^{-1/2}(\partial D) \to T_0$ is surjective, $L' : T_0 \to H_0^{1/2}(\partial D)$ is injective, and $F : H_0^{1/2}(\partial D) \to H_0^{-1/2}(\partial D)$ is bijective. Thus, $\dim(\mathcal{N}(L)^\perp) = \dim(\mathcal{R}(FL')) = \dim(T_0) < \infty$, and so the bijectivity of $Q : \mathcal{R}(FL') \to \mathcal{N}(L)^\perp$ follows from its injectivity that is guaranteed by (3.31), which provides also the needed norm estimate for Q^{-1}. ∎

To end this subsection, we make a few comments about the inverse operators of L and L' defined in (3.5) and (3.8), respectively. Since $L : H_0^{-1/2}(\partial D) \to T_0$ is noninjective and $L' : T_0 \to H_0^{1/2}(\partial D)$ is nonsurjective, they do not have inverse operators as such. However, the restrictions $L : \mathcal{R}(FL') \to T_0$ and $L' : T_0 \to \mathcal{R}(L')$ do have bounded inverses due to the bijectivity of $R_\sigma - R_{\mathbf{1}} = LFL' : T_0 \to T_0$ and finite-dimensionality of T_0. In what follows, we will denote by L^{-1} and $(L')^{-1}$ the inverses of these restrictions, i.e.,

$$L^{-1} : T_0 \to \mathcal{R}(FL'), \quad (L')^{-1} : \mathcal{R}(L') \to T_0, \tag{3.32}$$

with $LL^{-1} = \mathrm{id}$, $L^{-1}L|_{\mathcal{R}(FL')} = \mathrm{id}$ and $L'(L')^{-1} = \mathrm{id}$, $(L')^{-1}L' = \mathrm{id}$. With this notation, we can factorize $|R_\sigma - R_{\mathbf{1}}|^{-1} : T_0 \to T_0$, which exists according to Lemma 3.2, as

$$|R_\sigma - R_{\mathbf{1}}|^{-1} = (L')^{-1}|F|^{-1}L^{-1} = (L')^{-1}(F^{-1/2})'F^{-1/2}L^{-1},$$

due to Theorem 3.3 and Lemma 3.8.

3.3. Characterizing the inclusion. Before we can formulate and prove the main result of this section, we need to introduce some new concepts. Let $\{\mathcal{T}_M\}$ be a sequence of electrode configurations, meaning that

$$\mathcal{T}_M = \{e_1^M, \ldots, e_M^M \subset \partial\Omega \mid e_l^M \cap e_m^M = \emptyset \text{ if } l \neq m\}, \quad \Gamma_M = \cup_{m=1}^M e_m^M,$$

for each $1 \leq M < \infty$, satisfying the following conditions: $d(e_m^M) \leq \beta_M$ for all $1 \leq m \leq M$,

$$|\partial\Omega \setminus \Gamma_M|, \ \beta_M \to 0 \quad \text{when } M \to \infty, \tag{3.33}$$

where $d(e_m^M)$ is the diameter of e_m^M, i.e., $d(e_m^M) = \sup_{x,y\in e_m^M} |x-y|$. The subspaces T^M and T_0^M, corresponding to the electrode configuration $\mathcal{T}_M$, are defined in accordance with (2.5) and (2.6), respectively, and the associated orthogonal projections $P_1^M : L_0^2(\partial\Omega) \to L_0^2(\Gamma_M)$, $P_2^M : L_0^2(\Gamma_M) \to T_0^M$, and $P^M : L_0^2(\partial\Omega) \to T_0^M$ are given by obvious modifications of (2.3), (2.9), and (2.26). We also will use a similar index notation for other operators depending on the used electrode configuration.

Let $y \in \Omega$ be a parameter and $\hat{\alpha} \in \mathbb{R}^n$ a unit vector, and consider the solution Φ_y of the following homogenous Neumann problem:

$$\Delta\Phi(x) = \hat{\alpha} \cdot \nabla\delta(x-y) \quad \text{in } \Omega, \quad \frac{\partial\Phi}{\partial\nu} = 0 \quad \text{on } \partial\Omega, \quad \int_{\partial\Omega} \Phi dS = 0, \tag{3.34}$$

where δ is the delta functional. Physically Φ_y corresponds to the electromagnetic potential created by a dipole point source at y pointing in the direction $\hat{\alpha}$. It is a well-known fact that (3.34) is uniquely solvable with $\Phi_y \in C^\infty(\Omega \setminus \{y\})$ and Φ_y singular at y.

Assume that (3.33) is valid and let $\{\alpha_M\} \subset \mathbb{R}_+$ be a sequence of regularization parameters. Consider the minimizing sequence $\{I^M\} \subset L_0^2(\partial\Omega)$, $I^M \in T_0^M$, of the Tikhonov functionals

$$\left|\left||R_\sigma^M - R_{\mathbf{1}}^M|^{1/2} I - \Phi_y\right|\right|^2_{L^2(\partial\Omega)} + \alpha_M \, ||I||^2_{L^2(\partial\Omega)}, \quad I \in T_0^M, \ 1 \leq M < \infty, \tag{3.35}$$

where $|R_\sigma^M - R_{\mathbf{1}}^M|^{1/2} : T_0^M \to T_0^M$ is the unique, positive, self-adjoint, bijective square root of $|R_\sigma^M - R_{\mathbf{1}}^M|$ defined in (3.27). Since $R_\sigma^M - R_{\mathbf{1}}^M$ can be obtained through boundary measurements, so can $|R_\sigma^M - R_{\mathbf{1}}^M|^{1/2}$ and, hence, the behavior of the sequence

$\{I^M\}$ is something that can be observed by noninvasive methods. The following theorem characterizes the inclusion D by the limit behavior of $\{I^M\}$.

THEOREM 3.10. *Assume that* (3.33) *holds and the contact impedance z is smooth. Let $\{I^M\} \subset L^2_0(\partial\Omega)$, $I^M \in T^M_0$, be the minimizing sequence for the functionals* (3.35) *and assume that $\{\alpha_M\} \subset \mathbb{R}_+$ converges to zero but is such that the sequence*

$$\left\{ \frac{\inf_{V \in T^M} ||\Phi_y - V||^2_{L^2(\partial\Omega)}}{\alpha_M} \right\}$$

is bounded. Then $y \in D$ if and only if the sequence $\{I^M\}$ is bounded in $L^2_0(\partial\Omega)$.

In real life one is, naturally, not able to construct a sequence of electrode configurations with the properties given in (3.33). However, when conducting measurements with a fixed setting of electrodes that are relatively small and cover a large portion of the object boundary, Theorem 3.10 gives a reason to believe that the electrode currents needed for minimizing functional (3.35), with a fixed small $\alpha > 0$, are larger when $y \in \Omega \setminus D$ than when $y \in D$. This observation leads to a possibility of numerical implementation that will be considered in forthcoming articles.

The following simple lemma shows that the conditions of Theorem 3.10 are reasonable. One could also quite easily derive a quantitative estimate to suggest an a priori choice of regularization parameters α_M in Theorem (3.10) but for simplicity we content ourselves with a mere convergence result.

LEMMA 3.11. *Let $f \in C^\infty(\partial\Omega)$ and assume that $\{\mathcal{T}_M\}$ satisfies* (3.33). *Then it holds that*

$$\inf_{V \in T^M} ||f - V||_{L^2(\partial\Omega)} \to 0,$$

when M goes to infinity.

Proof. The claim is a straightforward consequence of the good behavior of $\{\mathcal{T}_M\}$ given by (3.33). ∎

The rest of this section is devoted to the proof of Theorem 3.10. Let L^M be the operator defined in (3.5) corresponding to the electrode configuration $\mathcal{T}_M$. We define the associated limit operator $\tilde{L} : H_0^{-1/2}(\partial D) \to H_0^{1/2}(\partial\Omega)$ by

$$\tilde{L}\phi = v|_{\partial\Omega}, \quad \phi \in H_0^{-1/2}(\partial D),$$

where $v \in H^1_{0,\partial\Omega}(\Omega \setminus \overline{D})$ is the unique solution of the boundary value problem

$$\Delta v = 0 \quad \text{in } \Omega \setminus \overline{D}, \quad \frac{\partial v}{\partial \nu}^+ = \phi \quad \text{on } \partial D, \quad \frac{\partial v}{\partial \nu} = 0 \quad \text{on } \partial\Omega. \tag{3.36}$$

The adjoint of $\tilde{L}$ is $\tilde{L}' : H_0^{-1/2}(\partial\Omega) \to H_0^{1/2}(\partial D)$ [3],

$$\tilde{L}'\phi' = v'|_{\partial D}, \quad \phi' \in H_0^{-1/2}(\partial\Omega), \tag{3.37}$$

where $v' \in H^1_{0,\partial D}(\Omega \setminus \overline{D})$ is the unique solution of the boundary value problem

$$\Delta v' = 0 \quad \text{in } \Omega \setminus \overline{D}, \quad \frac{\partial v'}{\partial \nu} = -\phi' \quad \text{on } \partial\Omega, \quad \frac{\partial v}{\partial \nu}^+ = 0 \quad \text{on } \partial D.$$

The first step of our proof is to characterize the inclusion D by the operator sequence $\{L_M\}$ with the help of known mapping properties of $\tilde{L}$ and $\tilde{L}'$.

Assume that $\{\alpha_M\}$ is a sequence of positive regularization parameters that converges to zero and consider the following Tikhonov functionals:

$$\left|\left|L^M\phi-\Phi_y\right|\right|^2_{L^2(\partial\Omega)}+\alpha_M\left|\left|\phi\right|\right|^2_{H^{-1/2}(\partial D)},\quad 1\le M\le\infty. \tag{3.38}$$

Since $L^M : H_0^{-1/2}(\partial D)\to T_0^M\subset L_0^2(\partial\Omega)$ is continuous for every $M\in\mathbb{N}$, it is well known that each of these functionals has a unique minimizer $\phi^M\in H_0^{-1/2}(\partial D)$. We intend to show that, for a correctly chosen sequence of regularization parameters $\{\alpha_M\}$, the behavior of the minimizer sequence $\{\phi^M\}$ at infinity determines uniquely whether y belongs to the inclusion D or not. We begin with the case when $y\in\Omega\setminus D$.

LEMMA 3.12. *Assume that $y\in\Omega\setminus D$, the contact impedance z is smooth, and $\{\alpha_M\}\subset\mathbb{R}_+$ converges to zero. Let $\{\phi^M\}\subset H_0^{-1/2}(\partial D)$ be the minimizing sequence for the functionals* (3.38). *Then it holds that*

$$\left|\left|\phi^M\right|\right|_{H^{-1/2}(\partial D)}\to\infty,$$

as M goes to infinity.

Proof. First, we will show that $L^M\phi^M$ converges to $\Phi_y|_{\partial\Omega}$ as M goes to infinity. Let $\epsilon>0$ be given. Since $\tilde{L}'$ defined in (3.37) is clearly injective, we have $\overline{\mathcal{R}(\tilde{L})}=\mathcal{N}(\tilde{L}')^\perp=H_0^{1/2}(\partial\Omega)$, where the orthogonal complement is taken with respect to the dual pairing between $H_0^{-1/2}(\partial\Omega)$ and $H_0^{1/2}(\partial\Omega)$. Hence, $\mathcal{R}(\tilde{L})$ is dense in $H_0^{1/2}(\partial\Omega)$ and, thus, also in $L_0^2(\partial\Omega)$. In consequence, we can choose $\phi^\epsilon\in H_0^{-1/2}(\partial D)$ such that

$$\left|\left|\tilde{L}\phi^\epsilon-\Phi_y\right|\right|^2_{L^2(\partial\Omega)}<\frac{\epsilon^2}{6}. \tag{3.39}$$

Note also that $\tilde{L}\phi^\epsilon\in C^\infty(\partial\Omega)\cap H_0^{1/2}(\partial\Omega)$ due to the regularity theory of elliptic partial differential equations [10].

Since $\tilde{L}\phi^\epsilon\in H_0^{1/2}(\partial\Omega)$ and $L^M\phi^\epsilon\in L_0^2(\Gamma_M)$, by using the projection $P_1^M : L_0^2(\partial\Omega)\to L_0^2(\Gamma_M)$, defined by (2.3), we can estimate

$$\begin{aligned}\left|\left|(\tilde{L}-L^M)\phi^\epsilon\right|\right|_{L^2(\Gamma_M)}&\le\left|\left|\tilde{L}\phi^\epsilon-P_1^M(\tilde{L}\phi^\epsilon)\right|\right|_{L^2(\Gamma_M)}+\left|\left|P_1^M(\tilde{L}-L^M)\phi^\epsilon\right|\right|_{L^2(\Gamma_M)}\\&\le C\left\{\frac{|\partial\Omega\setminus\overline{\Gamma}_M|^{1/2}}{|\Gamma_M|^{1/2}}\left|\left|\tilde{L}\phi^\epsilon\right|\right|_{L^2(\partial\Omega\setminus\overline{\Gamma}_M)}+\left|\left|(\tilde{L}-L^M)\phi^\epsilon\right|\right|_{L^2(\Gamma_M)/\mathbb{R}}\right\}\\&\le C\left\{\frac{|\partial\Omega\setminus\overline{\Gamma}_M|^{1/2}}{|\Gamma_M|^{1/2}}\left|\left|\tilde{L}\phi^\epsilon\right|\right|_{L^2(\partial\Omega\setminus\overline{\Gamma}_M)}+\inf_{V\in T^M}\left|\left|\tilde{L}\phi^\epsilon-V\right|\right|_{L^2(\Gamma_M)}\right\},\end{aligned}$$

where the second-to-last inequality follows from (2.3), by using the Schwarz inequality, and the fact that $P_1^M(\tilde{L}-L^M)\phi^\epsilon\in L_0^2(\Gamma_M)$, and the last inequality follows from Lemma 2.7 and Theorem 2.11 applied on boundary value problems (3.36) and (3.4). Thus, according to Lemma 3.11, we can choose $M_0\in\mathbb{N}$ in such a way that

$$\left|\left|(L^M-\tilde{L})\phi^\epsilon\right|\right|^2_{L^2(\partial\Omega)}\le\frac{C}{|\Gamma_M|}\inf_{V\in T^M}\left|\left|\tilde{L}\phi^\epsilon-V\right|\right|^2_{L^2(\partial\Omega)}<\frac{\epsilon^2}{6}, \tag{3.40}$$

and, in addition,

$$\alpha_M\left|\left|\phi^\epsilon\right|\right|^2_{H^{-1/2}(\partial D)}<\frac{\epsilon^2}{3} \tag{3.41}$$

for all $M \geq M_0$. Consequently, due to estimates (3.39), (3.40), (3.41), and the triangle inequality, for every $M \geq M_0$ it holds that

$$\begin{aligned}\left|\left|L^M\phi^M - \Phi_y\right|\right|^2_{L^2(\partial\Omega)} + \alpha_M \left|\left|\phi^M\right|\right|^2_{H^{-1/2}(\partial D)} \\ \leq \left|\left|L^M\phi^\epsilon - \Phi_y\right|\right|^2_{L^2(\partial\Omega)} + \alpha_M \left|\left|\phi^\epsilon\right|\right|^2_{H^{-1/2}(\partial D)} < \epsilon^2.\end{aligned}$$

In particular, since $\epsilon > 0$ was chosen arbitrarily, we have obtained

$$\left|\left|L^M\phi^M - \Phi_y\right|\right|_{L^2(\partial\Omega)} \to 0,$$

when M goes to infinity.

Next, we will use contradiction: Assume that the minimizing sequence $\{\phi^M\}$ is bounded in $H_0^{-1/2}(\partial D)$. In consequence, it follows from fundamental functional analysis [6] that $\{\phi^M\}$ has a subsequence $\{\phi^{M_k}\}_{k=1}^\infty$ that converges weakly to some distribution $\phi' \in H_0^{-1/2}(\partial D)$. Our goal is to show that $\tilde{L}\phi' = \Phi_y|_{\partial\Omega}$, which is a contradiction due to the singularity of Φ_y at $y \in \Omega \setminus D$ [3].

Let $g \in C^\infty(\partial\Omega) \cap L^2_0(\partial\Omega)$ be arbitrary and write it in two parts as $g = P_1^{M_k} g + (I - P_1^{M_k})g$, where $P_1^{M_k}$ is defined by (2.3). Then we have

$$\langle L^{M_k}\phi^{M_k}, g\rangle_{L^2(\partial\Omega)} = \langle L^{M_k}\phi^{M_k}, P^{M_k}g\rangle_{L^2(\partial\Omega)} + \langle L^{M_k}\phi^{M_k}, (I - P_1^{M_k})g\rangle_{L^2(\Gamma_{M_k})}, \tag{3.42}$$

where we used the fact that $L^{M_k}\phi^{M_k}$ is constant over each $e_m^{M_k}$ and zero elsewhere, and the way P^{M_k} is defined in (2.26). Due to the uniform boundedness of the operators $\{L^{M_k}\} \subset \mathcal{L}(H_0^{-1/2}(\partial D), L^2_0(\partial\Omega))$ (see Lemma 3.4) and of the sequence $\{\phi^M\} \subset H_0^{-1/2}(\partial D)$, the second term on the right-hand side of (3.42) can be estimated by the Schwarz inequality as follows:

$$\begin{aligned}\left|\langle L^{M_k}\phi^{M_k}, (I - P_1^{M_k})g\rangle_{L^2(\Gamma_{M_k})}\right| &\leq \frac{C}{|\Gamma_{M_k}|}\left\{\int_{\Gamma_{M_k}}\left|\int_{\partial\Omega\setminus\overline{\Gamma}_{M_k}} g dS\right|^2 dS\right\}^{1/2} \\ &\leq \frac{C}{|\Gamma_{M_k}|^{1/2}}|\partial\Omega \setminus \Gamma_{M_k}|\,||g||_\infty \to 0,\end{aligned}$$

when k goes to infinity due to (3.33). On the other hand, for the first term on the right-hand side of (3.42) we may write

$$\begin{aligned}\langle L^{M_k}\phi^{M_k}, P^{M_k}g\rangle_{L^2(\partial\Omega)} &= \langle \phi^{M_k}, (L^{M_k})' P^{M_k} g\rangle_{L^2(\partial D)} \\ &= \langle \phi^{M_k}, ((L^{M_k})' P^{M_k} - \tilde{L}')g\rangle_{L^2(\partial D)} + \langle \phi^{M_k}, \tilde{L}' g\rangle_{L^2(\partial D)}.\end{aligned} \tag{3.43}$$

Let $(v^{M_k}, V^{M_k}) \in H^1_{0,\partial D}(\Omega \setminus \overline{D}) \oplus T^{M_k}$ and $v \in H^1_{0,\partial D}(\Omega \setminus \overline{D})$ be the solutions corresponding to the operator current pairs $((L^{M_k})', P^{M_k}g)$ and $(\tilde{L}', g)$, respectively; i.e., by (3.8) and (3.37), $(L^{M_k})' P^{M_k} g = v^{M_k}|_{\partial D}$ and $\tilde{L}' g = v|_{\partial D}$. Since $\phi^{M_k} \in H_0^{-1/2}(\partial D)$, by a slight variation of Lemma 2.7 and Theorem 2.4, we have

$$\begin{aligned}\left|\langle \phi^{M_k}, ((L^{M_k})' P^{M_k} - \tilde{L}')g\rangle_{L^2(\partial D)}\right| &\leq \left|\left|\phi^{M_k}\right|\right|_{H^{-1/2}(\partial D)} \left|\left|v^{M_k} - v\right|\right|_{H^{1/2}(\partial D)/\mathbb{R}} \\ &\leq C\left|\left|v^{M_k} - v\right|\right|_{H^1(\Omega\setminus\overline{D})/\mathbb{R}} \\ &\leq C\left\{\inf_{V\in T^{M_k}} ||(v - zg) - V||_{L^2(\Gamma_{M_k})}\right. \\ &\quad \left. + \frac{1}{|\Gamma_{M_k}|^{1/2}}\,||g||_{\tilde{H}^{-1/2}(\partial\Omega\setminus\overline{\Gamma}_{M_k})}\right\} \to 0,\end{aligned}$$

when k goes to infinity by Lemma 3.11 and (3.33) since $v|_{\partial\Omega} - zg \in C^\infty(\partial\Omega)$ due to the assumptions on z and g, and the regularity theory of elliptic partial differential equations [10]. Finally, due to the weak convergence of $\{\phi^{M_k}\}$, the second term on the right-hand side of (3.43) satisfies

$$\langle \phi^{M_k}, \tilde{L}' g\rangle_{L^2(\partial D)} \to \langle \phi', \tilde{L}' g\rangle_{L^2(\partial D)} = \langle \tilde{L}\phi', g\rangle_{L^2(\partial\Omega)},$$

when k goes to infinity.

Putting the above estimates together, we have established that

$$\langle L^{M_k}\phi^{M_k}, g\rangle_{L^2(\partial\Omega)} \to \langle \tilde{L}\phi', g\rangle_{L^2(\partial\Omega)} = \langle \Phi_y, g\rangle_{L^2(\partial\Omega)} \quad \text{when } k \to \infty,$$

for all $g \in C^\infty(\partial\Omega) \cap L^2_0(\partial\Omega)$, by the first part of the proof. This means that $\tilde{L}\phi' = \Phi_y|_{\partial\Omega}$ almost everywhere on $\partial\Omega$, which is the contradiction we were looking for. ∎

Then it is the turn of $y \in D$.

LEMMA 3.13. *Assume that $y \in D$ and let $\{\alpha_M\} \subset \mathbb{R}_+$ be such that the sequence*

$$\left\{ \frac{\inf_{V\in T^M} ||\Phi_y - V||^2_{L^2(\partial\Omega)}}{\alpha_M} \right\} \tag{3.44}$$

is bounded. Then the sequence of the minimizers $\{\phi^M\} \subset H_0^{-1/2}(\partial D)$ for (3.38) *also is bounded.*

Proof. To begin with, note that $\frac{\partial\Phi_y}{\partial\nu}^+|_{\partial D} \in H_0^{-1/2}(\partial D)$ due to the divergence theorem. Since $L^M \frac{\partial\Phi_y}{\partial\nu}^+ \in L^2_0(\Gamma_M)$ and, clearly, $\tilde{L}\frac{\partial\Phi_y}{\partial\nu}^+ = \Phi_y|_{\partial\Omega} \in H_0^{1/2}(\partial\Omega)$, as in the proof of Lemma 3.12, we have the estimate

$$\left|\left| L^M \frac{\partial\Phi_y}{\partial\nu}^+ - \Phi_y \right|\right|^2_{L^2(\partial\Omega)} = \left|\left| (L^M - \tilde{L})\frac{\partial\Phi_y}{\partial\nu}^+ \right|\right|^2_{L^2(\partial\Omega)} \leq \frac{C}{|\Gamma_M|} \inf_{V\in T^M} ||\Phi_y - V||^2_{L^2(\partial\Omega)}.$$

Thus, due to the minimizing property of the sequence $\{\phi^M\} \subset H_0^{-1/2}(\partial D)$, for every $M \in \mathbb{N}$, we have

$$\begin{aligned} &\left|\left|L^M\phi^M - \Phi_y\right|\right|^2_{L^2(\partial\Omega)} + \alpha_M \left|\left|\phi^M\right|\right|^2_{H^{-1/2}(\partial D)} \\ &\qquad \leq \frac{C}{|\Gamma_M|} \inf_{V\in T^M} ||\Phi_y - V||^2_{L^2(\partial\Omega)} + \alpha_M \left|\left| \frac{\partial\Phi_y}{\partial\nu}^+ \right|\right|^2_{H^{-1/2}(\partial D)}. \end{aligned} \tag{3.45}$$

Forgetting the first term on the left-hand side of (3.45) and dividing by α_M, we get

$$\left|\left|\phi^M\right|\right|^2_{H^{-1/2}(\partial D)} \leq C\frac{\inf_{V\in T^M} ||\Phi_y - V||^2_{L^2(\partial\Omega)}}{\alpha_M|\Gamma_M|} + \left|\left| \frac{\partial\Phi_y}{\partial\nu}^+ \right|\right|^2_{H^{-1/2}(\partial D)},$$

for every $M \in \mathbb{N}$. Together with assumptions (3.44) and (3.33), this proves the claim. ∎

If the operator sequence $\{L^M\}$ is known, Lemmas 3.12 and 3.13 give us the means to find the inclusion D. However, to know $\{L^M\}$ is to know the shape of the boundary ∂D. Luckily, the operators L^M and $|R^M_\sigma - R^M_{\mathbf{1}}|^{1/2}$ are closely related, and so Lemmas 3.12 and 3.13 give us the weaponry to write out the proof for Theorem 3.10.

Proof of Theorem 3.10. Let us define a new sequence $\{\tilde{\phi}^M\} \subset H_0^{-1/2}(\partial D)$, $\tilde{\phi}^M \in \mathcal{R}(F^M(L^M)')$, by

$$\tilde{\phi}^M = (L^M)^{-1}|R^M_\sigma - R^M_{\mathbf{1}}|^{1/2} I_M, \quad 1 \leq M < \infty,$$

where $(L^M)^{-1}$ is given by (3.32). We get a simple relation between the norms of I^M and $\tilde{\phi}^M$:

$$\begin{aligned}\left|\left|I^M\right|\right|^2_{L^2(\partial\Omega)} &= \langle |R^M_\sigma - R^M_{\mathbf{1}}|^{-1/2} L^M \tilde{\phi}^M, |R^M_\sigma - R^M_{\mathbf{1}}|^{-1/2} L^M \tilde{\phi}^M \rangle_{L^2(\partial\Omega)} \\ &= \langle \tilde{\phi}^M, (L^M)' |R^M_\sigma - R^M_{\mathbf{1}}|^{-1} L^M \tilde{\phi}^M \rangle_{L^2(\partial D)} \\ &= \langle \tilde{\phi}^M, (L^M)' ((L^M)')^{-1} |F^M|^{-1} (L^M)^{-1} L^M \tilde{\phi}^M \rangle_{L^2(\partial D)} \\ &= \langle \tilde{\phi}^M, |F^M|^{-1} \tilde{\phi}^M \rangle_{L^2(\partial D)} = \left|\left|(F^M)^{-1/2} \tilde{\phi}^M\right|\right|^2_{L^2(\partial D)}.\end{aligned}$$

In consequence, since the sequence $\{I^M\}$ minimizes the functionals (3.35), the sequence $\{\tilde{\phi}^M\}$ minimizes the functionals

$$\left|\left|L^M\phi - \Phi_y\right|\right|^2_{L^2(\partial\Omega)} + \alpha_M \left|\left|(F^M)^{-1/2}\phi\right|\right|^2_{L^2(\partial D)}, \quad 1 \le M < \infty, \tag{3.46}$$

within the subspaces $\mathcal{R}(F^M(L^M)')$, respectively. Indeed, if $\hat{\phi}^M \in \mathcal{R}(F^M(L^M)')$ gave a smaller value for functional (3.46), then one easily sees that $|R^M_\sigma - R^M_{\mathbf{1}}|^{-1/2} L^M \hat{\phi}^M \in T^M_0$ would give a smaller value than I^M for functional (3.35), which is a contradiction.

We define yet a new sequence by $\{\phi^M\} = \{Q^M \tilde{\phi}^M\}$, where $Q^M : \mathcal{R}(F^M(L^M)') \to \mathcal{N}(L^M)^\perp \subset H^{-1/2}_0(\partial D)$ is defined by (3.29). Here and in the rest of this proof the orthogonal complement $\mathcal{N}(L^M)^\perp$ is taken with respect to the $H^{-1/2}$ inner product. By similar reasoning as above, one sees that this new sequence minimizes the functionals

$$\left|\left|L^M\phi - \Phi_y\right|\right|^2_{L^2(\partial\Omega)} + \alpha_M \left|\left|(F^M)^{-1/2}(Q^M)^{-1}\phi\right|\right|^2_{L^2(\partial D)}, \quad 1 \le M < \infty,$$

over the subspaces $\mathcal{N}(L^M)^\perp$, respectively. Now, Lemma 3.8 and Corollary 3.9 tell us that there exists a sequence of functionals $\{C_M\}$, $C_M : \mathcal{N}(L^M)^\perp \to \mathbb{R}$, such that

$$\left|\left|(F^M)^{-1/2}(Q^M)^{-1}\phi\right|\right|_{L^2(\partial D)} = C_M(\phi)\, ||\phi||_{H^{-1/2}(\partial D)}, \quad c \le C_M \le C,$$

for all $M \in \mathbb{N}$ and $\phi \in \mathcal{N}(L^M)^\perp$, where c and C are positive constants independent of M. Thus, the sequence $\{\phi^M\}$ also minimizes the functionals

$$\left|\left|L^M\phi - \Phi_y\right|\right|^2_{L^2(\partial\Omega)} + \alpha_M C^2_M(\phi)\, ||\phi||^2_{H^{-1/2}(\partial D)}, \quad 1 \le M < \infty, \tag{3.47}$$

over the subspaces $\mathcal{N}(L^M)^\perp \subset H^{-1/2}_0(\partial D)$, respectively. In particular, if we define $C_M(\phi) = C$ for $\phi \in H^{-1/2}_0(\partial D) \setminus \mathcal{N}(L^M)^\perp$, the sequence $\{\phi^M\}$ minimizes functionals (3.47) over the whole space $H^{-1/2}_0(\partial D)$. It is an easy consequence of the upper and lower bounds for $\{C_M\}$ that $\{\phi^M\}$ is bounded in $H^{-1/2}_0(\partial D)$ if and only if $y \in D$.

Let $\{\phi^M_c\}, \{\phi^M_C\} \subset H^{-1/2}_0(\partial D)$ be the minimizing sequences for the functionals

$$\left|\left|L^M\phi - \Phi_y\right|\right|^2_{L^2(\partial\Omega)} + \frac{1}{2}\alpha_M c^2\, ||\phi||^2_{H^{-1/2}(\partial D)}, \quad 1 \le M < \infty,$$

and

$$\left|\left|L^M\phi - \Phi_y\right|\right|^2_{L^2(\partial\Omega)} + 2\alpha_M C^2\, ||\phi||^2_{H^{-1/2}(\partial D)}, \quad 1 \le M < \infty,$$

respectively. It follows from Lemmas 3.12 and 3.13 that each of these sequences is bounded if and only if $y \in D$. Let us shorten our strenuous notations by $\Psi_M(\phi) = \left|\left|L^M\phi - \Phi_y\right|\right|_{L^2(\partial\Omega)}$ and note that due to the minimizing properties of the sequences $\{\phi^M\}$ and $\{\phi_c^M\}$, for every $M \in \mathbb{N}$, we have

$$\Psi_M^2(\phi^M) + \alpha_M C_M^2(\phi^M) \left|\left|\phi^M\right|\right|_{H^{-1/2}(\partial D)}^2 \leq \Psi_M^2(\phi_c^M) + \alpha_M C_M^2(\phi_c^M) \left|\left|\phi_c^M\right|\right|_{H^{-1/2}(\partial D)}^2,$$

and

$$\Psi_M^2(\phi^M) + \frac{1}{2}\alpha_M c^2 \left|\left|\phi^M\right|\right|_{H^{-1/2}(\partial D)}^2 \geq \Psi_M^2(\phi_c^M) + \frac{1}{2}\alpha_M c^2 \left|\left|\phi_c^M\right|\right|_{H^{-1/2}(\partial D)}^2.$$

By subtracting the second of these inequalities from the first one and arranging terms, we get

$$\left|\left|\phi^M\right|\right|_{H^{-1/2}(\partial D)}^2 \leq \frac{C_M^2(\phi_c^M) - \frac{1}{2}c^2}{C_M^2(\phi^M) - \frac{1}{2}c^2} \left|\left|\phi_c^M\right|\right|_{H^{-1/2}(\partial D)}^2 \leq \frac{2C^2 - c^2}{c^2} \left|\left|\phi_c^M\right|\right|_{H^{-1/2}(\partial D)}^2.$$

On the other hand, by similar means we deduce that

$$\left|\left|\phi^M\right|\right|_{H^{-1/2}(\partial D)}^2 \geq \frac{2C^2 - C_M^2(\phi_C^M)}{2C^2 - C_M^2(\phi^M)} \left|\left|\phi_C^M\right|\right|_{H^{-1/2}(\partial D)}^2 \geq \frac{C^2}{2C^2 - c^2} \left|\left|\phi_C^M\right|\right|_{H^{-1/2}(\partial D)}^2.$$

From the above estimates it follows that $\{\phi^M\} \subset H_0^{-1/2}(\partial D)$ is bounded if and only if $y \in D$.

Finally, walking the above path of reasoning backwards, one sees that

$$\left|\left|I^M\right|\right|_{L^2(\partial\Omega)} = \left|\left|(F^M)^{-1/2}(Q^M)^{-1}\phi^M\right|\right|_{L^2(\partial D)},$$

and so the claim follows from the uniform boundedness of the operator sequences $\{(F^M)^{1/2}\}$, $\{(F^M)^{-1/2}\}$, and $\{Q^M\}$, $\{(Q^M)^{-1}\}$ given in Lemma 3.8 and Corollary 3.9, respectively. □

We end this section, and at the same time the whole work, by noting that one could easily modify Theorem 3.10 for the case of multiple inclusions by using the means described in [3].

REFERENCES

[1] L. Borcea, *Electrical impedance tomography*, Inverse Problems, 18 (2002), pp. R99–R136.

[2] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.

[3] M. Brühl, *Explicit characterization of inclusions in electrical impedance tomography*, SIAM J. Math. Anal., 32 (2001), pp. 1327–1341.

[4] F. Cakoni and D. Colton, *The linear sampling method for cracks*, Inverse Problems, 19 (2003), pp. 279–295.

[5] R. Dautray and J-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 2, Springer-Verlag, Berlin, 1988.

[6] V. Hutson and J. S. Pym, *Applications of Functional Analysis and Operator Theory*, Academic Press, London, 1980.

[7] A. Kirsch, *Characterization of the shape of the scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.

[8] O. A. Ladyzhenskaya, *The Boundary Value Problems of Mathematical Physics*, Springer-Verlag, New York, 1985.

[9] E. Somersalo, M. Cheney, and D. Isaacson, *Existence and uniqueness for electrode models for electric current computed tomography*, SIAM J. Appl. Math., 52 (1992), pp. 1023–1040.
[10] M. E. Taylor, *Partial Differential Equations* I, Springer-Verlag, New York, 1996.
[11] M. Vauhkonen, D. Vadász, P. A. Karjalainen, E. Somersalo, and J. P. Kaipio, *Tikhonov regularization and prior information in electrical impedance tomography*, IEEE Trans. Med. Imaging, 17 (1998), pp. 285–293.
[12] P. Vauhkonen, M. Vauhkonen, T. Savolainen, and J. Kaipio, *Three-dimensional electrical impedance tomography based on the complete electrode model*, IEEE Trans. Biomed. Eng., 46 (1999), pp. 1150–1160.
[13] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, New York, 1980.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 932–960

SIMULATION BY DIFFUSION ON A MANIFOLD WITH BOUNDARY: APPLICATIONS TO ULTRASONIC PRENATAL MEDICAL IMAGING*

DANIEL M. KEENAN† AND PAULA A. SHORTER‡

Abstract. Consider the ultrasonic imaging of a fetal head, a standard procedure for assessing the growth and development of the fetus in utero. The size and shape of certain cross sections of the skull are particularly important in such assessments. The technician/radiologist spatially moves an ultrasonic transducer over the pregnant woman's abdomen and acquires three-dimensional information from the real-time two-dimensional video images. The ability to draw conclusions from such images is built upon an acquired practical knowledge as to what is bone, what is tissue, and what is sensor noise. Can one mechanize this intuitive understanding which allows the technician/radiologist to delineate between inherent biological variation and variation due to sensor noise, and in so doing, reconstruct the fetal head? The approach of the present paper is that of a probabilistic recovery of the fetal head. The "space of realizable fetal heads" is viewed as the result of similarity (translation × scale × rotation) transformations being applied to a given fetal head prototype. A model of the ultrasonic imaging of a fetal head is formulated and prior knowledge is incorporated, resulting in an a posteriori probability measure on the "space of realizable fetal heads." Sampling from this probability measure is achieved and justified via a time-homogeneous diffusion on a Riemannian, compact, simply connected, oriented manifold with boundary. The discretization of the diffusion is implemented into code, and the algorithm is applied to actual ultrasound prenatal images.

Key words. medical image processing, stochastic differential equations, manifold with boundary, stochastic optimization

AMS subject classifications. 92C50, 60M10, 92C30, 62M09, 93E03

DOI. 10.1137/S0036139902408904

1. Introduction.

Within a medical imaging context, consider the prenatal ultrasound imaging of a fetal head. The transducer is moved, resulting in a spatial sequence of two-dimensional ultrasound images. In Figure 1, a sequence of eight such images (taken in roughly parallel planes, perpendicular to the spine of the fetus) is displayed proceeding clockwise from the jawline (top left) upward to the top of the head (bottom left). The technician/radiologist is able to observe the images in "real time," and they are stored on video at a rate of 33 frames per second. The objectives of the ultrasonic exam are to detect any potential abnormalities and to monitor the growth of the fetus via certain measurements of the fetal skull. Can one automate this process by mechanizing the technician's intuitive understanding of the inherent biological variation as well as the variation due to sensor noise? We develop a beginning formulation for such a methodology and apply it to the displayed ultrasound images.

One measurement commonly taken is the diameter of the fetal skull in a specified plane, located just above the ears. This measurement is called the biparietal diameter (BPD), and the plane is referred to as the BPD plane (pictured in the top row of

*Received by the editors June 3, 2002; accepted for publication (in revised form) July 1, 2003; published electronically March 30, 2004.
http://www.siam.org/journals/siap/64-3/40890.html
†Department of Statistics, University of Virginia, Charlottesville, VA 22903 (dmk7b@cramer.stat.virginia.edu). The research of this author was supported by an NSF Interdisciplinary Grant in the Mathematical Sciences DMS-0107680 and NIH (NIA) K01 award AG19164-01.
‡Department of Mathematics, Computer Science, and Physics, Rockhurst University, Kansas City, MO 64110 (paula.shorter@rockhurst.edu).

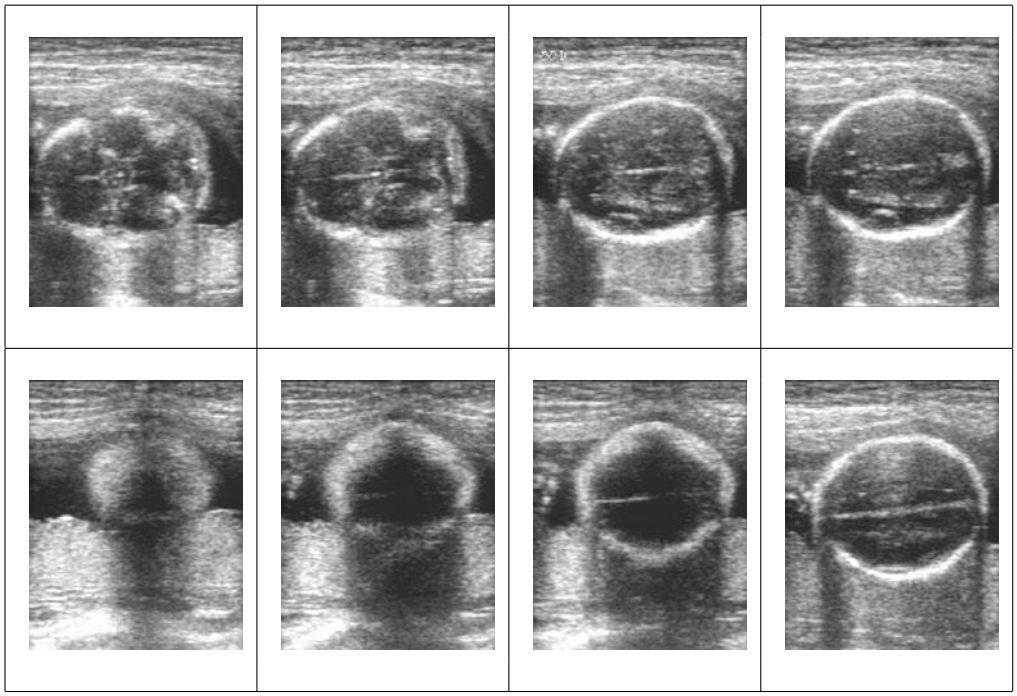


FIG. 1. *Ultrasound images of a fetal head. Images proceed clockwise from the jawline (top left) upward to the top of the head (bottom left).*

Figure 1, second from the right). Ideally, in order to automate measurements such as the BPD, the entire head would be reconstructed from the "ultrasound movie," and any measurements or other assessments would be made on the reconstructed head. That way, for example, even if one does not actually observe the BPD cross section, one would still be able to estimate the BPD measurement.

With the advent of telemedicine, where ultrasonic imaging may well be done remotely by the radiologist, possibly even in a semiautomated manner, being able to estimate the BPD under such conditions will be important. Another consideration is whether, by the incorporation of knowledge which allows for the distinction between what is bone, what is tissue, and what is sensor noise, it might be possible to perform prenatal ultrasonic assessments earlier in the development of the fetus. Traditionally, prenatal ultrasonic examinations of the fetal head are not performed until the bones of the fetal skull have calcified sufficiently to allow for visual recognition.

In the present context, imagine the collection of "realizable variations in form" of a fetal head. A basic strategy would then be to optimize over this collection, finding that variation which most likely could have produced the acquired two-dimensional ultrasound images. The difficulty is that it is computationally infeasible to store all such potential fetal head representations and search through them by brute force. The methodology presented in this paper is an attempt to do such a search in a semi-intelligent manner.

We consider this collection, $\overline{\Theta}(\mathcal{T})$, to be the result of a family of transformations, $\overline{\Theta}$, applied to a given fetal head template, $\mathcal{T}$. The transformations are assumed to be characterized by a compact manifold with boundary. In our particular application, the transformations are those of location $\times$ scale $\times$ rotation, but the formulation and

justification are established in a broader context. The resulting manifold structure is a consequence of the space of rotations $SO(3)$, and the boundary arises from the boundary of the space of allowable translations and/or from that of the space of possible scalings. A probability measure incorporating prior knowledge about fetal heads as well as the information provided by the acquired ultrasound images will be constructed on $\overline{\Theta}$. The ability to sample from this measure (stochastic relaxation; see [6], [7]) is the basis of the approach. The reconstruction will be obtained via such sampling.

More and more today, the stochastic modeling of complex phenomena requires formulations in terms of manifolds. The usual reason for this is that various nonlinear constraints inherent to the structure cannot be excluded or linearized. In pattern theory (see, e.g., [7], [8], [29]), a fundamental formulation of a parameter space is as a product of low-dimensional Lie groups, and when one takes into account certain global constraints, the resulting parameter space will be a submanifold (of the product) with boundary.

Below, our parameter space will be a manifold, assumed to be Riemannian, finite-dimensional, compact, connected, oriented, and with boundary. Having specified such a space (call it $\overline{\Theta} = \Theta \cup \partial\Theta$), one would then develop a likelihood function which describes how the data came about (i.e., a model of the technology) and a prior probability measure describing a priori knowledge of the parameter. Their joint interaction results in a posterior probability measure π on $\overline{\Theta}$.

After developing the posterior probability measure π, we develop methods for sampling from this posterior measure (with respect to a reference measure) via the "objective function"

$$h(\cdot) = -\ln q(\cdot),$$

where q is the posterior density function on $\overline{\Theta}$. The sampling will be achieved by running a discretization of a time-homogeneous diffusion on $\overline{\Theta}$. In a linear space, the diffusion would be defined by the Langevin equation

$$d\,X_t = -\frac{1}{2}\,\nabla h(X_t)dt \,+\, dW_t, \tag{1.1}$$

where $\{W_t, t \geq 0\}$ is a standard Wiener process. Heuristically thinking of h as a potential energy, the above diffusion would be defined by a Fokker–Planck equation for which $q = e^{-h}$ is the time-independent (i.e., steady-state) solution. To make this precise, one needs to prove an ergodic theorem (see section 8).

On a Riemannian manifold with boundary, the above diffusion in linear space is replaced by the diffusion $\{\theta_t, t \geq 0\}$ generated by the operator

$$A = \frac{1}{2}(\Delta - \nabla h) \tag{1.2}$$

and the boundary operator

$$L = \frac{\partial}{\partial n} \qquad \text{(derivative in the normal direction)},$$

where ∇ and Δ are the gradient and Laplacian on $\overline{\Theta}$, and there is reflection at the boundary.

In section 6, we discuss a practical issue that can arise in settings where the imaging is done in dimensions that are not specifically calculated (as can occur in

ultrasound). Since the original space $\overline{\Theta}$ is not completely determined in practice, in order to actually implement a discretization of the time-homogeneous diffusion process on $\overline{\Theta}$, we must run our algorithm on a normalized space, $\overline{\Theta}_0$, via a transformation of units. The consequence is that a scaling factor (from a change of variables) may be needed in order to put the gradient and the Brownian motion on the same scale.

2. Overview of algorithm and related literature.

2.1. Procedure: Simulation from the posterior measure π. The time-homogeneous diffusion $\{\theta_t, t \geq 0\}$ on $\overline{\Theta}$ will be shown to be ergodic with invariant measure π (whose density is q). Hence θ_t, for large t, represents (in an asymptotic sense) sampling from the desired posterior measure π, where π incorporates prior knowledge about θ and information from the observed data (in the present case, the ultrasound images).

2.2. Analysis: Automation of the BPD measurement. Recall that the motivating objective of our work is to automate various processes involved in the prenatal ultrasonic monitoring of fetal growth and development. In particular, the problem considered is that of automatically determining the BPD measurement of a fetus, given a sequence of two-dimensional ultrasound images of the fetal head. The approach taken here is to probabilistically recover the three-dimensional fetal head and then to acquire the BPD measurement from the recovered (estimated) head. An estimate for the "true" fetal head is obtained by sampling from the posterior measure on $\overline{\Theta}$ as described above.

We implement and apply a discretization of the diffusion $\{\theta_t, t \geq 0\}$ to actual ultrasound images of a fetal head. Each run of the algorithm results in an estimated "true" three-dimensional fetal head from which BPD measurements can be taken.

Note. To establish the theoretical basis for the present algorithm, an alternative approach could be pursued. One could imbed (by Whitney's theorem [34]) the d-dim manifold in $\mathbb{R}^N$ for some N, extend the function on the imbedded surface to a function on a larger N-dim rectangle without introducing any new global minima, and then apply the result of Geman and Hwang [5] in $\mathbb{R}^N$. (The present authors have, in fact, established such results.) The difficulty with such an approach is that the geometry of the imbedded surface could (and ordinarily would) be quite different from that of the original manifold, and the value of N needs to potentially be much larger ($2d$ or $2d+1$). One could consider an isometric imbedding (Nash's theorem [30]), but then the required dimension is even larger ($n(3n+11)/2$). The addition of a boundary in the present case complicates things even further. From a practical perspective, if one wants a theoretical structure for which the geometry is directly applicable to the construction of the algorithm, then it is most natural to establish the results directly, within the context of Riemannian geometry.

Our theoretical result can be viewed as a generalization to the time-homogeneous case for a manifold of the results of Geman and Hwang [5], who considered a compact rectangle in $\mathbb{R}^d$, and Chiang, Hwang, and Sheu [4], who extended the previous result to the case of all of $\mathbb{R}^d$. The focus of those papers was primarily on the time-inhomogeneous case, where dW_t is replaced by $c(t)dW_t$ with $c^2(t) \propto 1/(1+\ell n\, t)$ slowly decreasing to zero, as in simulated annealing. In Shorter [32] and Keenan and Shorter [21], we consider stochastic relaxation and simulated annealing on a manifold without boundary; the motivation of the latter is to perform maximum likelihood estimation (MLE) and MAP estimation for such examples as described in the next section. Holley, Kusuoka, and Stroock [14] consider continuous-time annealing on

a compact manifold, but with different goals in mind. Grenander and Miller [12] and Srivastava et al. [33] consider parameter spaces consisting of a countable union of Lie groups, each of a different dimension, and construct jump-diffusion processes, where the jumps are between the different Lie groups. There is no boundary to the manifolds within their formulations.

3. Prenatal ultrasonic assessment.

3.1. Ultrasound imaging. The underlying principles of ultrasound imaging are essentially the same as those in other echo ranging systems, where the range to an object is determined by knowing the speed at which sound is traveling in the given medium and measuring the amount of time required for a generated sound pulse to travel to and echo back from the object. Specifically in medical ultrasound, the transducer (the vibrating source that creates ultrasound waves) produces a short pulse of ultrasound, typically three or four vibrations, at a frequency in the 2–10 MHz range. The pulse propagates away from the transducer into the patient in the plane of the transducer.

One of the primary reasons that ultrasound can be used to accurately image structures within the human body is that the speed of sound (c) in different human soft tissues is very similar. Thus, in medical ultrasound c can be taken to be constant (usually 1540ms^{-1}), since in most cases the ultrasound wave is propagating through soft tissue of one kind or another. In prenatal ultrasonic examinations of the fetal head, the constant c is still taken to be 1540ms^{-1}, despite the significant presence of bone. In this case, the accuracy of the approximation is spared by the fact that the bones of the fetal skull are less calcified than in postnatal life.

As an ultrasound wave propagates through tissue, its direction and intensity change due to several processes, reflection being the most informative. Under certain circumstances, when an ultrasound wave encounters a boundary between different media (called an interface), it is redirected according to the laws characterizing optical reflection (the process occurring when light strikes a reflecting surface). The redirected wave, now referred to as an echo, reflects off of the interface at an angle equal to the angle of incidence. The various factors affecting the process of reflection at a particular interface include the size, roughness, and orientation of the interface. For optimum reflection, the interface must be large and smooth relative to the wavelength and oriented perpendicularly to the incident wave.

The intensity of the echo differs from that of the incident wave because not all of the incoming energy is reflected at an interface. A portion of the energy is reflected, but the rest is transmitted into the next medium, unaffected by the interface, where it is available for reflection by deeper structures. The fraction of the energy which is reflected depends upon the change in certain acoustic properties of the medium from one side of the boundary to the other.

For example, an echo reflecting from a boundary between two different soft tissues will have low intensity, while the transmitted wave will retain most of the original energy. At soft tissue/bone and soft tissue/air interfaces, however, most of the energy is reflected, resulting in strong (high intensity) echoes and weak (low intensity) transmitted waves. When the fraction of energy reflected is very large, "shadowing" occurs, where no structures beyond the interface are imaged because almost no energy is transmitted through the interface.

The same transducer that transmitted the original ultrasound pulse now acts as the receiver of its echoes. Many echoes, however, will not be detected. If the angle of reflection (equivalently, the angle of incidence) is very large at all, then the echo

may well miss the transducer and not be detected. Even if the reflected direction of the echo does point toward the transducer, the echo still may not be received if its intensity upon reflection is too low to sustain a detectable signal all the way back to the transducer. For those echoes that are received by the transducer, the imaging system is able to detect both the travel time and intensity of the wave. Travel time obviously gives the range to the object producing the echo, and intensity provides information about the reflectivity of the object. The form in which these values are displayed depends upon the particular type of ultrasound imaging system being used.

The imaging system considered here is called a B-scanner. The B-scanner produces recognizable two-dimensional images of anatomical sections. B-scan images are referred to as "slices" since what is being viewed in such images is the intersection of a plane (the plane of the transducer) with the three-dimensional object being imaged.

One ultrasound pulse creates one "scan line" of a B-scan image, extending from the transducer outward in the original direction of the wave. Each reflected echo from this pulse is plotted as a bright dot along the scan line at a distance equal to the calculated range of the object from the transducer. The intensity information is displayed using varying degrees of brightness (hence the name B (for brightness)-mode); the greater the amplitude, the greater the brightness of the dot. Therefore, a highly reflective interface will be imaged very brightly, while less reflective interfaces will be displayed less brightly.

A common arrangement of the scan lines is to make all of the scan lines parallel, creating a rectangular image. This format is produced by a "linear scanner" and is the format usually chosen for use in prenatal ultrasound. The simplest electronic arrangement that achieves the linear format consists of an array of small independent transducers lined up side by side, forming a "linear array probe." Each transducer in the array fires and receives its own scan line, corresponding to a single column in the image.

For the purposes of storage and display, the B-scan image is digitized and written into a matrix of memory (typically of sizes near 256×256 or 512×512). A given pixel corresponds to a single dot in the displayed image, and the integer gray scale value corresponding to the brightness of that dot is the measurement stored at that pixel location. The gray scale values are integers ranging from 0 to 255, where 0 denotes black, 255 denotes white, and the values between correspond to varying levels of gray. We henceforth can regard ultrasound images as $M \times K$ matrices whose components are integers ranging from 0 to 255.

The description given above of the process of ultrasonic imaging is, of course, greatly simplified. For example, there are several additional processes, besides reflection, which affect the direction and intensity of an ultrasound wave as it propagates through tissue. Two such processes are absorption (where a portion of the energy of a wave is lost as it is converted into heat) and scattering. Unlike reflection, which occurs at interfaces, scattering and absorption occur throughout the path of a wave.

Typically within regions of soft tissue there are many "targets" which are very small and rough. When an ultrasound wave strikes these targets, reflection does not occur since the process of reflection requires large flat targets. Scattering, however, does occur at such targets, resulting in relatively weak echoes redirected (scattered) in almost all directions. Most of the wave's energy is transmitted through the scattering point unchanged, but the intensity of the scattered echoes is often high enough to be detected by the receiver, producing dim discontinuous images of various soft tissue regions.

Regardless of the form of energy used by an imaging system, the digital images

produced are not "true" sampled versions of what is really there. For our purposes, the term "noise" will be used to describe all forms of degradation and distortion of the image. These include errors occurring in the sensing mechanism, blurring as a result of scattering, and artifacts such as shadowing, clutter (nonrandom scene structure other than an object of interest), and randomness inherent in both the physical system as well as the sensing and recording devices.

We can speak of a "true" or "ideal" image as the image we would obtain in an ideal system without noise, but the actual digital images produced by an imaging system will always contain some noise. We will call these actual images "noisy images." Specifically in ultrasound imaging systems, we notice that the images produced are very noisy.

(References for the material presented in this section include [13] and [24].)

3.2. BPD measurements. Ultrasonic imaging is particularly well suited for use in prenatal medicine because of its safety advantages and the membranous nature of some fetal bones. Doctors perform prenatal ultrasonic examinations to estimate gestational age as well as monitor the growth and development of the fetus in utero. Fetal growth is characterized by numerous anatomic changes, but measuring the change in size of various anatomic structures has become the primary means of assessing this growth. Since ultrasonic images are two-dimensional, the most practical measures of size are circumferences, diameters, and areas of anatomic planes. Of these three parameters, diameters are the least difficult to actually measure.

The most extensively studied diameter in the fetus is the BPD. This is a specific diameter of the fetal skull on the plane through the head located just above the ears, which contains both frontal and occipital bones of the skull. Referring again to Figure 1, the ultrasound image pictured in the top row, second from the right, was taken at the BPD plane.

BPD measurements for an individual fetus are useful in the growth evaluation or age estimation of that fetus only when compared to previous BPD measurements of the same fetus and/or standard growth curves and standard BPD charts. It is critical, therefore, that BPDs be consistently reproducible according to a standard measurement technique. The most difficult aspect of achieving this is reproducibly selecting the anatomical plane on which the measurement will be made.

An anatomic approach is commonly used as the standard method for selecting the BPD plane. In this approach, several intracranial landmarks are identified and used as guides in maneuvering the scan plane into the proper position. First, the relative position of the fetal spine and head is determined. The orbits are then identified, and the scan plane is guided into a lateral axial position at the base of the fetal skull. Once this is done, the scan plane is moved up the head, remaining parallel to the plane at the base of the skull. Various anatomical structures are noted and minor adjustments are made to keep the scan plane parallel until the BPD plane is finally reached, at approximately the level of the lateral ventricles.

A fundamental difficulty is the fact that BPD measurements are extremely sensitive to variations in the selected plane. Inaccuracies as large as one cm have been recorded for measurements taken on planes other than the ideal BPD plane. In order to obtain global reproducibility of these measurements, therefore, one must look toward automating the processes involved. The approach of this paper hopes to overcome the difficulties inherent in the present process by presenting, implementing, and justifying a procedure for automation.

(References for the material presented in this section include [3] and [23].)

4. A probabilistic model for ultrasonic imaging of a fetal head. Our approach uses the general methodology of a *deformable template* (see [9], [10], [11], [12]), which attempts to utilize high-level prior knowledge about an object being imaged by assuming the existence of a known reference object, called a template ($\mathcal{T}$). A space of possible transformations is defined in order to preserve certain characteristic features of the template, and it is assumed that the "true" object being imaged can be obtained by applying a transformation from this space to the template. Hence, the space of "deformed" templates (objects resulting from applying transformations to the template) constitutes the set of all possible "true" objects. An actual image acquired with a sensing device is then a noisy degradation of a "true" image generated from a particular deformed template via some model-specific mapping.

4.1. Construction of the template and space of transformations. The simplest possible formulation of the probabilistic recovery of the fetal head is to imagine the fetal head as being known, except for its location, orientation, and scale—that is, known up to a similarity transformation. Thus, in the framework described above, the template, $\mathcal{T}$, is taken to be a three-dimensional approximation of an "average" fetal head, and all possible locations, scales, and rotations of the fetal head make up the space of transformations, $\overline{\Theta}$. Given a sequence of ultrasound images, an estimate is made from this space of transformations, which, when applied to $\mathcal{T}$, constitutes the recovered fetal head.

4.1.1. The space of transformations, $\overline{\Theta}$. One can view the fetal skull as being contained in a known compact three-dimensional rectangle, say, $B_1 = \prod_{j=1}^{3} [0, L_j]$. If we assume that $B_2 = [L_4, L_5]$ is a known range of scale, then $\overline{\Theta}$ and the image of $\mathcal{T}$ under $\overline{\Theta}$ (call it $\overline{\Theta}(\mathcal{T})$) are given as

$$\overline{\Theta} = B_1 \times B_2 \times SO(3),$$
$$\overline{\Theta}(\mathcal{T}) = \{\theta(\mathcal{T}),\, \theta \in \overline{\Theta}\}.$$

Even this case is not as straightforward as it may seem. Methods can be developed which extract some crude information in the preprocessing stage concerning the location and scale of an object. However, this is not true in the case of an object's orientation. It is the rotation part of the parameter which ordinarily is very difficult to estimate. Systematic searches through the space of possible locations, scales, and rotations, even when crude initial estimates are made concerning location and scale, are extremely time-consuming. Additional difficulties arise due to the fact that biological and man-made objects often have many approximate symmetries which create local maxima for the likelihood. It is necessary, therefore, to search the space of transformations ($\overline{\Theta}$) in a semi-intelligent manner. Since $SO(3)$ is not a linear space, the search must also fully consider the geometry of the manifold $SO(3)$.

4.1.2. The template, $\mathcal{T}$. Two different fetal head templates were constructed. The first is a polyhedral approximation of an "average" fetal head, (V, P), where V contains 162 vertices and P contains 320 index triplets representing the 320 triangular faces of this polyhedron. The edge and face structures are those resulting from refinements of the icosahedral graph structure (see [10]). The fetal head template was constructed from a mold of an authentic fetal skull borrowed from the University of Virginia Medical Center Anatomy Laboratory. Using a Polhemus 3D Digitizer, the locations of 162 points on the surface of the skull model were recorded and taken as the 162 vertices of the polyhedron. The resulting polyhedron is displayed (from four different views) in the first row of Figure 2. As a three-dimensional object, the polyhedral

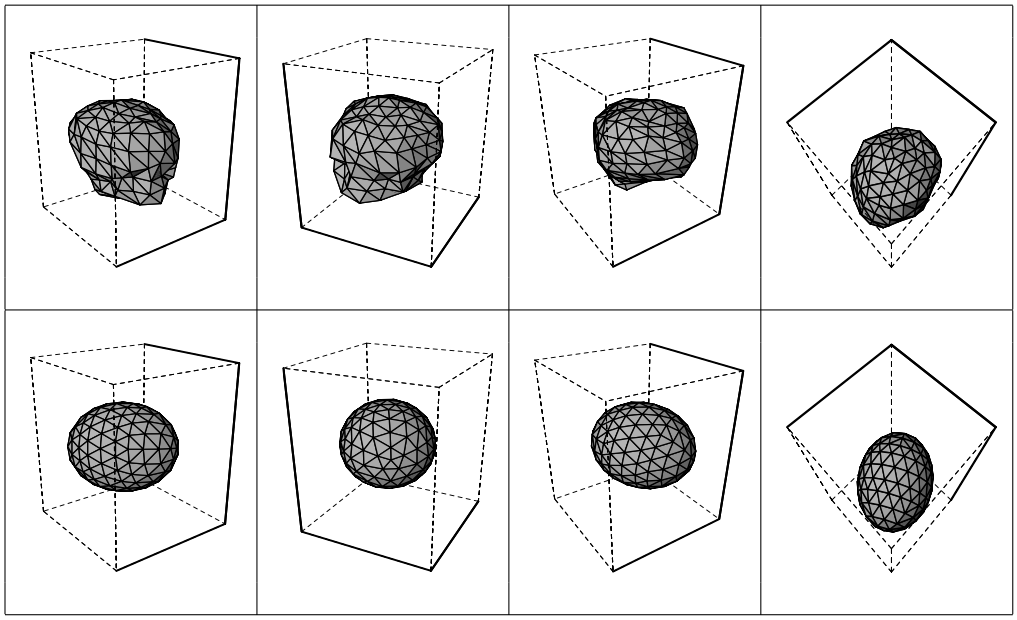


FIG. 2. *Fetal head templates. The polyhedral template is viewed from four different angles in the top row, and the ellipsoid template is viewed from four different angles in the bottom row.*

template provides a surprisingly good approximation of a fetal head. When generating "true" images, however, such a template results in a model-specific mapping that is computationally very complex, since it requires calculating the intersection of the polyhedron with a plane.

The second fetal head template is simply an ellipsoid. Representing the template as the solution to an implicit surface equation (e.g., as an ellipsoid) greatly reduces the level of complexity, and, depending upon the three-dimensional object, may result in a relatively small loss or even a gain in the accuracy of the approximation. The authors plan to use both templates in further investigations; in the present paper, only the ellipsoidal model is used.

The ellipsoid template, $\mathcal{T}$, is defined by

$$\mathcal{T} = \left\{ (x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \right\},$$

where a, b, and c are positive constants which determine the dimensions of the ellipsoid along the x, y, and z axes, respectively. These constants are taken to be a fixed part of the template and hence represent some prior knowledge about the general shape of a fetal head. Their values were estimated beforehand so that the dimensions of several ellipses, obtained by intersecting the ellipsoid with particular planes, most closely matched the corresponding dimensions of a fetal head in several two-dimensional ultrasound slices acquired under controlled conditions. The resulting ellipsoid (displayed from four different views in the second row of Figure 2) is a surprisingly accurate model for at least the top half of a fetal head (the relevant region for this problem). In addition, having an analytic expression for the template allows for quick and easy determination of normals to the surface, as well as whether a particular point lies on, inside, or outside the template, which greatly lowers the

complexity of certain crucial calculations.

4.2. A model for ultrasonic imaging (construction of a likelihood function). Given the ultrasound data images, we wish to estimate the unknown true parameter, $\theta^* \in \overline{\Theta}$ (say, with estimate $\hat{\theta} \in \overline{\Theta}$), and in so doing, indirectly recover the associated deformed template, $\hat{\theta}(\mathcal{T})$, and hence a plausible "true" object. As mentioned in the introduction of this paper, our estimation procedure involves sampling from the posterior measure, π, on $\overline{\Theta}$. In order to construct the posterior distribution, we must first model the ultrasonic imaging technology by defining the "model-specific mapping" which generates "true" images from a given deformed template, $\theta(\mathcal{T})$. Once this is done and a model for the degradation of an image is constructed, a family of density functions can be introduced which are indexed over the space of transformations, $\overline{\Theta}$, and which describe the distribution on the space of images, given a particular transformation, $\theta \in \overline{\Theta}$. The density functions then give rise to a likelihood function on $\overline{\Theta}$, given the data images. The details of the development described above are provided in this section.

4.2.1. The data images. Eight frames from a complete real-time ultrasonic fetal head examination are shown in Figure 1. These eight ultrasound images constitute our collection of data images. They were taken in roughly parallel planes. Approximate distances between the planes were calculated from information available on the video. Hence, it can be assumed that the ultrasound data images represent known planes within the three-dimensional rectangle, B_1.

Let $D^{(k)}$, $k = 1, \ldots, N$, be the N data images. Recall that the images produced are noisy degradations of "true" images, and they are stored as matrices (of size $M \times K$, say) whose components are integer gray scale values ranging from 0 to 255. Hence, for $k = 1, \ldots, N$,

$$D^{(k)} = \{d_{ij}^{(k)}\}_{M \times K},$$

where each $d_{ij}^{(k)}$ is an integer value in $[0, 255]$. We can assume, without loss of generality, that the box B_1 is oriented so that the data images, $D^{(k)}$, occur in parallel xz-planes within the box, specifically located at "slice planes," $y = y_k$, $k = 1, \ldots, N$.

4.2.2. A model for the degradation of an image. We will assume that degradation of an image is modeled by additive Gaussian noise at each pixel location. Given a particular $\theta \in \overline{\Theta}$, one can think of creating a noisy image by adding Gaussian noise (with mean zero) to each pixel value of the "true" image obtained from the template deformed by the transformation θ.

In addition, assume the independence of individual pixel measurements given the "true" image. This is not an unreasonable assumption, and since the "true" image is obtained from a particular transformation via the model-specific mapping, it is equivalent to assuming the independence of pixel measurements given a particular $\theta \in \overline{\Theta}$. Therefore, given the transformation $\theta \in \overline{\Theta}$, for $i = 1, \ldots, M$, $j = 1, \ldots, K$, $k = 1, \ldots, N$,

$$\{X_{ij}^{(k)} = \text{ integer gray level at pixel location } (i, j) \text{ of image taken in plane } y = y_k\}$$

are independent random variables with normal distributions given by

$$X_{ij}^{(k)} \sim N(\mu_{ij}^{(k)}(\theta), \sigma),$$

where the standard deviation, σ, is a fixed constant (estimated from the data images prior to analysis). Keep in mind that the actual acquired ultrasound image $D^{(k)} = \{d_{ij}^{(k)}\}$ is one realization of the matrix of random variables $\{X_{ij}^{(k)}\}$.

4.2.3. The model-specific mapping (modeling the technology). Clearly for $k = 1, \ldots, N$ the matrix of integer values $\{\mu_{ij}^{(k)}(\theta)\}$ constitutes the "true" image obtained from the template deformed by the transformation θ. In other words, it is the "expected" or "mean" image created by the model-specific mapping, given the transformation θ. Define

$$I(i, y_k, j|\theta) = \mu_{ij}^{(k)}(\theta)$$

for $i = 1, \ldots, M$, $j = 1, \ldots, K$, $k = 1, \ldots, N$. Let us refer to $\{I(i, y_k, j|\theta)\}_{M\times K}$ as the "mean image" at the slice plane $y = y_k$, given the transformation θ.

We will now consider the construction of the mean image, $\{I(i, y_k, j|\theta)\}_{M\times K}$. Recall that $\mathcal{T}$ is simply an ellipsoid,

$$\mathcal{T} = \left\{(x, y, z) \in \mathbb{R}^3 : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1\right\},$$

where a, b, and c are positive constants which determine the dimensions of the ellipsoid along the x, y, and z axes, respectively. Let Ψ be defined as

$$\Psi(x, y, z) = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1.$$

For $i = 1, \ldots, M$, $j = 1, \ldots, K$, $k = 1, \ldots, N$, and $\theta = (\bar{X}, Y) \in \overline{\Theta}$ ($\bar{X} = (X_1, X_2, X_3, X_4)$, where $(X_1, X_2, X_3) \in B_1$ is the location, $X_4 \in B_2$ is the scale, and $Y \in SO(3)$ is the rotation), define

$$I(i, y_k, j|\theta) = \beta \cdot \exp\left\{-\alpha \cdot \Psi^2\left(Y^{-1}\left(\frac{1}{X_4}\left((x_i, y_k, z_j) - (X_1, X_2, X_3)\right)\right)\right)\right\},$$

where $x_i = \frac{i}{M} \cdot L_1, z_j = \frac{j}{K} \cdot L_3$ (which splits the xz-plane into a lattice of pixel locations).

The function I, as it is defined above, assigns pixel values to points in the three-dimensional space, B_1. The brightest value, β (typically taken to be 255 = white), is assigned to those points which lie directly on the ellipsoid $\theta(\mathcal{T})$. The brightness values then decrease continuously (as on the normal density curve) as one moves away from $\theta(\mathcal{T})$ in the normal and negative normal directions. This model creates a two-dimensional mean image which is brightest at the intersection of the plane $y = y_k$ with the ellipsoid $\theta(\mathcal{T})$, and which has a slowly dimming band around this intersection giving the template some "thickness." Notice that the constant α controls the bandwidth and the constant β controls the maximum gray level assigned.

Upon actual implementation of this model, it was discovered that having the pixel values decrease all the way to zero (black) was creating images that, when compared to real ultrasound images, were far too dark. Therefore, the model was adjusted so that gray level values decrease down to a fixed value of $\beta_1 > 0$ as one moves inward in the normal direction away from the ellipsoid $\theta(\mathcal{T})$, and down to a fixed value of $\beta_2 > 0$ as one moves outward in the normal direction away from the ellipsoid. The particular values used for β_1 and β_2 are estimated from the actual images before the algorithm is run. The adjusted definition for the function I is given below. Notice

that the change increases the minimum gray level (to β_1 or β_2) while keeping the maximum gray level fixed at β.

$$I(i, y_k, j|\theta) = \begin{cases} (\beta - \beta_2) \cdot \exp\{-\alpha \cdot \Psi^2\} + \beta_2 & \text{if } (x_i, y_k, z_j) \text{ lies outside } \theta(\mathcal{T}), \\ (\beta - \beta_1) \cdot \exp\{-\alpha \cdot \Psi^2\} + \beta_1 & \text{if } (x_i, y_k, z_j) \text{ lies inside/on } \theta(\mathcal{T}). \end{cases}$$

See Figures 3 and 4 for illustrations of the model-specific mapping and degradation model described above. Notice in Figure 4 that neither the model-specific mapping nor the image degradation model attempts to capture any of the artifacts in the interior of the fetal head clearly present in the actual ultrasound images. Some of these artifacts are reflections from intracranial anatomical structures. Those structures that are consistently present and prominent in the typical fetal head during the appropriate age-range could be built into the model-specific mapping. We may attempt to implement such additional structures in the model-specific mapping for future refinements of our algorithm. In the current algorithm, however, we have found that focusing the model on the roughly ellipsoidal and well-formed fetal skull captures information that, because of its large continuous nature, provides strong direction to the gradient-descent algorithm.

4.2.4. The likelihood function. Given the distribution of the family of random variables $\{X_{ij}^{(k)}\}$, we can write down the likelihood function L on the parameter space $\overline{\Theta}$. For each $\theta \in \overline{\Theta}$,

$$L(\theta \mid D^{(1)}, D^{(2)}, \ldots, D^{(N)}) \propto \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{k=1}^{N} \sum_{i=1}^{M} \sum_{j=1}^{K} \left(d_{ij}^{(k)} - I(i, y_k, j|\theta) \right)^2 \right) \right\}$$

describes the likelihood L of the transformation θ, given the sequence of data images $D^{(1)}, D^{(2)}, \ldots, D^{(N)}$. Then the negative log likelihood is given by

$$-\ln\, L(\theta \mid D^{(1)}, D^{(2)}, \ldots, D^{(N)}) = \frac{1}{2\sigma^2} \left(\sum_{k=1}^{N} \sum_{i=1}^{M} \sum_{j=1}^{K} \left(d_{ij}^{(k)} - I(i, y_k, j|\theta) \right)^2 \right) + \text{ const.}$$

5. The algorithm.

5.1. Procedure: Generating realizations from the posterior distribution. A prior distribution on $\overline{\Theta}$ may also be constructed, incorporating additional prior knowledge about the fetal head. Given $p(\theta)$, a prior density on $\overline{\Theta}$, and the likelihood function defined above, Bayes' formula yields a posterior density function q on $\overline{\Theta}$, given the sequence of data images $D^{(1)}, D^{(2)}, \ldots, D^{(N)}$:

$$q(\theta|D^{(1)}, D^{(2)}, \ldots, D^{(N)}) \propto L(\theta|D^{(1)}, D^{(2)}, \ldots, D^{(N)}) \times p(\theta).$$

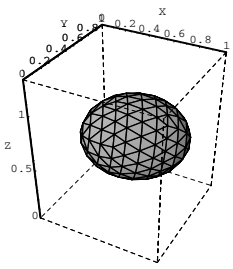
We may then define the "objective function" $h(\cdot)$ via the following relationship:

$$e^{-h(\theta)} = q(\theta) \qquad \text{or} \qquad h(\theta) = -\ln q(\theta).$$

In practice, because of the immense amount of information in the likelihood, we have used a uniform prior on $\overline{\Theta}$.

Recall that sampling from the posterior distribution will be achieved by running the manifold version of "$d\,\theta_t = -\frac{1}{2}\,\nabla h(\theta_t)dt + dW_t$." The resulting diffusion process on the manifold $\overline{\Theta}$ is generated by the operator

$$A = \frac{1}{2}(\,\Delta - \nabla h)$$



Ellipsoid template, $\mathcal{T}$.

$\mathcal{T} = \left\{(x, y, z) \in \mathbb{R}^3 : \ \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1\right\}$, where a, b, and c are positive constants.

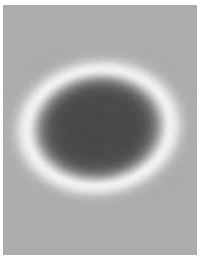
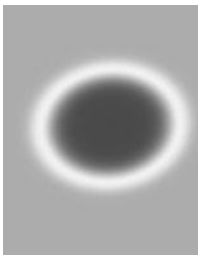
Slice plane y_1: $y = .39$	**Slice plane y_2:** $y = .50$	**Slice plane y_3:** $y = .61$	**Slice plane y_4:** $y = .77$
			
			

FIG. 3. *Ellipsoid template and the model-specific mapping. The constants a, b, and c in the ellipsoid template model are taken to be a fixed part of the template and hence represent some prior knowledge about the general shape of a fetal head. The actual values used in the algorithm,* $a = .73$, $b = .68$, $c = .86$, *were estimated beforehand so that the resulting "mean images," created by taking* $y = y_k$ *slices of the template via a model-specific mapping, most closely matched the corresponding true ultrasound images acquired under controlled conditions. The model-specific mapping creates "mean images" by modeling the ultrasound imaging technology (these images are shown in row* 1 *of the above table). The corresponding actual ultrasound images are shown in row* 2. *The model-specific mapping creates a mean image which is brightest at the intersection of the plane* $y = y_k$ *with the deformed ellipsoid template and which has a slowly dimming band around the intersection giving the template some "thickness." There are parameters which control the width of this band and the maximum gray level assigned. The gray level values decrease down to a fixed value of* $\beta_1 > 0$ *as one moves inward in the normal direction, and down to a fixed value of* $\beta_2 > 0$ *as one moves outward in the normal direction away from the ellipsoid. The particular values used for* β_1 *and* β_2 *are estimated from the actual ultrasound images before the algorithm is run.*

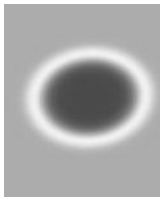
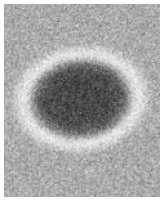
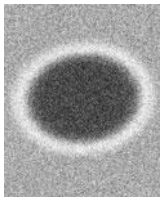
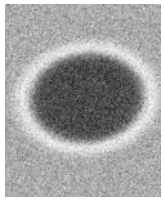
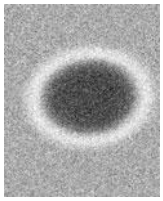
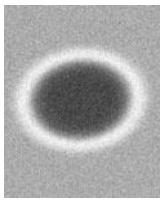
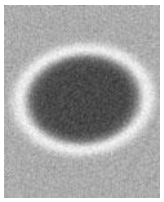
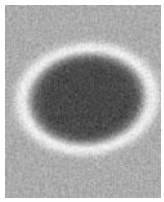
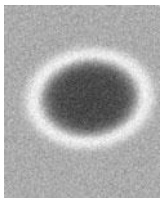
	Slice plane: $y_1 = .39$	**Slice plane:** $y_2 = .50$	**Slice plane:** $y_3 = .61$	**Slice plane:** $y_4 = .77$
Mean image				
Noisy mean image $\sigma = 50.0$				
Noisy mean image $\sigma = 25.35$				
True ultrasound image				

FIG. 4. *A model for the degradation of an image. We model the degradation of an image by additive Gaussian noise at each pixel location. Given a particular transformation, one can think of creating a noisy image by adding Gaussian noise (with mean zero) to each pixel value of the "mean image" obtained from the template deformed by the transformation. Such a process leads to the following construction: Given a transformation* θ*, the family of random variables* $\{X_{ij}^{(k)}\}$*, giving the gray level values for each pixel location* (i, j) *in slice plane* $y = y_k$*, are independent and normally distributed:* $X_{ij}^{(k)} \sim N(\mu_{ij}^{(k)}(\theta), \sigma)$*, where the matrix of integer values* $\{\mu_{ij}^{(k)}(\theta)\}$ *constitutes the mean image obtained from the template deformed by* θ*, and* σ *is the image degradation standard deviation.*

and boundary operator

$$L = \frac{\partial}{\partial n} \qquad \text{(derivative in the normal direction)}.$$

The function h is a real-valued function on $\overline{\Theta} = B_1 \times B_2 \times SO(3)$. At algorithmic time t, denote the value of $\theta(\in \overline{\Theta})$ as θ_t. Let $X_t \in B_1 \times B_2$ denote the location and scale part of θ_t and $Y_t \in SO(3)$ the rotation part. Then $\theta_t = (X_t, Y_t)$. Let $h_Y(X)$ represent the function $h(\theta)$ on $B_1 \times B_2$ with $Y \in SO(3)$ fixed and X varying over $B_1 \times B_2$. Similarly, let $h_X(Y)$ represent the function $h(\theta)$ on $SO(3)$ with $X \in B_1 \times B_2$ fixed and Y varying over $SO(3)$.

Let $so(3)$ denote the Lie algebra of skew-symmetric matrices and J the canonical isomorphism $J: \mathbb{R}^3 \to so(3)$ given by

$$J(x_1, x_2, x_3) = \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}.$$

For $A \in so(3)$, let $\exp(A)$ be defined as

$$\exp(A) = \sum_{n=0}^{\infty} \frac{A^n}{n!}.$$

Also, for $Y \in SO(3)$, let L_Y denote left translation by Y. Then the function $h_X : SO(3) \to \mathbb{R}^3$ can locally (near $Y \in SO(3)$) be viewed as the following composition:

$$\begin{array}{ccccccccc} \mathbb{R}^3 & \xrightarrow{J} & so(3) & \xrightarrow{\exp} & SO(3) & \xrightarrow{L_Y} & SO(3) & \xrightarrow{h_X} & \mathbb{R}. \\ \text{Near} & & \text{Near} & & \text{Near} & & \text{Near} & & \\ \text{zero} & & \text{zero} & & \text{identity} & & Y & & \end{array}$$

The gradient of h_X at a point $Y \in SO(3)$ is obtained by analytically calculating the (exact) gradient of the mapping diagrammed above, which then gets "pushed" across by the differential maps of J and exp.

Let $\delta t_i = t_i - t_{i-1}$ be the time step size and $N_p(0,1)$ be a realization from a p-variate standard normal distribution. The *algorithm* on the manifold $\overline{\Theta} = B_1 \times B_2 \times SO(3)$ is given by

$$\theta_{t_{i+1}} = (X_{t_{i+1}}, Y_{t_{i+1}}), \tag{5.1}$$

$$X_{t_{i+1}} = X_{t_i} - \frac{1}{2}(\delta t_i)\nabla h_{Y_{t_i}}(X_{t_i}) + \sqrt{\delta t_i} N_4(0,1), \tag{5.2}$$

$$Y_{t_{i+1}} = Y_{t_i} \times \exp\left[J\left\{ -\frac{1}{2}(\delta t_i)\nabla h_{X_{t_i}}(Y_{t_i}) \right\} + J\{\sqrt{\delta t_i} N_3(0,1)\} \right], \tag{5.3}$$

with reflection at the boundary of $\overline{\Theta}$ in the normal direction.

This algorithm is a discretization of the time-homogeneous diffusion process θ_t on the manifold $\overline{\Theta} = B_1 \times B_2 \times SO(3)$. The construction of a discrete approximation to Brownian motion on $SO(3)$ was considered by McKean [26]. Our algorithm is a version of that one with $-\frac{1}{2}\nabla h$ added as drift. For the $\{Y_t, t \geq 0\}$ part of θ_t, the above algorithm is the product injection formula for a stochastic flow on $SO(3)$ (see [26], [27]), except that now we have as the drift term a vector field which is not invariant, which makes a big difference.

For large t, θ_t is taken as a realization from the posterior distribution, π, on $\overline{\Theta}$, given the actual ultrasound images. This realization represents our estimate $\hat{\theta}$ of the "true" location, scale, and rotation of the fetal head. Applying the estimated transformation $\hat{\theta}$ to the ellipsoid template $\mathcal{T}$ yields an estimated (recovered) fetal head, $\hat{\theta}(\mathcal{T})$. In section 8, general theorems are proven which justify this methodology. Running the algorithm to obtain a sample from the distribution, π, is justified by proving that the time-homogeneous diffusion process converges weakly to the distribution π.

Figures 5 and 6 illustrate the time evolution of a single realization. The algorithm is run for 2000 steps in this example. In Figure 5, four of these steps are shown. In Figure 6, all 2000 steps of the same run are pictured via individual graphs of the time evolution of location, scale, and rotation. The final value, θ_{2000}, represents a single realization from the distribution π on $\overline{\Theta}$. Loosely speaking, the fetal head template will deform, trying to "best fit" the spatial sequence of ultrasound images while adhering to those constraints which define what are allowable deformations ($\overline{\Theta}$). Such a method can be expected to be broadly applicable to the general realm of object recognition in image processing and computer vision.

Note that different estimation procedures can be applied using the generated realizations from the posterior distribution π. It is clear from the graphs in Figure 6 that, even when the diffusion process converges quickly, choosing a specific large value of t for which $\hat{\theta} = \theta_t$ results in an estimated transformation that varies randomly (approximately according to the measure π). These estimates may vary a little (if the data is very strong) or they may vary a great deal. In either case, to account for the random variation, an "average" may be calculated from a simulated sample consisting of a finite number of generated realizations. However, how one calculates an average in the case of the rotation parameter is not clear. In our particular application, it would be possible (and appropriate) to recover an "average" fetal head by simply applying the average scale value obtained from the simulated sample to the fetal head template.

5.2. More general deformations. The above algorithm should move the template skull into the right location, overall size (scale), and orientation; this is really just the first stage. Local deformations would then be applied so that different regions of the template would be allowed to deform differently but still with certain local and/or global characteristics being preserved. For example, in the fetal skull context, the detection of Spina Bifida, where the cross section is *lemon shaped*, would require a deformation of the fetal skull template by means other than a similarity transformation. In [11] such local deformations, similar to elasticity theory, were proposed and appear to be applicable to this problem.

Recently, the general idea of transitions between global and local deformations has been made more rigorous and systematically implementable with the concept of the "group cascade." Its goal was not to speed up convergence, but rather to develop a framework, broadly applicable in biomedical settings, in which one could obtain a realistic representation of inherent biological variability [25].

6. Application of the algorithm: Reconstructing a fetal skull and calculating the BPD.

6.1. Analysis. In the general methodology of *deformable templates*, if one's aim is to gain some kind of structural understanding of an acquired two-dimensional image, then a "true" image can be generated from the (indirectly) estimated "true" object via the model-specific mapping. This particular "true" image is then taken as the noiseless

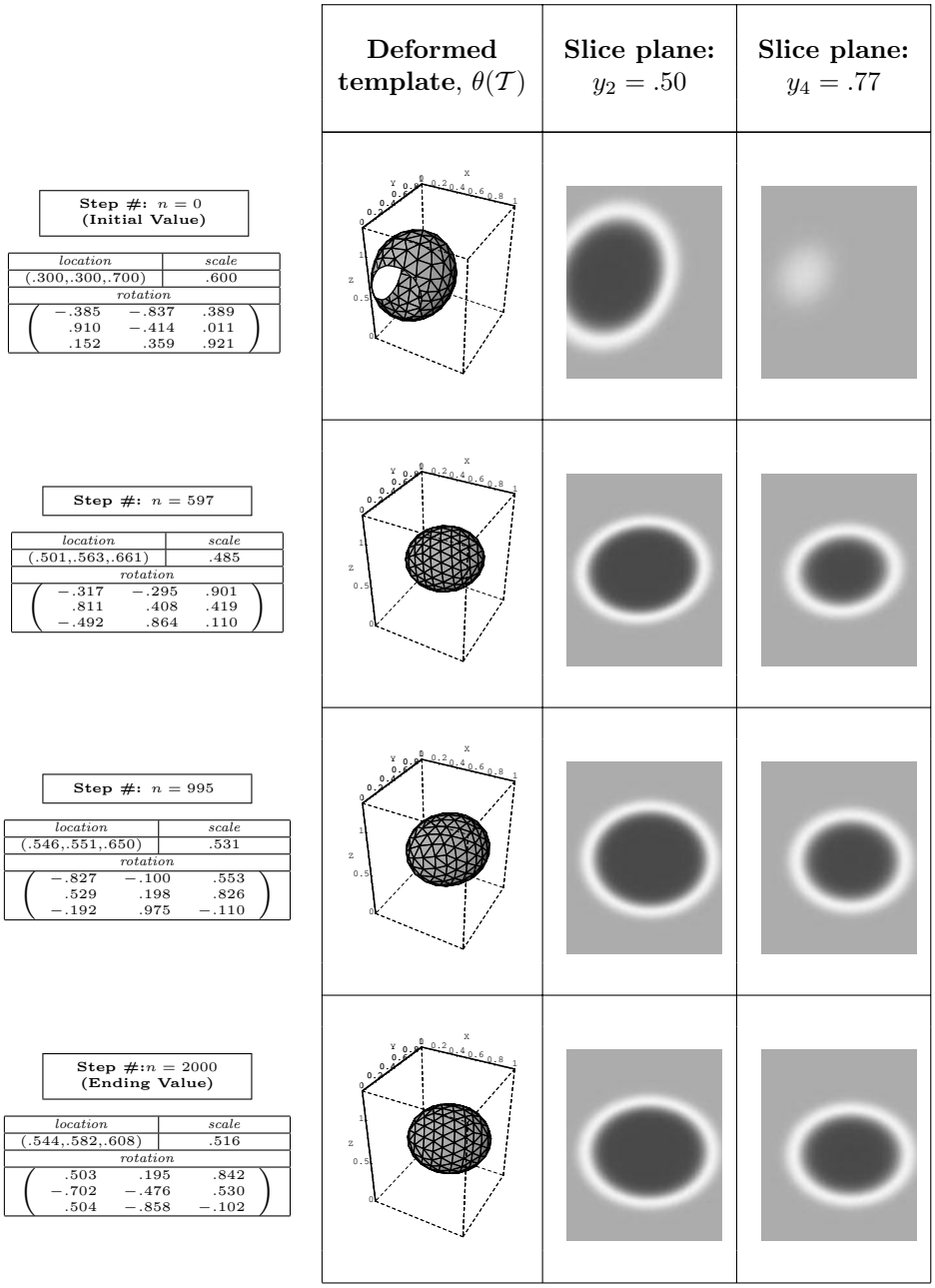


FIG. 5. *Time evolution of one realization. The figures shown above illustrate the value of the process at four distinct time steps within a single simulation. Column* 1 *shows the three-dimensional ellipsoid template transformed by the appropriate value of* θ. *Columns* 2 *and* 3 *show the intersection of the deformed template with the two slice planes,* y_2 *and* y_4, *thus creating "mean images," which are compared to the corresponding true ultrasound images in the algorithm. This particular simulation consists of* 2000 *steps of size* $\delta t = 1.0 \times 10^{-7}$.

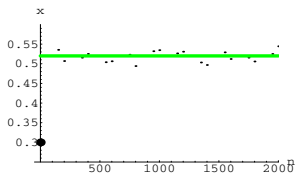
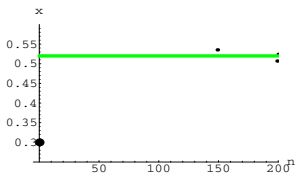
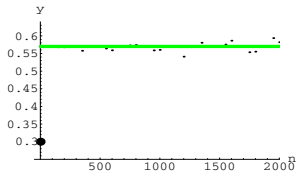
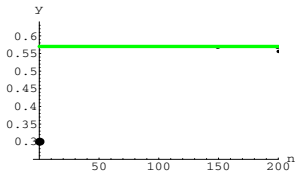
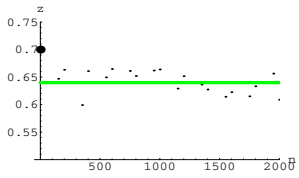
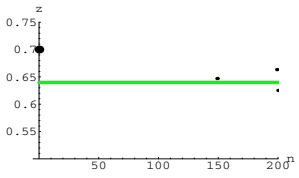
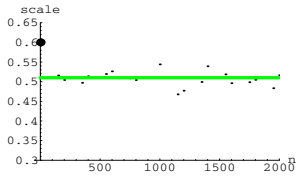
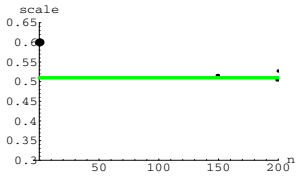
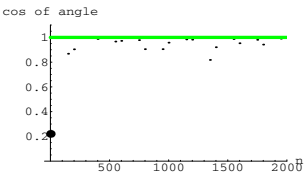
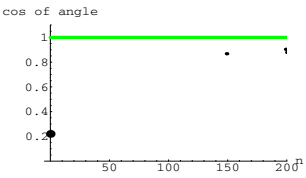
Transformation space coordinate	**Full evolution** (2000 steps)	**Initial segment** (First 200 steps)
Spatial location: x-coordinate		
Spatial location: y-coordinate		
Spatial location: z-coordinate		
Scale coordinate		
Rotation parameter		

FIG. 6. *Time evolution of one realization—a closer look. The figures shown above illustrate in greater detail the time evolution of the same realization as that pictured in Figure* 5. *Each row contains two graphs of the time evolution of a single coordinate of the parameter space. The second column contains a full plot of each parameter coordinate as a function of the iteration step. The third column contains just the first* 200 *steps of the process, providing a magnified view of the convergence. Each of these graphs also displays, as a reference, a line indicating the true value of the parameter coordinate. The three spatial location coordinates and the scale coordinate are displayed in the first four rows. In the last row, the evolution of the rotation parameter is illustrated via a single function giving the cosine of the angle between the major axes of the ellipsoids obtained by applying the true rotation and by applying the current value of the rotation parameter.*

restored image. One distinct advantage of this type of methodology is that, since the restored image is generated from a deformed template, it contains all structural information contained in the template (see [1], [9], [28]). For example, landmarks defined on the template are mapped (via the composition of a transformation and the model-specific mapping) into corresponding landmarks in the restored image and can therefore be automatically located, or associated geometric measurements can be automatically made.

In the present case, however, it is the structure of the three-dimensional fetal head itself which is of interest and *not* the structure of the two-dimensional image. Recall that an eventual objective is to automatically measure the BPD of the fetal head, which is a two-dimensional diameter measurement, but since the data images are taken in predefined planes, the appropriate ultrasound slice for measuring the BPD (at the level of the BPD plane) is not necessarily included among the acquired data images. Therefore, restoring the data images is of no use in this problem since these particular images do not necessarily contain the desired information.

The information must come from the structure of the recovered "true" fetal head. As in the image restoration case described above, when using this methodology we have access to structural information in the recovered fetal head because it is generated from the original fetal head template. This would not necessarily be the case if we were, for instance, "reconstructing" the fetal head from acquired two-dimensional ultrasound slices by, in a sense, filling in the gaps. The BPD plane can be defined on the fetal head template and can be mapped (via a transformation alone in this case) to the corresponding BPD plane on any deformed template. Thus, we stop short of generating a "true" image and instead take the BPD measurement on the indirectly estimated deformed three-dimensional fetal head template.

Notice that the transformations in the current model include only locating, scaling, and rotating. Of these, only scaling affects the structure of the template. Since structure is what we are interested in, it might seem that estimation of location and rotation is completely unnecessary. However, it becomes clear that the scale of the template cannot be accurately recovered without also simultaneously recovering the location and rotation.

6.2. Results.

6.2.1. A normalized parameter space. In practice, because ultrasound imaging is typically done in dimensions that are not specifically calculated, a transformation of units, ξ, is necessary to take us from the original actual parameter space, $\overline{\Theta} = B_1 \times B_2 \times SO(3)$, to a normalized parameter space, $\overline{\Theta}_0 = B_1^0 \times B_2^0 \times SO(3)$, where $B_1^0 = [0,1]^3$ and $B_2^0 = [0,1]$. Since the original space $\overline{\Theta}$ is not completely determined in practice, in order to actually implement a discretization of the time-homogeneous diffusion process on $\overline{\Theta}$, we must run our algorithm on the normalized space, $\overline{\Theta}_0$, via a transformation of units.

Consider the linear transformation, $\xi : \overline{\Theta} \to \overline{\Theta}_0$. Recall that q is the posterior density function on the original space, $\overline{\Theta}$. The posterior density on the normalized space is given by

$$q_0(u) = q(\xi^{-1}(u)) \cdot |J(u)| \quad \text{for } u \in \overline{\Theta}_0,$$

where $|J(u)|$ is the Jacobian of the transformation ξ.

If we define $h_0(u) = -\ln q_0(u) = h(\xi^{-1}(u)) + \text{const}$ for $u \in \overline{\Theta}_0$, then

$$\nabla_u h_0(u) = \nabla_u(h(\xi^{-1}(u))) = \nabla_x h(\xi^{-1}(u)) \cdot \frac{\partial \xi^{-1}}{\partial u_i}(u) = \tau \cdot \nabla_x h(\xi^{-1}(u))$$

for a constant τ, assuming for computational reasons that $\frac{\partial \xi^{-1}}{\partial u_i}(u) = \tau \cdot I$, where I is the identity matrix. This simplification involves assuming that the transformation of units, ξ, scales identically in all parameters (including $SO(3)$).

We can now write the diffusion on $\overline{\Theta}_0$ as

$$dZ_t = -\frac{1}{2}\nabla h_0(Z_t)dt + dW_t,$$

where $\{W_t, t \geq 0\}$ is a standard Wiener process on $\overline{\Theta}_0$. Equivalently,

$$dZ_t = -\frac{1}{2}\tau \nabla_x h(\xi^{-1}(Z_t))dt + dW_t.$$

We see that when our diffusion is run on the normalized parameter space, we must multiply the gradient piece of the algorithm by the unknown constant τ.

6.2.2. Running the algorithm. To test the algorithm itself, independent of our model for ultrasonic imaging, a sample location, scale, and rotation value (say, $\theta^* \in \overline{\Theta}$) were chosen, and noisy simulated data images were created from the deformed template $\theta^*(\mathcal{T})$ via our model-specific mapping and degradation model. The algorithm was then run using these simulated images as data, rather than the actual acquired ultrasound images. Various algorithm parameters were adjusted for each of these test runs on simulated data, and the runs were repeated until the realizations produced were fairly well concentrated around the chosen test value, θ^*.

One parameter of particular interest is the constant discussed above, τ. This constant, a result of a transformation of units, is multiplied by the gradient piece of the algorithm. Without this constant multiple, our Brownian motion and gradient terms would be on different scales. The constant τ, therefore, has the effect of balancing the deterministic and stochastic parts of the algorithm.

Since typically the constant τ cannot be calculated, it must be estimated via the process discussed in the first paragraph above, in which partial test runs of the algorithm are observed, and then the value of the constant is adjusted until the realizations produced are fairly well concentrated around the chosen test value. Recall that we made estimates for σ, the standard deviation of the image degradation model, by analyzing the data images prior to running the algorithm. If those estimates are inaccurate, there will be a correction factor, which will appear as a constant multiple in front of the gradient. Therefore, our constant τ may also potentially incorporate any correction necessary for our estimate of σ.

In Figure 7, some realizations run on simulated data are pictured which illustrate the importance of the parameter τ. For example, if τ is too small, then the Brownian motion is overemphasized in the process. The balancing effect of τ is evident when one views runs of the algorithm at various values of the constant (Figure 7). In Figure 8, the same runs are pictured via individual graphs of the time evolution of location, scale, and rotation. This provides a more detailed look at the effects of τ on convergence. Notice the varying effects among each coordinate of the parameter space. For example, the time evolution of the convergence in the x-coordinate of the location parameter may look quite different than that of the scale parameter. This

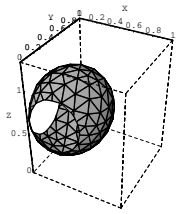
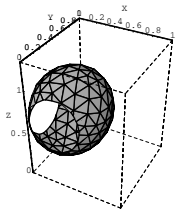
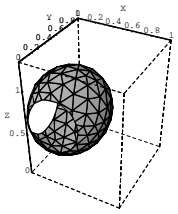
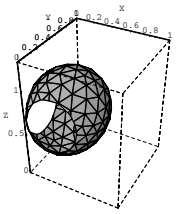
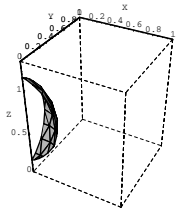
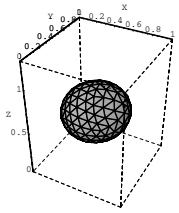
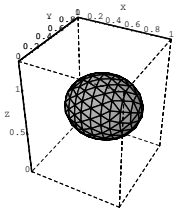
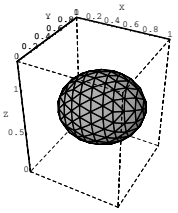
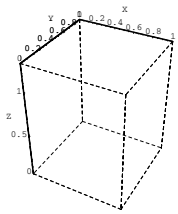
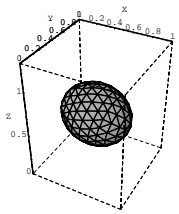
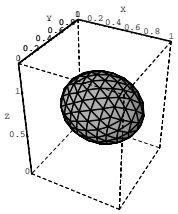
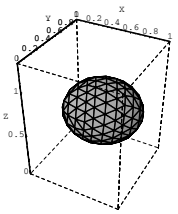
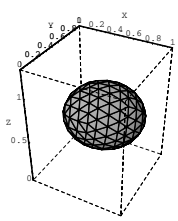
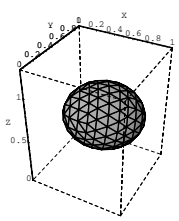
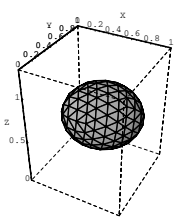
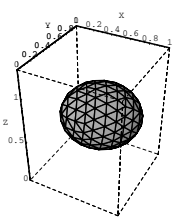
Step #	$\tau = .00005$	$\tau = .01$	$\tau = .0005$	$\tau = .001$
Initial value: $n = 0$				
$n = 297$				
Ending value: $n = 1000$				
Sample test value: θ^*				

FIG. 7. *The effects of varying the parameter τ. The figures shown above illustrate four different realizations running (on simulated data) at four different values of the parameter τ. Each column shows the ellipsoid template, $\mathcal{T}$, transformed by the nth value of θ_t in that realization (for three n values), and compares the ending ellipsoid (shown in row 3) to the sample "true" ellipsoid, $\theta^*(\mathcal{T})$ (shown in row 4). The realization pictured in column 1 has a very small τ value, overemphasizing the Brownian motion, which causes the template to jump out of the viewing box. The realization pictured in column 2 has a very large τ value. With the Brownian motion underemphasized, this realization falls into a local minimum and is unable to escape. The τ values used in the realizations pictured in columns 3 and 4 perform better. Column 3 illustrates a realization which jumps out of a local minimum that it falls into early on but then continues to jump around the test value θ^*. Column 4 illustrates a realization which converges nicely to a θ value which is very close to the test value θ^*.*

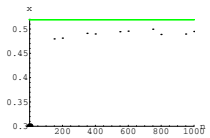
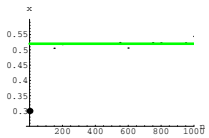
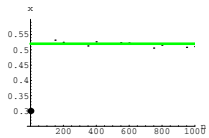
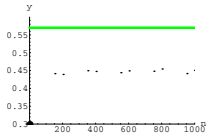
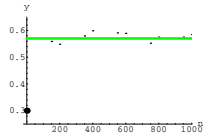
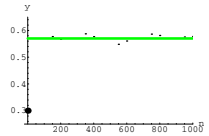
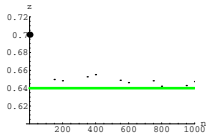
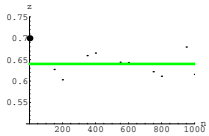
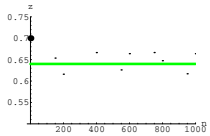
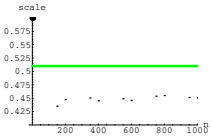
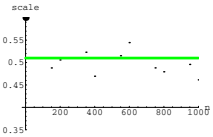
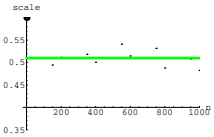
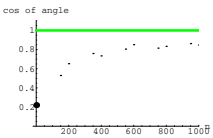
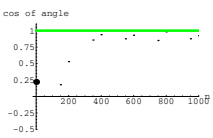
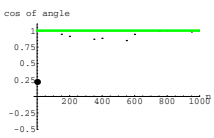
Transformation space coordinate	**Parameter value** $\tau = .01$	**Parameter value** $\tau = .0005$	**Parameter value** $\tau = .001$
Location: x-coordinate			
Location: y-coordinate			
Location: z-coordinate			
Scale coordinate			
Rotation parameter			

FIG. 8. *The effects of varying the parameter—a closer look. The figures shown above illustrate in greater detail the time evolution of three of the four runs pictured in Figure* 7. *Each row contains a plot of a single coordinate of the parameter space as a function of iteration step, and each column corresponds to the different values of the constant* τ *used. The full run in each case pictured above is* 1000 *steps. No further information was provided by these more detailed graphs in the case of* $\tau = .00005$ *from Figure* 7, *so it was not included in this figure. Note, however, the following details illuminated by the graphs provided for the other three* τ *values: For* $\tau = .01$, *one can clearly see convergence into a local minimum in each component of the paramter space. Comparing the effects of* $\tau = .0005$ *and* $\tau = .001$, *one can see (in each of the plots) slightly more scatter about the "true" value for* $\tau = .0005$.

is surely a result of using a single constant multiple for the constant τ rather than a matrix of values, a simplifying assumption made in the previous subsection. A single value cannot balance all dimensions of the parameter space equally well. However, making this practical approximation does not seem to significantly affect the overall success of the algorithm but merely alters the relative speed of convergence in the five distinct dimensions of the parameter space.

To test our model for ultrasonic imaging, the algorithm was run using the actual acquired ultrasound images as data. After adjusting various parameters in the test cases mentioned above, further refinement was necessary during the first runs on real ultrasound data since the change in the data images (from simulated data to real ultrasound data) caused changes in the shape of the likelihood function. Unlike the runs on simulated data, it was difficult to determine the degree of success for the runs on real ultrasound data because the true value of $\theta \in \overline{\Theta}$ is, of course, unknown. Therefore, in order to assess a particular estimate $\hat{\theta}$ (a realization from the posterior distribution π on $\overline{\Theta}$), noiseless "restored" images were generated from the deformed template $\hat{\theta}(\mathcal{T})$ via the model-specific mapping and were compared with the corresponding ultrasound images. If the slices generated from the estimated fetal head ($\hat{\theta}(\mathcal{T})$) match the corresponding ultrasound images fairly well, then one would assume that a slice at the BPD level would also well approximate the actual fetal head at the two-dimensional BPD plane. This would, in turn, indicate that the BPD measurement taken on the deformed template is a good estimate for the actual BPD.

Overall, the results of the algorithm appear quite promising. Again, this is merely a first step in our stated goal (see the introduction)—that is, to automate the process by which a technician, in a prenatal ultrasonic examination of the fetal skull, utilizes his or her intuitive understanding of what is bone, what is tissue, and what is sensor noise in assessing the developing fetal skull.

7. Summary. A general formulation was presented by which one could incorporate a priori information into modeling, where the parameterization of the models is described by a compact manifold with boundary. The formulation was driven by the specific problem of reconstructing a fetal head from a sequence of ultrasound frames obtained during a prenatal exam. An important question is whether one can automate the intuitive understanding of the technician/radiologist as to what is bone, what is tissue, and what is sensor noise. In the present application, one issue is whether ultrasound imaging could be performed prior to the development of the skull. To do this, one must use all available knowledge as to what is bone, what is tissue, and what is sensor noise. Moreover, with the advent of medical imaging being performed remotely (telemedicine), the issue of what can and cannot be reasonably and properly automated in medical imaging is becoming more and more important.

The approach of the present paper is that of a probabilistic recovery of the fetal head. The "space of all possible fetal heads" is viewed as the result of similarity (translation, scale, rotation) transformations being applied to a given fetal head template. A model of the ultrasonic imaging of a fetal head is formulated and prior knowledge is incorporated, resulting in an a posteriori probability measure. A method for sampling from this measure is then constructed and justified. The complication is that the measure is on a Riemannian, compact, simply connected, oriented manifold with boundary. A discretization of a time-homogeneous diffusion process was implemented into code in the particular case of interest, that of $\overline{\Theta} = B_1 \times B_2 \times SO(3)$. The algorithm was then applied to simulated images, as well as to actual ultrasound prenatal images.

Two particular future applications concern robotics and protein structure. Consider "active" robotics, where the movement of a robot arm is not preprogrammed but acts dynamically within a changing environment. A representation of the arm is then provided by the "product" of the location and orientation of each of the N links of the robot arm. Moreover, each link of a robot arm is constrained not only by the fact that the links are connected end-to-end but, more importantly, that one link cannot pass through another. The result is a configuration space which is a submanifold with boundary of a finite product of such Lie groups.

Second, consider the computational biological approach to protein structure determination. For a variety of reasons, tertiary and quarternary structures have been determined for only a relatively small number of proteins. An alternative approach is to infer structural characteristics by comparison to potential evolutionary relatives whose three-dimensional structures are known. Many monomer proteins are built up from motifs (α-helices, β-strands) which are connected by links whose structure is very flexible. Similarly, the subunits of multimer proteins are usually held together by weak interactions, with flexibility in their realizable relative orientations. In the alternative approach, a potential energy is defined and a distance is calculated between the $C_\alpha - C_\alpha$ backbone of the protein of interest and that of a candidate of known three-dimensional structure from within a protein database. This distance is taken as a measure of structural dissimilarity. One aspect of such programs as Dali and FSSP [15], [16] is to find a candidate from the database and an approximate rigid motion of the configuration which minimizes the distance between the two. Ideally, one would like to have an algorithm which would allow for both rotations and scale differences and which would treat each submotif or subunit separately, with the allowable combinations of transformations being somewhat constrained. The present methodology may provide some basis for such an approach.

8. Stochastic relaxation on a manifold with boundary. We assume that $\overline{\Theta}$ is a compact regular domain (see [2]), in the sense that there exists an n-dimensional, orientable manifold N of class C^∞, with Θ being a d-dimensional domain in N, whose closure $\overline{\Theta}$ is compact and whose boundary $\partial\Theta = \overline{\Theta} - \Theta$ consists of a finite number of hypersurfaces, each of dimension $d-1$ and of class C^3. Let $\overline{\Theta}$ be a compact, connected, oriented, smooth, d-dimensional, Riemannian manifold with boundary and with metric g.

Let $C^\infty(\overline{\Theta})$ denote the class of infinitely differentiable functions from $\overline{\Theta}$ to $\mathbb{R}$. Let Ω be the natural volume element of the oriented Riemannian manifold $\overline{\Theta}$, and let ν be the Borel measure associated with Ω. Integration of functions on $\overline{\Theta}$ (which is integration of d-forms on $\overline{\Theta}$) can be viewed as Lebesgue–Stieltjes integration on $\overline{\Theta}$ with respect to ν:

$$\int_{\overline{\Theta}} f = \int_{\overline{\Theta}} f\ \Omega = \int_{\overline{\Theta}} f(\theta)\ \nu(d\theta) \qquad \text{for all } f \in C(\overline{\Theta}).$$

Let π be a Borel probability measure on $\overline{\Theta}$ which has a strictly positive density function, $q(x) \in C^\infty(\overline{\Theta})$, with respect to ν, and define $h(x) = -\ln q(x)$. Let

$$A = \frac{1}{2}(\,\Delta - \nabla h)$$

with domain $\mathcal{D}(A) = C^2(\overline{\Theta})$ and where ∇h is the gradient of the function h. What is to be shown is that the algorithm is sampling from the measure $q = e^{-h}$. In the

situation described in section 6.2.1, where the normalized parameter space $\overline{\Theta}_0$ is used and a change of variables is involved, all the above conditions are satisfied, where h is replaced by h_0 (properly balancing the gradient and stochastic terms).

Consider the Cauchy problem for the following parabolic equation: For a given $f \in C(\overline{\Theta})$, find u such that

$$\frac{\partial}{\partial t}u(t,x) = A\,u(t,x), \quad t > 0, x \in \Theta; \tag{8.1}$$

boundary condition (reflecting barrier):

$$Lu(t,x) = \frac{\partial}{\partial n}u(t,x) = 0, \quad t > 0, x \in \partial\Theta; \tag{8.2}$$

initial condition:

$$\lim_{t \downarrow 0} u(t,x) = f(x) \ \text{ (uniformly in) } x \in \overline{\Theta}. \tag{8.3}$$

PROPOSITION 1 (see Ito [18], [19], [20]; Sato and Ueno [31]). *The fundamental solution for the above Cauchy problem, constructed by Ito,*

$$p : (0,\infty) \times \overline{\Theta} \times \overline{\Theta} \to \mathbf{R},$$

satisfies the following properties:

1. *For a fixed $(t,x) \in (0,\infty) \times \overline{\Theta}$, $p(t,x,y)$ is continuous in y;*
2. *for any $f \in C(\overline{\Theta})$,*

$$u(t,x) = \int_{\overline{\Theta}} p(t,x,y) f(y)\, d\nu(y) \tag{8.4}$$

is continuous on $(0,\infty) \times \overline{\Theta}$ and continuously differentiable in $t \in (0,\infty)$ and $C^2(\overline{\Theta})$ for a fixed $t \in (0,\infty)$;

3. *$p(t,x,y)$ is strictly positive (in our case of reflection at the boundary) and satisfies*

$$p(t+s,x,y) = \int_{\overline{\Theta}} p(t,x,z) p(s,z,y)\, d\nu(y);$$

4.

$$\int_{\overline{\Theta}} p(t,x,y) d\nu(y) = 1. \tag{8.5}$$

Proof. For the proof, see Ito [18], [19], [20] and Sato and Ueno [31]. □

Note. Under boundary conditions other than reflection at the boundary, strict positivity of the density is not ensured, in that the process can then spend "local time" on the boundary (see [18, section 9]).

Let $\overline{A}$ denote the smallest closed extension in $C(\overline{\Theta})$ of A and denote its domain by $\mathcal{D}(\overline{A})$. To apply the Hille–Yosida theory in the present case of (A, L), one needs to construct the semigroup on $C(\overline{\Theta})$ whose Green's operators $\{G_\alpha\}$ satisfy $(\alpha - \overline{A})G_\alpha u = u$ and $LG_\alpha u(x) = 0$ for $x \in \partial\Theta$. In [31], Sato and Ueno present such a construction. For $f \in C(\overline{\Theta})$ and $\alpha \geq 0$, from the equation $(\alpha - \overline{A})u = f$ and $u = 0$ on $\partial\Theta$, one obtains the minimal resolvent for this translated Poisson equation, $\{G_\alpha^{min}\}$. Similarly,

for $\phi \in C(\partial\Theta)$ and $\alpha \geq 0$, from the equation $(\alpha - \overline{A})u = 0$ and $u = \phi$ on $\partial\Theta$, one obtains the resolvent for this translated Dirichlet problem, $\{H_\alpha\}$. Let the extensions of LG_α^{min} and LH_α be denoted by $\overline{LG_\alpha^{min}}$ and $\overline{LH_\alpha}$. Denote the domain of $\overline{LH_\alpha}$, which does not depend on α, by $\tilde{\mathcal{D}}$. Let $\hat{L}$ be the unique operator for which $\hat{L} = L$ on $C^{2,\kappa}(\overline{\Theta})$, $\hat{L}G_\alpha^{min} f = \overline{LG_\alpha^{min}} f$ for $f \in C(\overline{\Theta})$, and $\hat{L}H_\alpha\phi = \overline{LH_\alpha}\phi$ for $\phi \in \tilde{\mathcal{D}}$. Let the domain $\mathcal{D}(\hat{L})$ (which $\supset C^{2,\kappa}(\overline{\Theta})$) be given by all functions of the form $\sum_{i=1}^n G_{\alpha_i}^{min} f_i + H_{\beta_i}\phi_i$, $f_i \in C(\overline{\Theta})$, $\phi_i \in \tilde{\mathcal{D}}$, $\alpha_i \geq 0, \beta_i \geq 0, 1 \leq i \leq n$. Finally, let $\overline{A}_{\hat{L}}$ be the restriction of $\overline{A}$ to the subset $\mathcal{D}(\overline{A}_{\hat{L}}) = \{u | u \in \mathcal{D}(\hat{L})$ and $Lu = 0\}$ of $\mathcal{D}(\overline{A})$. In [31], Sato and Ueno show that $\overline{A}_{\hat{L}}$ is the infinitesimal generator for the semigroup on $C(\overline{\Theta})$ whose Green's operators $\{G_\alpha\}$ satisfy $(\alpha - \overline{A})G_\alpha u = u$ and $LG_\alpha u(x) = 0$ for $x \in \partial\Theta$. It is further shown that it is the infinitesimal generator for our desired Markov process. The collection of C^∞ functions f on $\overline{\Theta}$ such that $Lf = 0$ for $x \in \partial\Theta$ are dense in the $dom(\overline{A}_{\hat{L}})$.

PROPOSITION 2 (see Sato and Ueno [31]). *Under the conditions of Proposition 1 above, there is a time-homogeneous (Feller) Markov process* $X = (X_t, W, B_t, \{P(t,x,\cdot) : x \in \overline{\Theta}\}, t \geq 0\}$ *with infinitesimal generator* $\overline{A}_{\hat{L}}$, *and the probability transition measures* $P(t,x,\cdot)$ *of* X_t *have the fundamental solution* $p(t,x,y)$ *as a probability density with respect to the volume element measure* ν: $P(t,x,dy) = p(t,x,y)\nu(dy)$ *for all* $t \geq 0$ *and all* $x \in \overline{\Theta}$.

Proof. For the proof, see Sato and Ueno [31]. □

The goal of this section is to show the weak convergence of the transition probability of X_t, $\{P(t,x,\cdot)\}$ to the probability measure π. The following theorem gives the desired result.

THEOREM 3. *Let the family of measures* $\{P(t,x,\cdot) : x \in \overline{\Theta}, t \geq 0\}$ *be the transition probability of the diffusion process* X_t, *with generator* $A = \frac{1}{2}(\Delta - \nabla h)$ *and boundary operator* $L = \frac{\partial}{\partial n}$. *Then for each* x, $\{P(t,x,\cdot)\}$ *converges weakly to the probability measure* π *as* $t \to \infty$, *where* $q(x) \in C^\infty(\overline{\Theta})$ *is the strictly positive density function of* π *with respect to* ν *and* $h(x) = -\ln q(x)$. *More precisely, for all* $x \in \overline{\Theta}$, *and for any bounded measurable function* f *on* $\overline{\Theta}$,

$$T_t f(x) \stackrel{\text{def}}{=} \int_{\overline{\Theta}} f(y)P(t,x,dy) \to \int_{\overline{\Theta}} f(y)\pi(dy) \quad \text{as } t \to \infty.$$

Proof. The manifold $\overline{\Theta}$ is a compact, complete, separable metric space. Moreover, by the above propositions, there exist a Borel measure ν on $\overline{\Theta}$ and a strictly positive function $p(t,x,y)$ continuous in $(t,x,y) \in (0,\infty) \times \overline{\Theta}^2$ such that the transition probability $P(t,x,dy)$ equals $p(t,x,y)\ \nu(dy)$. By Theorem 1.3.4 of [22], in the case of a time-homogeneous Feller process (X_t) satisfying this condition on a locally compact, complete, separable metric space, X_t is either transient or recurrent in the sense of Harris. With $\overline{\Theta}$ compact, X_t cannot be transient, since $I_{\overline{\Theta}(X_t)} \equiv 1$, and the definition of transient would require that for any compact $K \subset \overline{\Theta}$

$$\sup_{x \in K} E_x \left[\int_0^\infty I_K(X_t)dt \right] < \infty \qquad (I_K \text{ is the indicator function of } K).$$

Therefore, X_t is recurrent in the sense of Harris, and by Theorem 1.3.5 of [22], X_t has a (T_t)-invariant probability measure Λ, and in particular, it is unique. Finally, since X_t is a Feller process satisfying the above stated condition, it has the following property given by Theorem 1.3.10 of [22]. For any bounded measurable function

f on $\overline{\Theta}$,

$$T_t f(x) \to \int_{\overline{\Theta}} f d\Lambda \quad \text{as t} \to \infty$$

holds for every $x \in \overline{\Theta}$. Then it is clear that for every $x \in \overline{\Theta}$, $\{P(t,x,\cdot)\}$ converges weakly to the unique invariant probability measure $\Lambda(\cdot)$.

Therefore, it remains only to show that $\Lambda = \pi$. Recall that (A, L) and h are

$$A = \frac{1}{2}(\Delta - \nabla h), \quad L = \frac{\partial}{\partial n}, \quad \text{and} \quad h(x) = -\ln q(x),$$

where $q(x) \in C^\infty(M)$ is the strictly positive density function of π with respect to ν, the Borel measure associated with the Riemannian volume element, Ω. The operator $\overline{A}_{\hat{L}}$ is the infinitesimal generator of the Feller semigroup $\{T_t, t \geq 0\}$ from Propositions 1 and 2. Recall that $f \in dom(\overline{A}_{\hat{L}})$ satisfy $Lf = 0$.

A measure μ is an invariant measure of an (A, L)-diffusion if and only if

$$\int_{\overline{\Theta}} \overline{A}_{\hat{L}} f(x) \mu(dx) = 0 \quad \text{for all } f \in dom(\overline{A}_{\hat{L}}) \qquad \text{(Ikeda and Watanabe [17]).}$$

Because the collection of $C^\infty(\overline{\Theta})$ functions f on $\overline{\Theta}$ such that $Lf = 0$ for $x \in \partial\Theta$ are dense in the $dom(\overline{A}_{\hat{L}})$, it suffices to verify for such f that

$$\int_{\overline{\Theta}} Af(x)\mu(dx) = 0 \quad \text{for all } f \in C^\infty(\overline{\Theta}), Lf = 0.$$

Notice that

$$\begin{aligned}
\int_{\overline{\Theta}} \overline{A}_{\hat{L}} f(x)\ \pi(dx) &= \int_{\overline{\Theta}} (Af) q(x)\ \nu(dx) = \frac{1}{2}\int_{\overline{\Theta}} q(\Delta - \nabla h) f(x)\, \nu(dx) \\
&= \frac{1}{2}\int_{\overline{\Theta}} \left(q\Delta f - q(\nabla h)\, f\right)(x)\ \nu(dx) \\
&= \frac{1}{2}\int_{\overline{\Theta}} \left(q\Delta f - q \cdot g(\nabla f, \nabla h)\right)(x)\, \nu(dx) \\
&= \frac{1}{2}\int_{\overline{\Theta}} \left(q\Delta f - q \cdot g\left(\nabla f, -\frac{1}{q}(\nabla q)\right)\right)(x)\ \nu(dx) \\
&= \frac{1}{2}\int_{\overline{\Theta}} \left(q\Delta f + g(\nabla f, \nabla q)\right)(x)\ \nu(dx) \\
&= \frac{1}{2}\int_{\overline{\Theta}} \operatorname{div}\left(q(\nabla f)\right)(x)\ \nu(dx) \\
&= \frac{1}{2}\int_{\overline{\Theta}} \operatorname{div}\left(q(\nabla f)\right)\ \Omega,
\end{aligned}$$

since $h = -\ln q$ and $\nabla(-\ln q) = -\frac{1}{q}(\nabla q)$. However, for a C^∞ vector field X on $\overline{\Theta}$, $\operatorname{div} X\ \Omega = d(i(X)\Omega)$, where $i(X)\Omega$ is the interior product of X with Ω. Hence,

$$\begin{aligned}
\int_{\overline{\Theta}} (\overline{A}_{\hat{L}} f)\ d\pi &= \frac{1}{2}\int_{\overline{\Theta}} d(i(q\,\nabla f)\Omega) \\
&= \frac{1}{2}\int_{\partial\Theta} i(q(\nabla f))\Omega \qquad \text{by Stokes' theorem.}
\end{aligned}$$

Most importantly, since $\frac{\partial f}{\partial n} = 0$, it follows that ∇f has a zero normal component and hence, at each $x \in \partial\Theta$, ∇f lies in the tangential component of the tangent space $T_x(\overline{\Theta})$, which is $T_x(\partial\Theta)$. Consequently, the interior product $i(q\,\nabla f)\Omega$ at each $x \in \partial\Theta$ consists of n vectors in an $(n-1)$-dimensional space, and thus $i(q\,\nabla f)\Omega = 0$. Consequently,

$$\int_{\overline{\Theta}} (\overline{A}_{\hat{L}} f)\, d\pi = \frac{1}{2} \int_{\partial\Theta} i(q(\nabla f))\Omega = 0.$$

Therefore, since X_t is the diffusion process generated by the operators (A, L), the measure π is an invariant probability measure of X_t. By the uniqueness of the invariant probability measure Λ, we must have $\Lambda = \pi$. □

REFERENCES

[1] Y. Amit, U. Grenander, and M. Piccioni, *Structural image restoration through deformable templates*, J. Amer. Statist. Assoc., 86 (1991), pp. 376–387.

[2] W. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

[3] P. W. Callen, ed., *Ultrasonography in Obstetrics and Gynecology*, 2nd ed., Saunders, Philadelphia, 1988.

[4] T.-S. Chiang, C.-R. Hwang, and S. J. Sheu, *Diffusion for global optimization in* $\mathbb{R}^n$, SIAM J. Control Optim., 25 (1987), pp. 737–753.

[5] S. Geman and C.-R. Hwang, *Diffusions for global optimization*, SIAM J. Control Optim., 24 (1986), pp. 1031–1043.

[6] U. Grenander, *Tutorial in Pattern Theory*, Reports in Pattern Analysis, Division of Applied Mathematics, Brown University, Providence, RI, 1983.

[7] U. Grenander, *General Pattern Theory*, Oxford University Press, Oxford, UK, 1993.

[8] U. Grenander, *Geometries of knowledge*, Proc. Natl. Acad. Sci. USA, 94 (1997), pp. 783–789.

[9] U. Grenander, Y. Chow, and D. Keenan, *HANDS: A Pattern Theoretic Study of Biological Shapes*, Springer-Verlag, New York, 1990.

[10] U. Grenander and D. Keenan, *On the shape of plane images*, SIAM J. Appl. Math., 53 (1993), pp. 1072–1094.

[11] U. Grenander and D. Keenan, *Understanding variable shapes in* 3*-D*, Stoch. Models, 11 (1995), pp. 51–78.

[12] U. Grenander and M. I. Miller, *Representations of knowledge in complex systems*, J. Roy. Statist. Soc. Ser. B, 56 (1994), pp. 549–603.

[13] C. R. Hill and A. Kratochwil, eds., *Medical Ultrasonic Images: Formation, Display, Recording, and Perception*, Elsevier, North–Holland, Amsterdam, Princeton, New York, 1981.

[14] R. Holley, S. Kusuoka, and D. Stroock, *Asymptotics of the spectral gap with application to the theory of simulated annealing*, J. Funct. Anal., 83 (1989), pp. 333–347.

[15] L. Holm and C. Sander, *Mapping the protein universe*, Science, 273 (1996), pp. 595–602.

[16] L. Holm and C. Sander, *Touring protein fold space with Dali/FSSP*, Nucl. Acids Res., 26 (1998), pp. 316–319.

[17] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, 2nd ed., North–Holland, Amsterdam, 1989.

[18] S. Ito, *Fundamental solutions of parabolic differential equations and boundary value problems*, Japan. J. Math., 27 (1957), pp. 55–102.

[19] S. Ito, *A boundary value problem of partial differential equations of parabolic type*, Duke Math. J., 24 (1957), pp. 299–312.

[20] S. Ito, *A remark on my paper "A boundary value problem of partial differential equations of parabolic type" in Duke Mathematical Journal*, Proc. Japan Acad., 34 (1958), pp. 463–465.

[21] D. Keenan and P. Shorter, *Stochastic global optimization on a manifold with applications in image processing*, 2004, in preparation.

[22] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, New York, 1990.

[23] A. B. Kurtz, R. J. Wapner, R. J. Kurtz, D. D. Dershaw, C. S. Rubin, C. Cole-Beuglet, and B. B. Goldberg, *Analysis of biparietal diameter as an accurate indicator of gestational age*, J. Clin. Ultrasound, 8 (1980), pp. 319–326.

[24] R. A. Lerski, ed., *Practical Ultrasound*, IRL Press, Oxford, UK, Washington, D.C., 1988.
[25] L. Matejic, *A mathematical representation of biological variability in medical images*, Quart. Appl. Math., 61 (2003), pp. 1–16.
[26] H. P. McKean, *Brownian motion on the* 3*-dimensional rotation group*, Mem. Coll. Sci. Univ. Kyoto Ser. A Math., 33 (1960/1961), pp. 25–38.
[27] H. P. McKean, *Stochastic Integrals*, Academic Press, New York, 1969.
[28] M. I. Miller, G. E. Christensen, Y. Amit, and U. Grenander, *Mathematical textbook of deformable neuroanatomies*, Proc. Natl. Acad. Sci. USA, 90 (1993), pp. 11944–11948.
[29] D. Mumford, *Pattern theory: A unifying perspective*, in First European Congress of Mathematics, A. Joseph et al., eds., Birkhäuser, Basel, 1994, pp. 187–224.
[30] J. F. Nash, *The imbedding problem for Riemannian manifolds*, Ann. of Math. (2), 63 (1956), pp. 20–63.
[31] K. I. Sato and T. Ueno, *Multi-dimensional diffusion and the Markov processes on the boundary*, J. Math. Kyoto Univ., 4 (1965), pp. 529–605.
[32] P. Shorter, *Diffusion Processes for Stochastic Global Optimization on a Manifold with Applications in Image Processing*, Ph.D. dissertation, Department of Mathematics, University of Virginia, Charlottesville, VA, 1996.
[33] A. Srivastava, U. Grenander, G. Jensen, and M. I. Miller, *Jump-diffusion processes on orthogonal groups for object pose estimation*, J. Statist. Plann. Inference, 103 (2002), pp. 15–37.
[34] H. Whitney, *Differentiable manifolds*, Ann. of Math. (2), 37 (1936), pp. 645–680.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 961–976

EXACT SIMILARITY SOLUTIONS OF COUPLED NONSTEADY NAVIER–STOKES AND ENERGY EQUATIONS IN LIQUIDS*

SHIMON HABER†

Abstract. Exact similarity solutions are obtained for coupled Navier–Stokes and energy equations that govern the time-dependent motion of a gravitation-free viscous liquid with variable density in one- and three-dimensional (1D and 3D) spaces. The 1D case deals with propagation and diffusion of an initial algebraic density hump C/x that is subjected to a constant flow rate at the origin. We demonstrate that the initial temperature distribution is convected without being diffused at very long times. We also demonstrate that two different propagation velocities exist for short and long times. The 3D case addresses the implosion of an insulated closed system with an initial radially symmetric algebraic density hump C/r^3. We demonstrate that if viscous dissipation and liquid compressibility terms are neglected in the energy equation, very strong shock-like pressure distributions may occur that may lead to a "black hole" within a finite time.

A comprehensive analysis is also carried out for density fields with an initial, r^n, radially symmetric distribution in 1D, 2D, and 3D spaces. A first integral is obtained for all n's in a 2D space. A phase-space solution is utilized to depict the system evolution and stability for any value of n. It also allows us to consider intriguing aphysical negative density fields, manifesting a peculiar periodic solution for the 3D ($n = 0$) case that mimics a prey-predator problem.

Key words. general fluid mechanics, Navier–Stokes equations, liquids, similarity solutions

AMS subject classification. 76M55

DOI. 10.1137/S0036139902420456

1. Introduction.

To date, very few *exact* analytical solutions are known for the nonsteady motion of viscous *liquids* with variable densities, owing to the fact that the governing Navier–Stokes equations are nonlinear and coupled with the energy equation. To circumvent this difficulty, numerical solvers have frequently been used but have rarely been verified against exact analytical solutions of nontrivial flow problems that encompass the interplay between viscosity, inertia, temperature, and density fields. Analytical expressions have been obtained in the past using extremely effective similarity methods; alas, in most previous cases, only an approximate set of the Navier–Stokes equations has been employed. Similarity solutions were also criticized due to the fact that, to quote Barenblatt (1979), "In nonlinear problems, exact special solutions sometimes *appear* to be useless; since there is no principle of superposition, one cannot immediately find a solution of the problem with arbitrary initial conditions." Notwithstanding, Barenblatt emphasizes the importance of similarity solutions, as they may "represent the asymptotics of a wide class of other more general solutions." Burger's nonlinear equation (see Whitham (1974)) is a celebrated attempt to describe approximately the combined effects of nonlinear density propagation and diffusion under the assumption of small amplitude disturbances. A similarity solution was obtained for the case of an initial hump, a problem that is also addressed in this paper, only here we employ the full Navier–Stokes equations. Barenblatt (1979) addresses the problem of thermal flame propagation in gaseous mixtures. An approximate equation is obtained for the propagating wave front (the

*Received by the editors December 27, 2002; accepted for publication (in revised form) August 27, 2003; published electronically March 30, 2004.
http://www.siam.org/journals/siap/64-3/42045.html
†Department of Mechanical Engineering, Technion-Israel Institute of Technology, Haifa, 32000, Israel (mersh01@techunix.technion.ac.il).

inner flow approximation), and the analysis indicates how to obtain the long time propagation velocity. It is interesting to note that, since the propagating velocity is much smaller than the speed of sound, the author assumes that the gas density depends only upon temperature, and the pressure field is treated as an independent variable. In addition, the functional forms of the velocity and temperature fields are generally obtained before employment of the momentum equation. The latter is used to obtain the pressure field and the value of the long time propagation velocity. We follow to some extent a similar approach, where the velocity and density (temperature) fields are addressed first and the pressure distribution is consequently obtained. More recently, similarity methods were also employed by Canright and Morris (1993) (addressing the buoyant instability of a viscous film over a passive fluid, an attempt to describe the behavior of lava lakes and mantle convection), Romero and Yost (1996) (who obtained a similarity solution for a capillary-driven flow in a V-shaped surface groove), Woods and Fitzgerald (1997) (who described the temperature field as liquid spreads from a line source into a porous rock with a time-dependent liquid injection), Koehler et al. (1998) (who obtained a similarity solution for the drainage of liquid foams under gravity), Witelski and Bernoff (1999) and Zhang and Lister (1999) (who addressed the stability of van der Waals–driven thin film rupture), and Christopher and Wang (2001) (who simulated the Marangoni convection around a vapor bubble during nucleation and growth), to name a few.

The objective of this communication is to present an *exact* analytic solution for *liquid* fields or *incompressible* gases such that their density is a function of temperature only while the pressure field can be treated as an independent variable. The fields are radially symmetric and are governed by the fully coupled nonlinear Navier–Stokes and energy equations under a minimal set of assumptions. The solution illustrates the interplay between the convective, diffusive, and temporal terms of the equations and the effects of the phenomenological coefficients on the flow variables. To achieve this goal, we assume that the flow and temperature fields satisfy a particular set of initial and boundary conditions and that the density is inversely proportional to the temperature field, a relation commonly satisfied by liquids and by ideal incompressible gases. Despite these obvious limitations of the solution class, we shall demonstrate that the results possess remarkable features (mentioned in the abstract) of general physical significance. In particular, in cases where a moving sharp interface is assumed to divide the flow field into two regions of distinct densities, the model elucidates the effect of a diffused density front. (Obviously, far from the interface the model is no longer valid and can no longer describe the flow field faithfully.) Thus, for instance, the exact 1D solution shows a surprising resemblance to a certain region of the flow generated in a shock tube despite the fact that the latter addresses compressible flow (see a broader discussion in section 5). Moreover, one of the possible 3D radially symmetric solutions elucidates a process that resembles the behavior of the flow field near the liquid interface surrounding an expanding bubble, only here the adiabatic interface is considered to be diffused. We also believe that the radially symmetric solutions are a necessary first step if stability to small nonradially symmetric initial density disturbances, heat sources, Coriolis accelerations, etc., is to be explored. Incorporation of the radial gravity term in the momentum equation and the foregoing asymmetric disturbances (not accounted for in this paper) can be employed to model diffused asymmetrical interfacial phenomena in earth science (see Gerald (1971), Davies (1999)).

Finally, the exact solutions can probably be utilized as benchmarks to verify the accuracy of numerical solvers since they encompass the combined effects of inertia, viscosity, heat convection, and diffusion in an unsteady flow field. (This latter suggestion

must, for the time being, be tested with care, since the solution's stability has not been addressed in this paper.)

The paper is divided into four main parts. Section 2 defines the governing equations and the simplifying assumptions used to derive them. Section 3 presents the method we used to derive the particular exact similarity solutions of the coupled Navier–Stokes and energy equations in 1D, 2D, and 3D spaces. Section 4 analyzes a broader family of radially symmetric problems. Utilizing a phase map analysis proves that these similarity solutions can never be chaotic. Section 5 discusses and summarizes the main results.

2. Statement of problem. A common set of equations governing the motion of Newtonian liquids includes the mass conservation equation,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{2.1}$$

the linear momentum equation,

$$\rho \left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right) = -\nabla p + \mu \nabla^2 \mathbf{v} + \left(\kappa + \frac{\mu}{3} \right) \nabla (\nabla \cdot \mathbf{v}) + \rho \mathbf{g}, \tag{2.2}$$

the energy conservation equation,

$$\rho c_p \left(\frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T \right) = k \nabla^2 T, \tag{2.3}$$

and the equation describing the relation between temperature and density,

$$\rho = \frac{\rho_0}{[1 + \gamma \rho_0 (T - T_0)]}, \tag{2.4}$$

where ρ, p, T, and $\mathbf{v}$ denote the density, pressure, temperature, and velocity fields; μ and κ stand for the fixed fluid shear and bulk viscosities; and c_p and k stand for the fixed specific heat and thermal conductivity coefficients. Notice that in (2.3) we neglected the viscous dissipation and the temperature increase with compression terms. The exclusion of the energy dissipation term is justified in most cases, except for flow fields that are subjected to very high shear rates (e.g., flows in journal bearings). The omission of the compressibility term is suited to liquids that satisfy (2.4), and consequently, the ratio of their specific heats, c_p/c_v, is unity.

For γ, ρ_0, and T_0 constants, (2.4) applies to many liquids and tacitly implies that the liquid is incompressible in the sense that the density does not vary with pressure. Accordingly, the pressure field must be regarded as an independent variable (Aris (1962, p. 111)). Thus, (2.1)–(2.4) constitute a consistent set of four equations for the four unknown fields ρ, p, T, and $\mathbf{v}$.

3. Method of solution. We assume that the velocity field is irrotational and guess that it possesses the following form:

$$\mathbf{v} = \frac{k}{c_p} \nabla \frac{1}{\rho}. \tag{3.1}$$

Substituting (3.1) into the mass conservation equation (2.1) results in a second order nonlinear equation for the density field,

$$\frac{\partial \rho}{\partial t} + \frac{k}{\rho^2 c_p} \nabla \rho \cdot \nabla \rho = \frac{k}{\rho c_p} \nabla^2 \rho. \tag{3.2}$$

Remarkably, (3.2) is also recovered if we rewrite the energy equation (2.3) utilizing (2.4) and (3.1) for the velocity field. Thus, the energy and mass conservation equations collapse into a single equation (3.2). *This result is the crux of the analysis.* Notice that Hopf (1950), Cole (1951), and Camacho and Brenner (1995) utilized a similar idea to derive a solution for the diffusion equation, where they assumed that the velocity field is proportional to $\nabla(\ln\rho)$. If (3.1) and (3.2) are substituted into (2.2), the following equation for the pressure is obtained:

$$\left(\frac{c_p}{k}\right)^2\nabla p = P_r\nabla[\nabla^2\rho^{-1}] - \rho\nabla[\rho^{-1}\nabla^2\rho^{-1} - 0.5\nabla\rho^{-1}\cdot\nabla\rho^{-1}] + \left(\frac{c_p}{k}\right)^2\rho\mathbf{g}, \tag{3.3}$$

where the Prandtl number $P_r = \tilde{\mu}c_p/k$ is based on the modified viscosity $\tilde{\mu} = \kappa + 4\mu/3$. Equation (3.3) cannot always be solved for the pressure since an irrotational solution for the velocity field would generally not satisfy the momentum equation. However, a solution for the pressure may be found in 1D Cartesian problems, in radially symmetric 2D and 3D problems, and in case the gravity field is radially symmetric. The latter is commonly encountered in earth sciences such as mantle convection under the earth lithosphere (Davies (1999)). Henceforth, we shall focus on these simplified cases. We also define the following dimensionless variables: $\hat{\rho} = \rho/\rho_0$, $\hat{\mathbf{r}} = \mathbf{r}/a$, $\hat{t} = kt/(\rho_0 c_p a^2)$, $\hat{\mathbf{v}} = (\rho_0 c_p a/k)\mathbf{v}$, and $\hat{p} = (c_p^2\rho_0 a^2/k^2)p$, where a is a given length-scale. (The caret symbol will be used to describe dimensionless quantities.)

3.1. A similarity solution for the propagation and diffusion of an *n*th order polynomial hump in an *m*-dimensional space. From (3.2), the radially symmetric mass/energy conservation equation in an m-dimensional space is

$$\frac{\partial\hat{\rho}}{\partial\hat{t}} + \frac{1}{\hat{\rho}^2}\left(\frac{\partial\hat{\rho}}{\partial\hat{r}}\right)^2 = \frac{1}{\hat{\rho}}\hat{\nabla}^2\hat{\rho}, \tag{3.4}$$

where the dimensionless Laplace operator is

$$\hat{\nabla}^2 = \frac{\partial^2}{\partial\hat{r}^2} + \frac{m-1}{\hat{r}}\frac{\partial}{\partial\hat{r}}.$$

From (2.2), the corresponding momentum equation is

$$\frac{\partial\hat{p}}{\partial\hat{r}} = (P_r - 1)\frac{\partial}{\partial\hat{r}}\hat{\nabla}^2\frac{1}{\hat{\rho}} - \hat{\rho}\frac{m-1}{\hat{r}}\left[\frac{\partial(1/\hat{\rho})}{\partial\hat{r}}\right]^2 + B\hat{\rho}, \tag{3.5}$$

where $B = a^3c_p^2\rho_0^2 g_r/k^2$ is the dimensionless gravity number based upon the radial gravity acceleration g_r (considered negative if pointing toward the origin). Generally, g_r will depend on the whole field distribution of ρ and thereby on the global geometrical configuration of the flow field. Such an additional complication is avoided in this paper (to be addressed in a subsequent paper), and henceforth we shall investigate the microgravity case for which B is negligibly small (i.e., the flow is *not* buoyancy driven). Notice that if the local distribution of the density field is known and $B = 0$, the pressure field can readily be obtained. Indeed, for the 1D ($m = 1$) case, the explicit distribution of the density field is not required, and a straightforward integration of (3.5) is possible.

To obtain a solution of the nonlinear partial differential equation (3.4) is a formidable task. Indeed, it is well known that a general solution representation does not exist for partial differential equations. At most, what we hope to obtain is a solution

class that can be applied to a given domain and boundary conditions. Our task is made easier if we tailor the boundary conditions a posteriori to make them fit the solution. In what follows we focus on some similarity solutions of (3.4) and, in particular, on the case in which the initial density field is a hump described by $\hat{\rho} = C_1 \hat{r}^n$ at $\hat{t} = 0$. In this case a similarity solution exists that possesses the form

$$\hat{\rho} = \hat{r}^n g(\eta), \qquad \eta = \frac{\hat{r}}{\hat{t}^{1/(n+2)}}, \tag{3.6}$$

where η is the appropriate similarity variable. Introducing (3.6) into (3.4) results in a second order, ordinary, nonlinear differential equation for g,

$$gg'' - (g')^2 + \frac{\eta^{n+1} g^2 g'}{n+2} + \frac{n(m-2)g^2}{\eta^2} + \frac{(m-1)g'g}{\eta} = 0, \tag{3.7}$$

where the prime symbol stands for differentiation with respect to η.

Equation (3.7) possesses a simple *particular* solution[1] that contains no free constants of integration,

$$g = (4 - 2m)\eta^{-n-2}. \tag{3.8}$$

A *general* solution of (3.7) for any n and m that includes two independent integration constants is probably impossible. Notwithstanding, (3.7) can be solved analytically for the parameter sets $(m, n) = (1, -1)$ and $(m, n) = (3, -3)$. The first set describes the case of an initial density hump $\rho(\hat{x}, t = 0) = C_1/\hat{x}$ propagating in 1D space. The second case addresses the propagation of an initial density hump $\hat{\rho}(\hat{r}, t = 0) = C_1/\hat{r}^3$ in 3D space.

3.2. The exact 1D solution for the propagation of an initial hump.

$$\rho(\hat{x}, t = 0) = \frac{C_1}{\hat{x}} \quad \text{or} \quad \rho(\hat{x}, t = 0) = \frac{C_1}{\hat{x} + \alpha}.$$

In case $n = -1$ and $m = 1$, namely for an initial 1D density hump $\hat{\rho}(\hat{x}, 0) = C_1/\hat{x}$ ($\hat{x}$ stands for the 1D spatial coordinate instead of $\hat{r}$), a *general analytical* solution of (3.7) can be obtained. A step-by-step derivation is as follows: Divide (3.7) by g^2 and define $u = g'/g$; consequently, rewrite (3.7) as $u' + g' + 1/\eta^2 = 0$. Thus, a first integral $u + g - 1/\eta = C_1$ exists, where C_1 is an arbitrary constant. In terms of the unknown function g, the latter equation can be rewritten in the form $g' - (C_1 + 1/\eta)g = -g^2$, which is a nonlinear Bernoulli equation. Defining a new dependent variable $h = 1/g$, the latter equation transforms into the simple linear equation, $h' + (C_1 + 1/\eta)h = 1$. Its straightforward solution eventually yields

$$g = \left[\frac{C_2 e^{-C_1 \eta}}{\eta} + \frac{1}{C_1} - \frac{1}{(\eta C_1^2)} \right]^{-1}, \tag{3.9}$$

where C_1 and C_2 are arbitrary constants. Consequently, the solution for the density field with $n = -1$ is

$$\hat{\rho} = \left(C_2 \hat{t} e^{-C_1 \hat{x}/\hat{t}} + \frac{\hat{x}}{C_1} - \frac{\hat{t}}{C_1^2} \right)^{-1}.$$

[1] The solution is particular in the sense that it does not possess any free constants of integration and cannot be obtained from the general solution by a particular choice of the constants.

A more general expression can be obtained if $\hat{x}$ is replaced with $\hat{x}+\alpha$:

$$\hat{\rho}=\left(C_2\hat{t}e^{-C_1(\hat{x}+\alpha)/\hat{t}}+\frac{\hat{x}+\alpha}{C_1}-\frac{\hat{t}}{C_1^2}\right)^{-1}, \tag{3.10}$$

where $\alpha>0$ is an arbitrary constant. Equation (3.10) satisfies a more general initial condition $\hat{\rho}(\hat{x},\hat{t}=0)=C_1/(\hat{x}+\alpha)$ that is nonsingular for all $\hat{x}>0$. (This transformation is possible since, in case $m=1$, (3.4) contains only differentials of x.)

Notice that the particular solution (3.8) is still valid here but cannot be obtained from a specific choice of the constants, a common occurrence in nonlinear equations (Bender and Orszag (1978, p. 4)).

From (3.1) and (3.5) the solutions for the velocity and pressure fields are

$$\hat{v}_x=-C_1C_2\exp\left[\frac{-C_1(\hat{x}+\alpha)}{\hat{t}}\right]+\frac{1}{C_1}, \tag{3.11}$$

$$\frac{\partial\hat{p}}{\partial\hat{x}}=-(P_r-1)C_1^3C_2\hat{t}^{-2}e^{-C_1(\hat{x}+\alpha)/\hat{t}}. \tag{3.12}$$

A specific choice of the constants C_1 and C_2 defines a nontrivial solution for the propagation and diffusion of an initial density hump $\hat{\rho}(\hat{x},0)=C_1/(\hat{x}+\alpha)$, with liquid injected at $x=0$ at a rate $\hat{v}_x=-C_1C_2\exp(-C_1\alpha/\hat{t})+1/C_1$ that readily reaches a fixed value. Notice that if $C_2=1/C_1^2$ and $\alpha=0$, the solution describes the behavior of an insulated closed system at $x=0$ for all times.

Equation (3.10) is a remarkable solution of the Navier–Stokes and energy equations. It possesses short *and* long time behavior of a propagating wave. For short times, the density distribution propagates with velocity $\hat{u}_P=1/C_1$, while for long times the propagation velocity changes into $\hat{u}_P=1/C_1-C_2C_1$. Figure 1 illustrates the density distribution for various times and locations. It vividly demonstrates that the maximum density value (normalized for position) propagates and reaches downstream locations as time evolves.

3.3. The exact 3D solution of a propagating initial hump $\boldsymbol{\rho(\hat{r},t=0)=C_1/\hat{r}^3}$. For an initial density distribution $\hat{\rho}(\hat{r},0)=C_1/\hat{r}^3$, a general *analytical* solution of (3.7) in a 3D space exists,[2]

$$g(\eta)=\eta^3e^{C_1/\eta}\left\{C_2+\int_\eta^{\eta_1}\eta e^{C_1/\eta}d\eta\right\}^{-1}, \tag{3.13}$$

where $\eta=\hat{r}\hat{t}$ and C_1 and C_2 are two arbitrary constants of integration. Hence,

$$\hat{\rho}=\hat{t}^3e^{C_1/\eta}\left\{C_2+\int_\eta^{\eta_1}\eta e^{C_1/\eta}d\eta\right\}^{-1}. \tag{3.14}$$

Substitution of (3.13) into (3.1) and (3.3) yields the expressions for the velocity and the pressure fields:

$$\hat{v}_r=\frac{\hat{r}(C_1/g-1)}{\hat{t}}, \tag{3.15}$$

$$\frac{\partial\hat{p}}{\partial\hat{r}}=(P_r-1)\left[\frac{C_1^2(C_1-4\eta)}{g\eta^3}-\frac{C_1^2}{\eta^3}+\frac{C_1}{\eta^2}\right]-\frac{2g}{\eta^2}\left(\frac{C_1}{g}-1\right)^2. \tag{3.16}$$

[2]The solution process is similar to that used for the 1D case (see section 3.2), only here define $u=\eta^2g'/g$ and proceed accordingly.

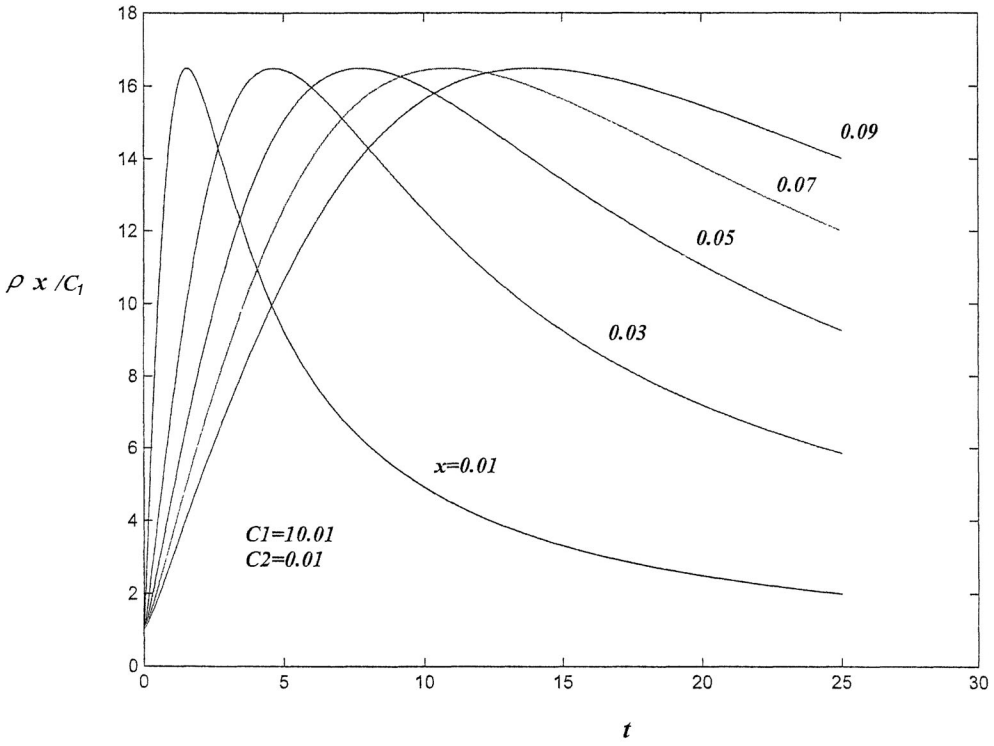


FIG. 1. *The evolution of the density field for a* 1*D case in which the initial density is given by* $\hat{\rho}(\hat{x}, \hat{t} = 0) = C_1/\hat{x}$. *The similarity parameter is* $\eta = \hat{x}/\hat{t}$ *(see* (3.10)*).*

At the limit $\eta \to 0_+$ and $C_1 > 0$ the inverse of g can asymptotically be expanded, $1/g = [1 + 3\eta/C_1 + 12(\eta/C_1)^2 + 60(\eta/C_1)^3 + \cdots]/C_1$. Hence for $\hat{t} \to 0_+$ and r held fixed, $\hat{\rho} \to C_1\hat{r}^{-3}$, $\hat{v}_r \to 3\hat{r}^2/C_1$, and $\partial\hat{p}/\partial\hat{r} \to 6(2P_r - 5)/C_1$. Similarly, for $\hat{r} \to 0$ and $\hat{t}$ held fixed, the radial velocity $\hat{v}_r$ vanishes. Hence, (3.14) is a nontrivial solution for the propagation and diffusion of an initial radially symmetric density hump $\hat{\rho}(\hat{r}, \hat{t} = 0) = C_1/\hat{r}^3$ with no liquid sources or sinks at $r = 0$. The solution may experience a blow-up at a finite time when the expression in curly brackets in (3.14) vanishes. Before blow-up occurs, for long times $\eta > \eta_1$, where η_1 satisfies the condition $\exp(C_1/\eta_1) = O(1)$, the solution no longer depends directly upon the initial condition (the value of C_1), and the density distribution possesses approximately the following functional form:

$$\hat{\rho} = \hat{t}^3 \left\{C - 0.5\eta^2\right\}^{-1}, \tag{3.17}$$

which eventually becomes singular after a finite time. If, however, the time-space domain is limited to $0 < \eta \leq \eta_1$ and $C_2 > 0$, the solution yields finite density and velocity fields. An interesting case arises if we choose a solution domain $0 < \eta < \eta_1$ with $g(\eta_1) = C_1$. In this case $v_r(\eta_1)$ and $[\partial T/\partial r]_{\eta=\eta_1}$ vanish at the domain boundaries, and the solution describes a *closed insulated* system with an initial algebraic density hump, $C_1/\hat{r}^3$, that *implodes* within a finite time at a rate proportional to $1/\hat{t}^2$. In Figures 2 and 3 we picked $\eta_1 = C_1 = 1$ and $C_2 = e$, which results in $g(\eta_1 = 1) = C_1 = 1$. Figure 2 describes the finite value of g obtained for all η in the space-time

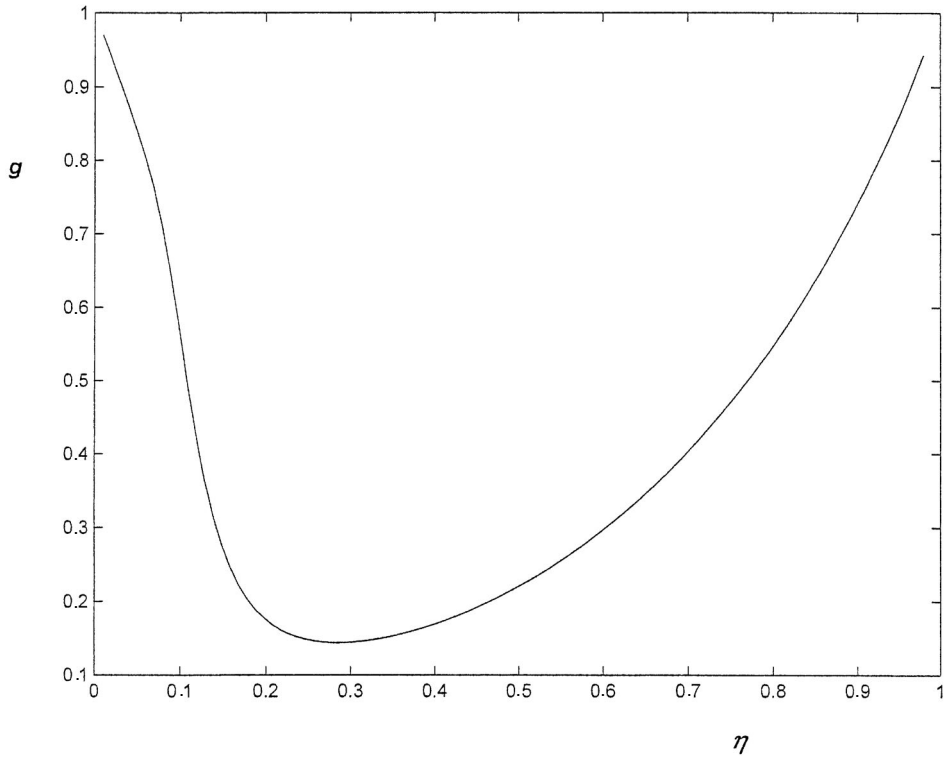


FIG. 2. *The density field for a 3D case in which the initial density is given by* $\hat{\rho}(\hat{r}, \hat{t} = 0) = C_1/\hat{r}^3$, *where* $C_1 = 1$, $\eta_1 = 1$, *and* $C_2 = e$. *The similarity parameter is* $\eta = \hat{r}\hat{t}$.

domain $0 < \hat{r}\hat{t} < 1$ (the density distribution is equal to $g/\hat{r}^3$). Figure 3 illustrates the resulting pressure gradient distribution. The figures exhibit steep gradients in the density and pressure fields that resemble diffused shock waves. Initially, the space extends to infinity, storing infinite mass. As time evolves, the system's spatial domain shrinks, a process that is accompanied by a strong flow respreading the density field, while no flow is permitted through the boundaries.

These results raise an interesting question. Can "black holes" exist in liquid systems obeying the Navier–Stokes equations? Obviously, this cannot be the case, since, at the last stages of the implosion process, strong velocity gradients exist, and it is no longer correct to assume that viscous dissipation and heating due to compressibility effects are negligibly small. Thus, the energy equation (2.3) no longer faithfully represents the energy balance in the system.

Equation (3.14) with negative $\hat{t}$ values ($-\infty < \hat{t} < 0$) describes a whole new class of solutions. Figure 4 illustrates the solution for a density field that resembles the behavior of the fluid near the adiabatic interface of an expanding bubble in which the interface is described by a diffused domain rather than by a distinct jump in the density field.

4. Phase-space analysis. Equation (3.7) can be drastically simplified utilizing the following transformations:

$$G = \frac{\eta^{n+2} g}{n+2}, \quad U = \frac{\eta g'}{g}, \quad w = \ln \eta. \tag{4.1}$$

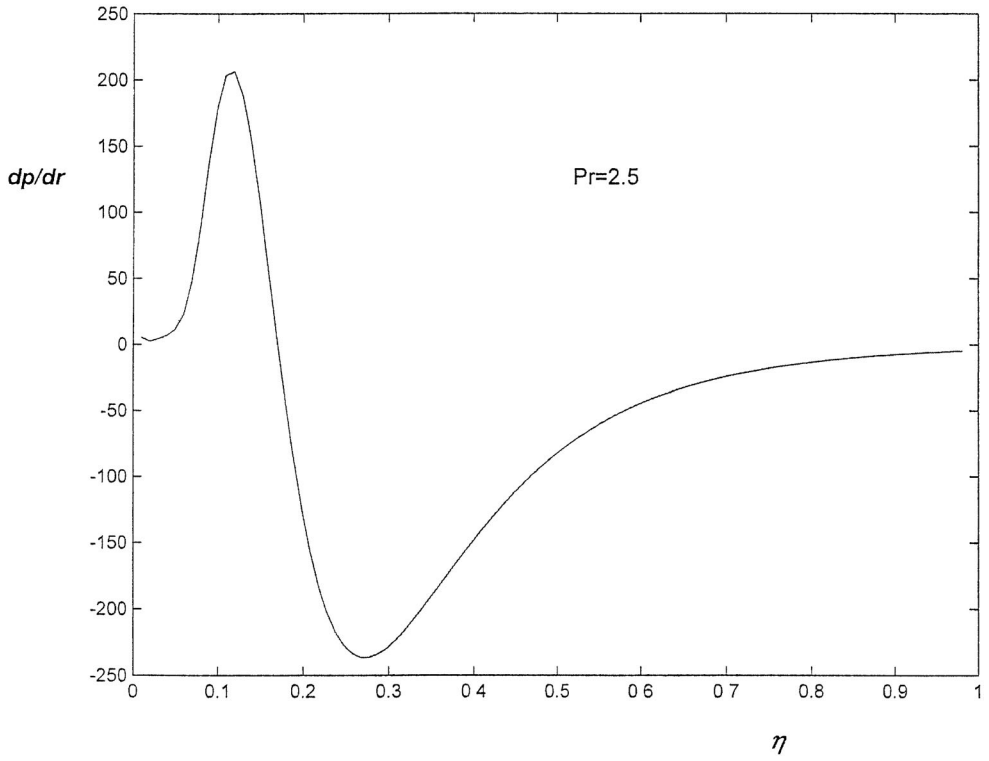


FIG. 3. *The pressure field for a 3D case in which the initial density is given by* $\hat{\rho}(\hat{r}, \hat{t} = 0) = C_1/\hat{r}^3$, *where* $C_1 = 1$, $\eta_1 = 1$, *and* $C_2 = e$. *The similarity parameter is* $\eta = \hat{r}\hat{t}$.

Substituting (4.1) into (3.7) yields an equivalent set of two first order nonlinear equations in G and U that is amenable to a simple analysis in the G-U phase-space:

$$\begin{aligned} \frac{dG}{dw} &= (n+2)G + UG, \\ \frac{dU}{dw} &= (2-m)U - UG + n(2-m). \end{aligned} \tag{4.2}$$

System (4.2) possesses a close resemblance to the well-known equations governing the prey-predator problem. For U positive and $n > -2$, the growth-rate of predator G depends upon its natural birth-rate $(n+2)$ and upon the abundance of the prey population U. For $m < 2$ the growth-rate of the prey population depends upon its rate of birth $(2-m)$ and a fixed rate, $n(2-m)$, of influx or outflow of its members for $n > 0$ or $-2 < n < 0$, respectively. Its decrease in population is directly related to predator population. For U negative, the roles of the predator and prey are exchanged. System (4.2) also manifests that chaotic solutions are impossible for any value of n or m since the order of the system is exactly 2 (no explicit dependency on w exists). System (4.2) possesses two equilibrium points A and B for which G and U have the following values:

$$\begin{aligned} \overline{G}_{\mathrm{A}} &= 0, & \overline{U}_{\mathrm{A}} &= -n, \\ \overline{G}_{\mathrm{B}} &= -2(m-2)/(n+2), & \overline{U}_{\mathrm{B}} &= -n-2. \end{aligned} \tag{4.3}$$

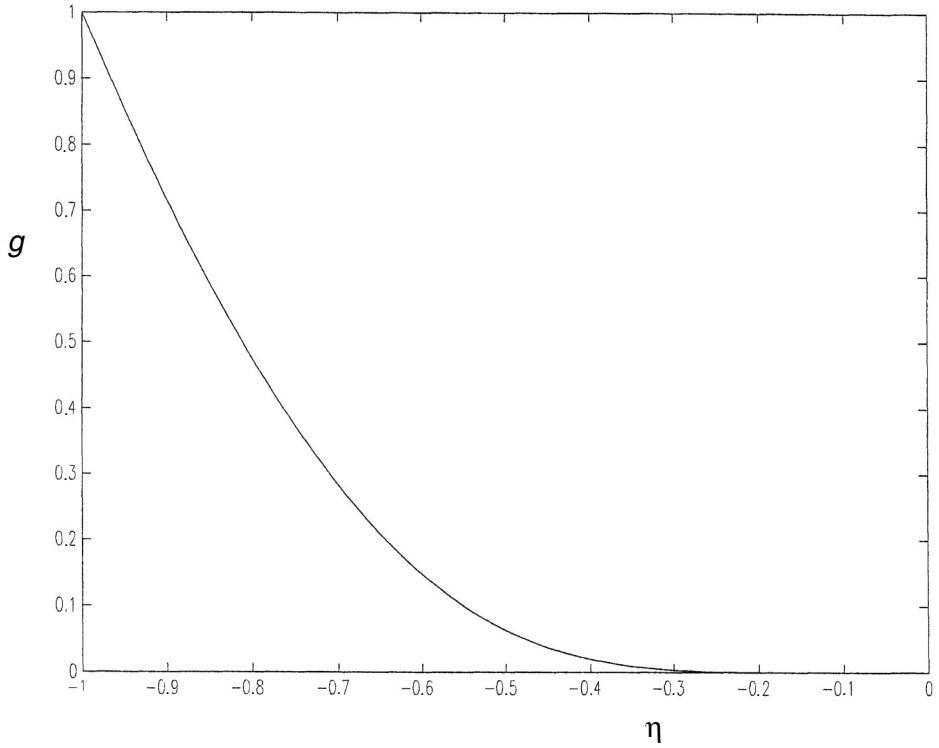


FIG. 4. *The density field for a 3D case for negative times* $-\infty < \hat{t} < 0$*;* $C_1 = 1$*,* $\eta_1 = -1$*, and* $C_2 = -1/e$*. The similarity parameter is* $\eta = \hat{r}\hat{t}$*.*

A simple linear stability analysis of system (4.2) proves that A is an unstable node if $m < 2$, and a saddle point if $m > 2$ (for any value of n). Point B demonstrates a much richer behavior. If $m < 2$, B is a saddle point. If $m > 2$, point B could be a center, a stable node or a stable spiral, an unstable node or an unstable spiral, depending upon the values of n and m. If $n < n_2$, B is a stable spiral. If $n_2 < n < -2$, B is a stable node. If $-2 < n < n_1$, B is an unstable node. If $n_1 < n < 0$, B is an unstable spiral. If $n = 0$, B is a center, and if $n > 0$, B is a stable spiral. The parameters n_1 and n_2 ($n_1 > n_2$) are $n_{1,2} = (16 \pm \sqrt{32m - 64})/(m - 10)$. Notice that $n = -2$ is excluded from the analysis since it degenerates to the particular solution. In case $m = 2$ (the 2D axisymmetric problem) linear stability analysis is not feasible, and we treat it separately in section 4.2. Figure 5 summarizes the above conclusions and depicts the type and stability of points A and B in the $m \times n$ plane. Notice that we treated m as a continuous parameter despite the fact that it may normally possess integral values only. In the following we explicitly explore the similarity solutions for $m = 1, 2, 3$ that pertain to flow fields in 1D, 2D, and 3D spaces, respectively.

4.1. Similarity solutions for 1D flows. A phase-plane analysis of (4.2) for $m = 1$ yields trajectories for flows in a 1D space. The phase-maps are depicted in Figure 6 for typical values of n. Notice that positive values of the density field correspond to positive values of G if $n > -2$ and negative values of G if $n < -2$. It is interesting to note that the two spaces of positive and negative densities do not

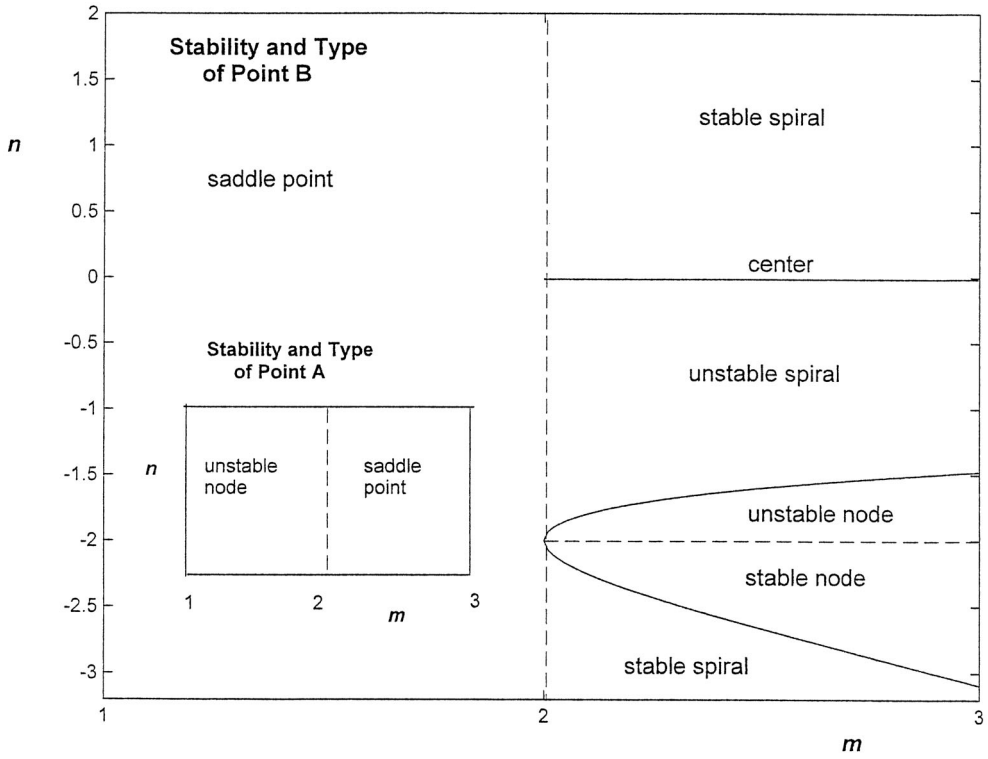


FIG. 5. *Type and stability of the fixed points A and B of system* (4.2) *in the* $m \times n$ *plane (*m *defines the dimension of the flow space, whereas* n *stands for the polynomial order of the initial density distribution).*

interact, and a trajectory that started initially in one space will always remain in that space (a comforting idea?).

In case $n = 0$ (depicted in Figure 6(c)) a simple separation of variables makes it possible to obtain a first integral of (4.2),

$$Ge^{-G} = CU^2 e^U, \tag{4.4}$$

where C is a constant of integration. Further analytic integration of this transcendental equation is a formidable task.

4.2. Similarity solutions for 2D axisymmetric flows. For 2D axisymmetric flows ($m = 2$), a simple separation of variables of (4.2) yields a general first integral for all $n \neq -2$,

$$G = -(n+2)\ln|U| - U + C, \tag{4.5}$$

where C is an arbitrary constant of integration. Hence, from (4.2),

$$w = \int \frac{dU}{U[(n+2)\ln|U| + U - C]} + w_0. \tag{4.6}$$

Figures 7(a,b) illustrate a single trajectory in phase space for $n = -3$ and $n = 1$, respectively. A particular value of C is chosen so that the $G = 0$ coordinate is

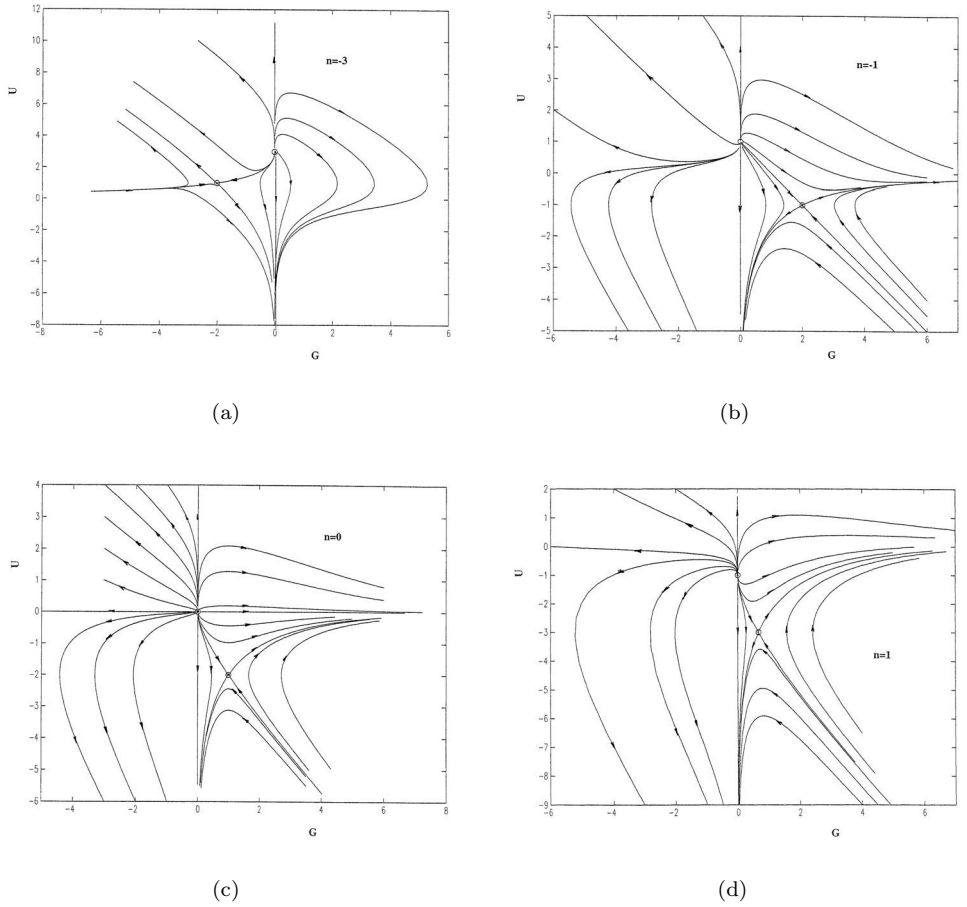
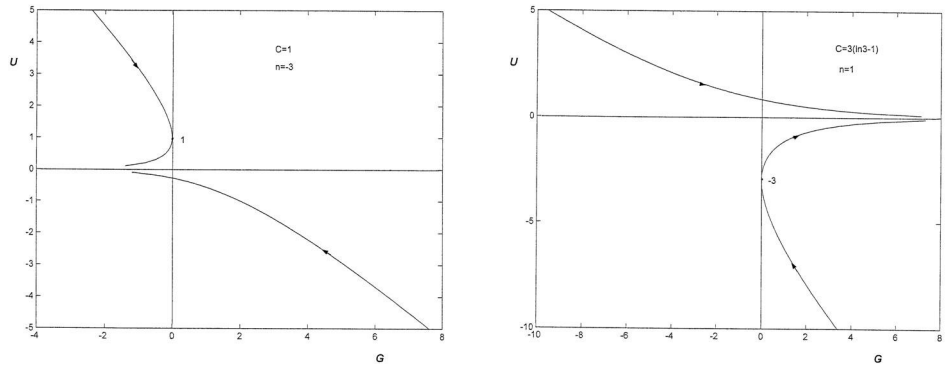


FIG. 6. *Phase-maps of one-dimensional flow cases for various values of* n. (a) $n = -3$, (b) $n = -1$, (c) $n = 0$, *and* (d) $n = 1$. *Positive density fields are obtained in the left-half plane in case* (a) *and right-half plane in cases* (b), (c), *and* (d). *Case* (b) *was solved analytically (see* (3.10)*).*

tangent to the trajectory. For $n = -3$ only the left-half plane, $G < 0$, corresponds to positive density values. Hence, the shown trajectory is a separatrix that approaches a degenerate saddle point $(0, 1)$ and partitions the phase space into two regions: Points above the separatrix reach a zero density within a finite time, while points below it approach infinity. Similar conclusions can be drawn for the $n = 1$ case. Here, however, the right-half plane, $G > 0$, corresponds to positive density values, and the region below the separatrix, which approaches $(0, -3)$, pertains to points that reach zero density within a finite time. Varying C in (4.5) yields similar trajectories that are shifted along the G-axis and may cross the $G = 0$ line. Namely, in 2D cases trajectories may cross from positive to negative aphysical density spaces.

4.3. Similarity solutions for 3D radially symmetric flows. A general similarity solution for radially symmetric problems in three dimensions can be derived analogously. A phase-plane analysis of (4.2) for $m = 3$ yields the desired trajectories. Phase-maps are depicted in Figure 8 for typical values of n that validate the classification of point B (see Figure 5). Again, positive values of the density field correspond



(a) (b)

FIG. 7. *Phase maps of 2D flow cases for two values of n.* (a) $n = -3$, (b) $n = 1$. *Only one trajectory is shown. Other trajectories are identical to the one shown and are a parallel shift along the G-axis. For the particular chosen value C* (4.5) *the trajectory touches the G-axis.*

to positive values of G for $n > -2$ and negative values of G for $n < -2$. Notice that, similar to the 1D cases, no interaction between positive and negative density spaces is possible. Very intriguing phase-maps are obtained for negative densities, an obviously aphysical domain of the solution. A periodic solution is obtained for $n = 0$ (see Figure 8(e)); alas, it survives only if we allow negative densities.

5. Conclusions. Explicit nontrivial analytical solutions (3.10) and (3.14) are obtained for 1D ($n = -1$) and 3D ($n = -3$) initial density humps in liquid flow fields. The 1D solution shows that the initial density distribution C_1/x is convected without being diffused at very short times, then diffuses at intermediate times, and then is recovered at long times; the last effect is remarkable behavior of a solution for the nonlinear convection-diffusion equation. In case $C_2 = C_1^{-2}$ and $\alpha = 0$ the solution describes the propagation velocity of a density (temperature) disturbance in the positive half space with zero back pressure at $\hat{x} \to \infty$. The effect of infinite initial density at $x = 0$ is short-lived and can be neglected, quite similar to analyses which employ the Dirac delta function to describe particle concentrations at $t = 0$ in diffusion problems, etc. A comparison with Barenblatt's (1979) general analyses of similarity solutions reveals that the long time propagation velocity should depend on n and can be predicted quite simply by exploring the similarity parameter η. Namely, the density distribution can be expressed as a function of $\ln(\eta)$ or $\ln(\hat{x}) - (1/(n+2))\ln(\hat{t}) + c$. Thus, a propagation velocity in the $\ln(\hat{x})$, $\ln(\hat{t})$ space is equal to $1/(n+2)$. However, in the x, t domain the constant c is paramount and is generally unknown. We show that two essential propagation velocities exist for short and long times ($\hat{u}_P = 1/C_1$ and $\hat{u}_P = 1/C_1 - C_2C_1$, respectively) that cannot be predicted unless a full solution is obtained.

It is also interesting to note that the flow region near the contact line in a shock tube may asymptotically be compared with our simple 1D model despite the fact that the former deals with compressible fluid. The initial density distribution in a shock tube possesses two distinct regions, separated by a diaphragm. However, after the diaphragm is broken, the known theory predicts that *four* main regions exist and a

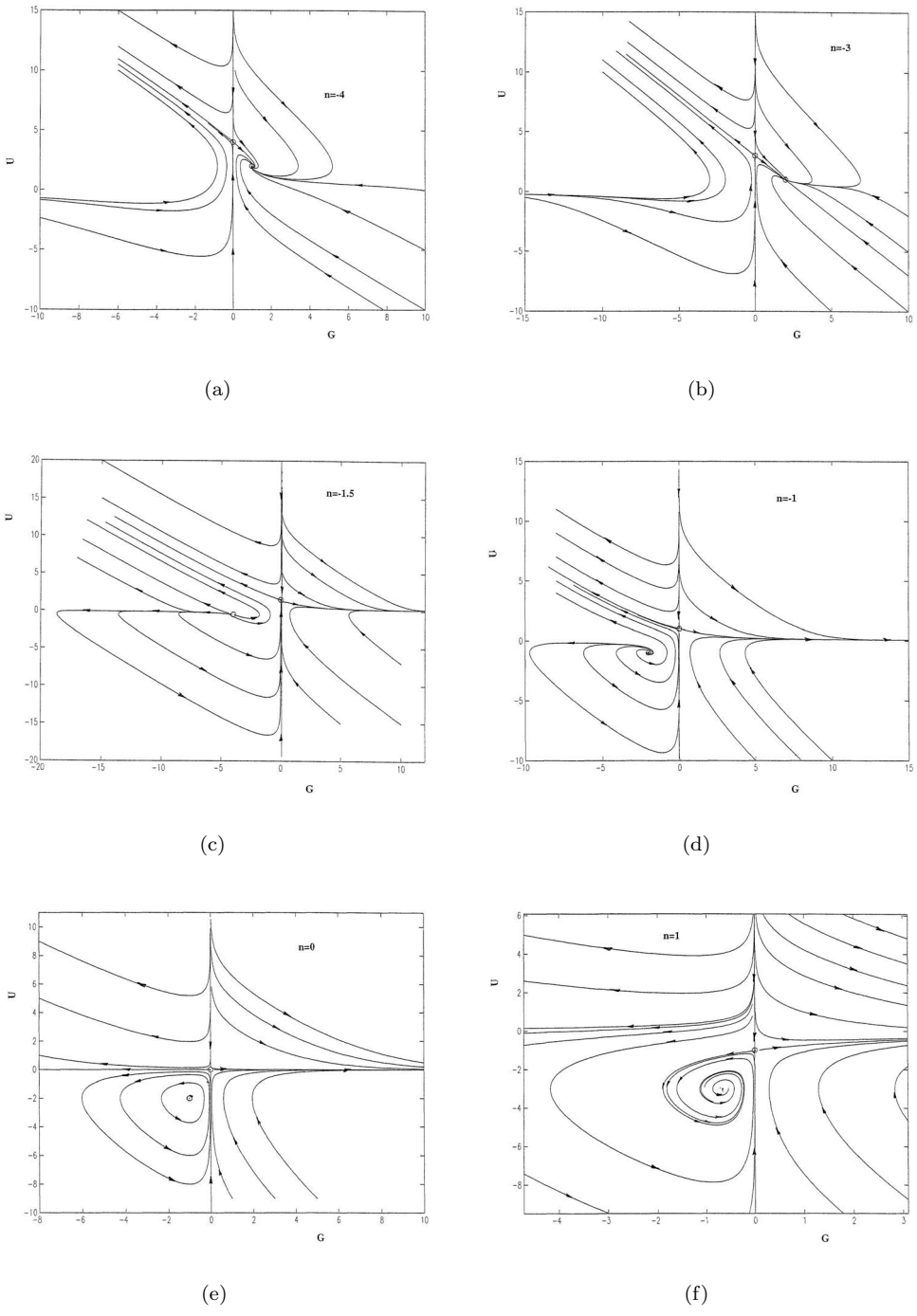


FIG. 8. *Phase maps for 3D flow cases for prototypical values of n.* (a) $n = -4$, *point B is a stable spiral;* (b) $n = -3$, *point B is a stable node;* (c) $n = -1.5$, *point B is an unstable node;* (d) $n = -1$, *point B is an unstable spiral;* (e) $n = 0$, *point B is a center;* (f) $n = 1$, *point B is a stable spiral. Positive density fields are obtained at the left-half plane in cases* (a), (b) *and the right-half plane in cases* (c), (d), (e), *and* (f). *Case* (b) *was solved analytically (see* (3.14)*).*

traveling distinct contact line (not to be confused with the pressure shock; see Shapiro (1954, p. 1007)) separates the two inner regions between the rarefaction and the pressure shock waves. In these two regions the pressure is constant, while the temperature and density fields are discontinuous across the contact line, the latter moving with a constant downstream velocity. Notice that, despite the fact that the fluid is compressible, within the two foregoing regions and for ideal gases, the temperature is inversely proportional to the density since the pressure is uniformly distributed. At the boundaries of these two regions (the rarefaction and the shock waves) a constant flow enters and exits the system. All of the foregoing boundary conditions are nearly satisfied by our 1D model if we assume that $P_r = 1$, keeping the pressure field uniform (see (3.12)). The initial density distribution is obviously different. However, the contact line is, in reality, a diffused region, and a sharp contact line that separates the two density fields is an approximate theoretical description of the field. In our 1D example we focus on the effect of an initially diffused density shock that may mimic the flow field within the two regions adjacent to the contact line. The analytical results of the 1D example show that a flow that possesses an initially diffused density profile has an additional feature; the traveling velocity of the density disturbance has different values for short and long times, unlike the case of a distinct contact line that possesses a unique constant traveling velocity.

The 3D solution exhibits extremely steep gradients in the density and pressure fields that resemble diffused shock waves. A prototypical solution describes a *closed insulated* system with an initial algebraic density hump, $C_1/\hat{r}^3$, that *implodes* within a finite time at a rate proportional to $1/\hat{t}^{\,2}$. Such an unexpected behavior may give rise to a rather interesting question: Are "black holes" possible in liquid systems satisfying the Navier–Stokes equations? In reality this unusual result is an artifact of the assumptions that the terms describing the energy dissipation and fluid compressibility can be omitted from the energy conservation equation. Similarly, a solution was derived for negative times that may provide an insight about the flow field near a diffused adiabatic interfacial region of an expanding bubble.

Similarity solutions for density fields that possess initially r^n (n-arbitrary) humps are also analyzed. The particular form of the governing set of two first order ordinary differential equations (4.2) proves that these similarity solutions can never be *chaotic*. Phase-maps illustrate the trajectories of such systems in 1D, 2D, and 3D flow fields. A formal analytical solution is also obtained for 2D axisymmetric flows (see (4.5), (4.6)). Of particular interest is the 3D ($n = 0$) periodic solution that is obtained for negative densities and mimics the well-known prey-predator problem. It is interesting to note that trajectories do not cross the $G = 0$ axis in 1D and 3D flow problems. Hence, positive and negative density spaces do not interact and are completely separated worlds. Also note that there is no qualitative difference between the different cases shown in Figure 8 for the half plane that pertains to positive (physical) densities, despite the fact that the critical point changes type. In 2D flow problems, trajectories may cross the $G = 0$ axis, and negative and positive density spaces may interact.

Finally, we suggest that (3.10) and (3.14) may also be used to define benchmark problems by which the accuracy of numerical solvers can be verified. One must, however, be cautious in employing the latter, since the stability of solutions (3.10) and (3.14) has not been investigated.

Acknowledgment. The author would like to thank Professor Howard Brenner from MIT for his useful comments.

REFERENCES

R. Aris (1962), *Vectors, Tensors and the Basic Equations of Fluid Mechanics*, Prentice–Hall, Englewood Cliffs, NJ.

G. I. Barenblatt (1979), *Similarity, Self Similarity, and Intermediate Asymptotics*, Consultants Bureau, New York.

C. M. Bender and S. A. Orszag (1978), *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York.

J. Camacho and H. Brenner (1995), *On convection induced by molecular diffusion*, Ind. Eng. Chem. Res., 34, pp. 3326–3335.

D. Canright and S. Morris (1993), *Buoyant instability of viscous film over a passive fluid*, J. Fluid Mech., 255, pp. 349–372.

D. M. Christopher and B. X. Wang (2001), *Similarity simulation of Marangoni convection around a vapor bubble during nucleation and growth*, Int. J. Heat Mass Trans., 44, pp. 799–810.

J. D. Cole (1951), *On a quasi-linear parabolic equation occurring in aerodynamics*, Quart. Appl. Math., 9, pp. 225–236.

G. F. Davies (1999), *Dynamic Earth, Plates, Plumes and Mantle Convection*, Cambridge University Press, Cambridge, UK.

G. D. Gerald (1971), *Introduction to Geophysics, Mantle Core and Crust*, W. B. Saunders, Philadelphia.

E. Hopf (1950), *The partial differential equation* $u_t + u_{xx} = \mu u_{xx}$, Comm. Pure Appl. Math., 3, pp. 201–230.

S. A. Koehler, H. A. Stone, M. P. Brenner, and J. Eggers (1998), *Dynamics of foam drainage*, Phys. Rev. E, 58, pp. 2097–2106.

L. A. Romero and F. G. Yost (1996), *Flow in open channel capillary*, J. Fluid Mech., 322, pp. 109–129.

A. H. Shapiro (1954), *The Dynamics and Thermodynamics of Compressible Fluid Flow*, The Ronald Press Company, New York.

G. B. Whitham (1974), *Linear and Nonlinear Waves*, John Wiley & Sons, New York.

T. P. Witelski and A. J. Bernoff (1999), *Stability of self-similar solutions for van der Waals driven thin film rupture*, Phys. Fluids, 11, pp. 2443–2445.

A. W. Woods and S. D. Fitzgerald (1997), *The vaporization of a liquid front moving through a hot porous rock. Part* 2. *Slow injection*, J. Fluid Mech., 343, pp. 303–316.

W. W. Zhang and J. R. Lister (1999), *Similarity solutions for van der Waals rupture of a thin film on a solid substrate*, Phys. Fluids, 11, pp. 2454–2462.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 977–1001

DISPERSIVE WAVE ATTENUATION DUE TO OROGRAPHIC FORCING*

JUAN CARLOS MUÑOZ GRAJALES† AND ANDRÉ NACHBIN‡

Abstract. The O'Doherty–Anstey (ODA) approximation was originally formulated in the seismological literature for acoustic pulse propagation through a disordered stratified medium [*Geophys. Prospecting*, 19 (1971), pp. 430–458]. It explains the mechanism for amplitude attenuation (and pulse shaping) promoted by the variable coefficient, conservative hyperbolic model. This work generalizes the one-dimensional ODA theory for linear weakly dispersive water waves forced by a disordered orography. The analysis is performed through the recently formulated terrain-following Boussinesq system. This is achieved by applying the invariant imbedding method. As a result, dispersion alters the medium's correlation function which controls the apparent attenuation mechanism. On the other hand, orography affects the dispersive mechanism for the Airy function–like formation. A nonlinear Boussinesq solver was implemented, and theoretical results were validated for different values of the parameters of interest. The theoretical results are in very good agreement with the small amplitude simulations. In particular, the approximate theory was able to capture a good part of the forward scattering radiation. Moreover, through numerical experiments the theory is pushed beyond its expected regime and captures the attenuation of small amplitude solitons due to orographic forcing.

Key words. dispersive waves, inhomogeneous media, asymptotic theory

AMS subject classifications. 76B07, 76B15, 35Q

DOI. 10.1137/S0036139902412769

1. Introduction.

Wave-topography interaction has been the subject of considerable mathematical research. The physical applications range from coastal surface waves [18] to atmospheric flows over mountain ranges [2, 7]. In particular, the interaction of waves with fine features of the topography is of great interest. As pointed out in the introduction to [7], the "representation ... of subgrid-scale orographic processes is recognized as crucial to numerical weather prediction at all time ranges." In the atmospheric literature *orography* implies mountain ranges [2]. Our study is therefore focused on the effect of small-scale orographic features, which we call the *microstructure*. A mathematical theory is described and its robustness validated numerically. As surface gravity waves propagate from deep to shallow waters, they are transformed due to shoaling, refraction, diffraction, and reflection. In order to concentrate on the main scattering mechanism connected with the pulse shaping phenomenon to be described, we consider the normal incidence of surface pulse shaped waves. These waves propagate over topographies containing a smooth slowly varying profile together with disordered small-scale features. Our goal is to capture the wave-microstructure interaction.

The main result is that the disordered medium fluctuations cause the propagating pulse to broaden as it travels. Due to multiple scattered energy, the pulse appears to

*Received by the editors August 6, 2002; accepted for publication (in revised form) September 25, 2003; published electronically March 30, 2004. This work was supported by CNPq/Brazil under grant 300368/96-8. Part of this work was supported by NSF grant DMS97-09320 through the Mathematical Geophysics Summer Schools (1998–2002) at Stanford University.

http://www.siam.org/journals/siap/64-3/41276.html

†Instituto de Matemática Pura e Aplicada, Est. D. Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (juanc@impa.br). Current address: Department of Mathematics, Universidad del Valle, A. A. 25360, Cali, Colombia.

‡Instituto de Matemática Pura e Aplicada, Est. D. Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (nachbin@impa.br).

diffuse about a moving center. The amount of broadening and attenuation is proportional to the traveling distance and depends on the disorder's correlation function. In what follows we refer to the transformation of the pulse due to the medium's microscale fluctuations as *pulse shaping*.

The theory for pulse shaping was originally derived in the context of acoustic wave propagation in the earth's crust [25]. It is known as the O'Doherty–Anstey (ODA) approximation. In the acoustic wave applications several authors have analyzed the spreading of a pulse due to microscale variations in the medium parameters (cf., for example, the work by Clouet and Fouque [9], Papanicolaou and Sølna [27], and Lewicki, Burridge, and de Hoop [15]). The motivation for modeling in terms of a random medium is that a detailed description of microscale medium fluctuations is often not known. Using a stochastic model, uncertainties about a specific medium are translated into uncertainties about a transmitted pulse shape in a systematic way [1]. Stochastic modeling has been used by Nachbin for long weakly dispersive surface wave problems [22, 24].

This work generalizes the one-dimensional ODA theory for linear weakly dispersive water waves when forced by a disordered orography. The analysis is performed through the recently formulated terrain-following Boussinesq system [23]. This is achieved by applying the invariant imbedding method. As a result, dispersion alters the medium's correlation function, which controls the apparent attenuation mechanism. On the other hand, orography affects the dispersive mechanism for the Airy function–like formation. A nonlinear Boussinesq solver was implemented, and theoretical results were validated for different values of the parameters of interest. Details of the numerical method and a larger range of experiments will be presented elsewhere [20]. The theoretical results presented here are in very good agreement with the small amplitude simulations. This amounts to solving the nonlinear Boussinesq system with data on a small amplitude-to-depth ratio. In particular, the approximate theory was able to capture a good part of the forward scattering radiation.

This paper is organized as follows. In section 2 we present the terrain-following Boussinesq system. In section 3 the pulse shaping ODA theory is formulated in detail. Section 4 contains five sets of numerical experiments validating the theory and generating further insight into it. The conclusions are given in section 5. A three-part appendix is intended to provide further detail for the reader.

2. The linearized terrain-following Boussinesq model. We start by presenting the potential theory formulation for Euler's equations with a free surface and an impermeable bottom topography. In the potential theory model the fluid is assumed to be inviscid, incompressible, and irrotational. Let variables with physical dimensions be denoted with a tilde. We introduce the length scales σ (a typical pulse width or wavelength), h_0 (a typical depth), a (a typical wave amplitude), ℓ (the horizontal length scale for bottom irregularities), and L (the total length of the rough region or the total propagation distance). The acceleration due to gravity is denoted by g, and the reference shallow water speed is $c_0 = \sqrt{gh_0}$. Dimensionless variables are then defined in a standard fashion [23, 29]:

$$\tilde{x} = \sigma x, \quad \tilde{y} = h_0\, y, \quad \tilde{t} = \left(\frac{\sigma}{c_0}\right) t, \quad \tilde{\eta} = a\, \eta, \quad \tilde{\phi} = \left(\frac{g\sigma a}{c_0}\right) \phi, \quad \tilde{h} = h_0\, H\left(\frac{\tilde{x}}{\ell}\right).$$

The velocity potential $\phi(x, y, t)$ and wave elevation $\eta(x, t)$ satisfy the dimensionless equations (see [29])

$$\beta\, \phi_{xx} + \phi_{yy} = 0 \quad \text{for} \quad -H(x/\gamma) < y < \alpha\eta(x, t),$$

with the nonlinear free surface conditions

$$\eta_t + \alpha\phi_x\eta_x - \frac{1}{\beta}\phi_y = 0,$$

$$\eta + \phi_t + \frac{\alpha}{2}\left(\phi_x^2 + \frac{1}{\beta}\phi_y^2\right) = 0$$

at $y = \alpha\eta(x,t)$. The parameter $\alpha = a/h_0$ controls the strength of nonlinear effects, and $\beta = h_0^2/\sigma^2$ the level of dispersion. The length scale for the bottom inhomogeneities ℓ leads to the dimensionless parameter $\gamma = \ell/\sigma$. The Neumann condition at the impermeable bottom is

$$\phi_y + \frac{\beta}{\gamma}\, H'(x/\gamma)\phi_x = 0.$$

The bottom topography is described by $y = -H(x/\gamma)$, where

$$H(x/\gamma) \;=\; \begin{cases} 1 + n(x/\gamma) & \text{when } \; 0 < x < L, \\ 1 & \text{when } \; x \le 0 \;\text{ or }\; x \ge L. \end{cases}$$

The bottom profile is described by the function $-n(x/\gamma)$. The topography is rapidly varying when $\gamma \ll 1$. The undisturbed depth is given by $y = -1$, and the topography can be of large amplitude, provided that $|n| < 1$. We do not need to assume that the fluctuations n are small, nor continuous, nor slowly varying.

In past years Boussinesq-type equations have been employed to model surface wave propagation in shallow channels. Such models are weakly nonlinear, weakly dispersive approximations to the full potential theory equations. Peregrine [26] in 1967 deduced a model valid for channels of slowly varying depth, whereas Hamilton [14] in 1977 derived a set of equations valid for arbitrary orographies. The latter used an appropriate curvilinear coordinate system and applied a perturbation approach to the linear potential theory equations retaining the lowest-order effects of dispersion. The change of variables uses a conformal mapping from a strip in the complex $(\xi, \tilde{\zeta})$-plane to the complex z-plane ($z = x + i\tilde{y}$; $\tilde{y} \equiv \beta^{1/2} y$) (see [23]). Hamilton also outlines the formulation of a weakly nonlinear model by considering a Lagragian functional. In this direction, using dimensionless variables and Hamilton's conformal mapping strategy, the potential theory equations are first cast in a fixed orthogonal curvilinear coordinates $(\xi, \tilde{\zeta})$ as (see [23])

$$\phi_{\xi\xi} \;+\; \phi_{\tilde{\zeta}\tilde{\zeta}} \;=\; 0, \qquad\qquad -\sqrt{\beta} < \tilde{\zeta} < \alpha\sqrt{\beta}N(\xi,t), \tag{2.1}$$

with free surface conditions

$$|J|N_t + \alpha\phi_\xi N_\xi - \frac{1}{\sqrt{\beta}}\phi_{\tilde{\zeta}} = 0 \tag{2.2}$$

and

$$\phi_t + \eta + \frac{\alpha}{2|J|}\left(\phi_\xi^2 + \phi_{\tilde{\zeta}}^2\right) = 0 \tag{2.3}$$

at $\tilde{\zeta} = \alpha\sqrt{\beta}N(\xi,t)$. The bottom condition is

$$\phi_{\tilde{\zeta}} = 0 \qquad \text{at } \; \tilde{\zeta} = -\sqrt{\beta}. \tag{2.4}$$

The Jacobian for the $(\xi, \tilde{\zeta}) \to (x, \tilde{y})$ coordinate transformation is represented by $|J|$. Details can be found in Nachbin [23], where the potential theory equations given above are approximated as the ($O(\alpha)$, $O(\beta)$: respectively, weakly nonlinear, weakly dispersive) terrain-following Boussinesq system

$$M(\xi)\eta_t + \left(\left(1 + \frac{\alpha\eta}{M(\xi)}\right)u\right)_\xi = 0,$$
$$u_t + \eta_\xi + \frac{\alpha}{2}\left(\frac{u^2}{M(\xi)^2}\right)_\xi - \frac{\beta}{3}u_{\xi\xi t} = 0.$$

The metric term $M(\xi)$ is defined by

$$M(\xi) \equiv \tilde{y}_{\tilde{\zeta}}(\xi, 0).$$

When $\alpha = 0$ the system above reduces to

$$M(\xi)\eta_t + u_\xi = 0, \tag{2.5}$$
$$u_t + \eta_\xi - \frac{\beta}{3}u_{\xi\xi t} = 0.$$

Here ξ and t are the space and time coordinates, respectively; $\eta = \eta(\xi, t)$ is the wave elevation; and $u = u(\xi, t)$ is the weighted depth-averaged terrain-following velocity [23]. The variable coefficient $M(\xi)$ is a smooth orography-dependent function which appears as a consequence of the changes of variables [23]. When the parameter β is small, we are in the shallow channel/long wave regime. Equivalently, this is called the weakly dispersive regime.

In the next section we will analyze the linear ($\alpha = 0$) Boussinesq system for a large range of dispersive effects, expressed through the parameter β. It is very important to note that for $\beta > 0.25$ we are in the *deep water* regime. The particle-orbits decay exponentially with depth as shown in Dean and Dalrymple [5, Chapter 4, on linear potential theory]. Hence in this regime there is no wave-topography interaction. Thus the fact that the Boussinesq system is an $O(\beta)$ approximation to the potential theory is not much of a limitation for dispersive wave-topography interaction. Moreover, for large time intervals dispersive effects will be strongly displayed in the solutions even for small values of β (cf. Appendix A). Even though the theory we have developed is linear, in our numerical experiments we will consider nontrivial values of α, as will be discussed in the corresponding subsections.

3. The linear pulse shaping theory. In this section we generalize the ODA approximation from linear acoustic waves to the linear weakly dispersive system (2.5). The weakly dispersive Boussinesq system is forced by a rapidly varying orography expressed through the variable coefficient $M(\xi)$. We consider a technique analogous to Berlyand and Burridge's acoustic work [3], which we apply successfully to dispersive waves for two reasons. First, system (2.5) can be written equivalently as two coupled KdV-type equations for the transmitted and reflected fields. These linearized KdV-type equations can be viewed as a dispersive perturbation to those obtained by Berlyand and Burridge; namely, a propagating pulse will slowly disperse with a given (known) rate (cf. Appendix A, (A.4)). Second, the pulse's effective propagation velocity is bounded, and all Fourier modes have positive phase speeds bounded by 1. These two properties described allow us to apply the *invariant imbedding approach* [19], taking the propagation distance as the *imbedding parameter*.

In what follows, solutions u, η to (2.5) will be assumed to be smooth enough and absolutely integrable. This requirement is necessary to justify the use of the Fourier transform and the Fourier inversion formula.

We start by performing the change of variables

$$x = \int_0^{\xi} \frac{1}{C_o(s)} ds, \quad t' = t$$

in system (2.5). Here $C_o(s) = \sqrt{1/M(s)}$ is the *local wave speed* and x is the *travel time*. Dropping the primes, system (2.5) becomes

$$\begin{aligned} & C_o^{-1/2}\, \eta_t + C_o^{1/2}\, u_x = 0, \\ & C_o^{1/2}\, u_t + C_o^{-1/2}\, \eta_x - \frac{\beta}{3C_o^{1/2}} \left(u_{xxt}\, \frac{1}{C_o} + \left(\frac{1}{C_o} \right)_x u_{xt} \right) = 0. \end{aligned} \tag{3.1}$$

We adopt the *wave mode splitting*

$$R = C_o^{1/2} u + C_o^{-1/2} \eta, \tag{3.2}$$

$$L = -C_o^{1/2} u + C_o^{-1/2} \eta. \tag{3.3}$$

Differentiating (3.2) and (3.3) with respect to x and t and using (3.1) and the fact that $u = (R - L)/(2C_o^{1/2})$ leads to the coupled wave mode system

$$\begin{aligned} R_t + R_x &= p(x) \left(\frac{R - L}{2C_o^{1/2}} \right)_{xxt} + q(x) \left(\frac{R - L}{2C_o^{1/2}} \right)_{xt} - r(x) L, \\ L_x - L_t &= p(x) \left(\frac{R - L}{2C_o^{1/2}} \right)_{xxt} + q(x) \left(\frac{R - L}{2C_o^{1/2}} \right)_{xt} - r(x) R, \end{aligned} \tag{3.4}$$

where the variable coefficients are

$$p(x) = \frac{\beta}{3C_o^{3/2}(x)}, \quad q(x) = -\frac{\beta C_{o,x}(x)}{3C_o^{5/2}(x)}, \quad \text{and} \quad r(x) = \frac{C_{o,x}(x)}{2C_o(x)}.$$

The initial conditions for system (3.4) are

$$R(x, 0) = R_o(x), \quad L(x, 0) = 0.$$

Note that when the bottom is flat ($r \equiv 0$) and $\beta = 0$ (no dispersion), equations (3.4) identify R with a wave propagating to the right (transmitted wave) and L with a wave propagating to the left (reflected wave). For variable depths we adopt the same terminology. Moreover, this terminology is also consistent for system (3.4), in the weakly dispersive regime ($0 < \beta \ll 1$) and $r \equiv 0$, since the left-propagating signal L is negligible if the initial data corresponds to a right-going wave (see Appendix A).

Several decompositions of the Boussinesq equations into a pair of KdV equations were introduced by Mattioli in [16] and [17]. We point out that system (3.4) has some advantages with respect to those decompositions. First, unlike system (3.4), the Boussinesq equations used by Mattioli are not valid as an asymptotic approximation of the potential theory equations for arbitrary rapidly varying or nondifferentiable orographies. Second, in contrast to system (3.4), the linear dispersion relation for Mattioli's model results in unstable short waves with amplitude tending to infinity.

In order to set our forthcoming results in a wave propagating frame let us introduce the change of variables

$$\tau = t - x, \quad x' = x,$$

where τ is the time-delay variable. Again abandoning the primes, equations (3.4) become

$$\begin{aligned} R_x = {} & p\left(\frac{R-L}{2C_o^{1/2}}\right)_{\tau\tau\tau} - 2p\left(\frac{R-L}{2C_o^{1/2}}\right)_{x\tau\tau} + p\left(\frac{R-L}{2C_o^{1/2}}\right)_{xx\tau} \\ & - q\left(\frac{R-L}{2C_o^{1/2}}\right)_{\tau\tau} + q\left(\frac{R-L}{2C_o^{1/2}}\right)_{x\tau} - rL, \\ L_x - 2L_\tau = {} & p\left(\frac{R-L}{2C_o^{1/2}}\right)_{\tau\tau\tau} - 2p\left(\frac{R-L}{2C_o^{1/2}}\right)_{x\tau\tau} + p\left(\frac{R-L}{2C_o^{1/2}}\right)_{xx\tau} \\ & - q\left(\frac{R-L}{2C_o^{1/2}}\right)_{\tau\tau} + q\left(\frac{R-L}{2C_o^{1/2}}\right)_{x\tau} - rR. \end{aligned} \tag{3.5}$$

Let $\widehat{f}(\omega)$ denote the Fourier transform of $f(\tau)$ in the time-delay variable τ:

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} e^{-i\omega\tau} f(\tau)\, d\tau.$$

Take the Fourier transform in τ of (3.5). Manipulate the resulting system in order to obtain first-order equations for the travel time-evolution of the Fourier modes. This goal is achieved by first subtracting the two equations giving

$$\widehat{R}_x - \widehat{L}_x = r\widehat{R} - 2i\omega\widehat{L} - r\widehat{L}. \tag{3.6}$$

Moreover, by differentiating this equation with respect to x, we find

$$\begin{aligned} \widehat{R}_{xx} - \widehat{L}_{xx} = {} & r'(x)\widehat{R} - r'(x)\widehat{L} - 2iw\widehat{R}_x + 2iwr(x)\widehat{R} \\ & + 4w^2\widehat{L} - 4iwr(x)\widehat{L} + r^2(x)\widehat{R} - r^2(x)\widehat{L}. \end{aligned} \tag{3.7}$$

This expression can be used to eliminate the second order x-derivatives in the first Fourier transformed equation arising from (3.5). This is a crucial step in order to apply the invariant imbedding approach, which requires a first-order system of ordinary differential equations (ODE). Using the fact that the time frequency range $|\omega| < C_o(x)\sqrt{3/\beta}$ (see Appendix A), we solve for $\widehat{R}_x$ to obtain

$$\widehat{R}_x = \zeta(x,\omega)\widehat{L} + \gamma(x,\omega)\widehat{R}, \tag{3.8}$$

where

$$\zeta(x,\omega) = \frac{id(x)\omega^3 + e(x)\omega^2 - 2r(x)}{2(1 - d(x)\ \omega^2)}, \quad \gamma(x,\omega) = \frac{\frac{-i\beta\omega^2}{3C_o^2(x)}(\omega - 4ir(x))}{2\left(1 - \frac{\beta\omega^2}{3C_o^2(x)}\right)},$$

$$d(x) = \frac{\beta}{3C_o^2(x)} \quad \text{and} \quad e(x) = -\frac{\beta}{3C_o^3(x)}C_{o,x}(x).$$

Putting (3.8) into (3.6) and solving for $\widehat{L}_x$ results in

$$\widehat{L}_x = (2i\omega + \bar{\gamma}(x,\omega))\widehat{L} + \bar{\zeta}(x,\omega)\widehat{R}, \tag{3.9}$$

where $\bar{\zeta}(x,w)$ and $\bar{\gamma}(x,w)$ denote the complex conjugates of $\zeta(x,w)$ and $\gamma(x,w)$, respectively. As claimed above, (3.8) and (3.9) give the evolution in x (travel time) of each Fourier mode, corresponding to the transmitted and reflected fields.

Next, suppose that we want to calculate $L(x_o,\tau) = L(x_o,\tau;T)$ for $0 \le \tau \le T$. At this step we apply the *invariant imbedding approach* [19] to system (3.8) and (3.9). To do so we imbed the relevant inhomogeneous region inside a homogeneous medium so that we can give appropriate boundary conditions at the border of the medium's slab $[x_o, x_o + X]$, say

$$\widehat{R}(x_o,\omega;T) = \widehat{h}(\omega), \tag{3.10}$$

$$\widehat{L}(x_o + X,\omega;T) = 0. \tag{3.11}$$

Boundary condition (3.11) means that no reflection is expected at the downstream travel time location x_o+X when $0 < \tau = t - x_o < T$. In other words, the signal did not have enough time to arrive at this point and to produce a medium's response. This is true at least for some $X > 0$ large enough (depending on T) due to the pulse's finite (effective) velocity mentioned earlier. Linearity of ODE system (3.8)–(3.9) and the invariant imbedding technique [19] guarantees the existence of a function $K(x,\omega;T)$ (called the *reflection kernel*) such that

$$\widehat{L}(x,\omega;T) = \widehat{K}(x,\omega;T)\ \widehat{R}(x,\omega;T) \tag{3.12}$$

and satisfying the Riccati-type equation

$$\widehat{K}_x = \bar{\zeta}(x,\omega) + 2i\omega\Gamma(x,\omega)\widehat{K} - \zeta(x,\omega)\widehat{K}^2, \tag{3.13}$$

with

$$\Gamma(x,\omega) = \frac{2 - d(x)\omega^2}{2(1 - d(x)\omega^2)}.$$

From (3.11) we obtain an appropriate initial condition for (3.13):

$$\widehat{K}(x_o + X,\omega;T) = 0. \tag{3.14}$$

Notice that (3.12) allows us to solve for the reflected signal $L(x_o,\tau;T)$ in terms of the reflection kernel $K(x_o,\tau;T)$ and the transmitted pulse $R(x_o,\tau;T)$ for $0 \le \tau \le T$. This function K contains all the information about medium's reflection properties. We also remark that the boundary value problem (3.8), (3.9), (3.10), (3.11) has been reduced to solving the initial value problem (3.13)–(3.14) in reversed direction from the travel time location $x_o + X$ up to the time of interest x_o.

In general it is not possible to solve explicitly for $\widehat{K}$ from (3.13), so we adopt an approximation. To this end we present the following generalization to the lemma given by Berlyand and Burridge [3].

LEMMA 3.1. *Let y satisfy the Riccati equation*

$$y'(s) = -A(s) - 2i\omega B(s)y(s) + \bar{A}(s)y^2(s), \quad 0 \le s \le s_o, \tag{3.15}$$

subject to the initial condition $y(0) = 0$. *Here* $\bar{A}(s)$ *denotes the complex conjugate of* $A(s)$. *Hence the solution can be expressed as*

$$y(s_o) = -\int_0^{s_o} e^{-2i\omega \int_{s'}^{s_o} B(\xi)d\xi} A(s')ds' + E(s_o), \tag{3.16}$$

where the error term is

$$E(s_o) = \int_0^{s_o} \bar{A}(s') e^{-2i\omega \int_{s'}^{s_o} B(\xi)d\xi} y^2(s')ds'.$$

Set

$$v(s_o) = \sup_{0\le s\le s_o} \left| \int_0^s e^{2iw \int_0^{s'} B(\xi)d\xi} A(s')ds' \right| \quad \textit{and} \quad V(s) = \int_0^s |A(s')|\, ds'.$$

If $v(s_o)V(s) < 1$, *then*

$$|E(s)| \le \frac{v^2(s_o)V(s)}{1 - v(s_o)V(s)}, \quad 0 \le s \le s_o. \tag{3.17}$$

Proof. Multiplying (3.15) by its integrating factor and integrating both sides from 0 to s yields

$$\begin{aligned} &\int_0^s e^{2i\omega \int_0^{s'} B(\xi)d\xi} y'(s')ds' + 2iw \int_0^s B(s') e^{2i\omega \int_0^{s'} B(\xi)d\xi} y(s')ds' \\ &= -\int_0^s e^{2iw \int_0^{s'} B(\xi)d\xi} A(s')ds' + \int_0^s \bar{A}(s') e^{2i\omega \int_0^{s'} B(\xi)d\xi} y^2(s')ds'. \end{aligned}$$

Consequently,

$$\begin{aligned} &\int_0^s \frac{d}{ds'} \left(e^{2i\omega \int_0^{s'} B(\xi)d\xi} y(s') \right) ds' = e^{2i\omega \int_0^s B(\xi)d\xi} y(s) \\ &= -\int_0^s e^{2i\omega \int_0^{s'} B(\xi)d\xi} A(s')ds' + \int_0^s \bar{A}(s') e^{2i\omega \int_0^{s'} B(\xi)d\xi} y^2(s')ds'. \end{aligned} \tag{3.18}$$

Now solving for $y(s)$ and evaluating at $s = s_o$, in (3.18) we arrive at (3.16).

To achieve estimate (3.17) we start with

$$\begin{aligned} E(s) &= \int_0^s \bar{A}(s') e^{-2i\omega \int_{s'}^{s} B(\xi)d\xi} y^2(s')ds' \\ &= e^{-2i\omega \int_0^s B(\xi)d\xi} \int_0^s \bar{A}(s') e^{2i\omega \int_0^{s'} B(\xi)d\xi} y^2(s')ds'. \end{aligned}$$

Let

$$Y(s) = \int_0^s |A(s')|\, |y(s')|^2\, ds'. \tag{3.19}$$

Hence

$$|E(s)| = \left| \int_0^s \bar{A}(s') e^{2i\omega \int_0^{s'} B(\xi)d\xi} y^2(s')ds' \right| \le Y(s). \tag{3.20}$$

Differentiating (3.19) with respect to s,

$$Y'(s) = |A(s)|\,|y(s)|^2. \tag{3.21}$$

Note that from (3.16) and (3.20)

$$|y(s)| \le v(s_o) + Y(s), \quad 0 \le s \le s_o. \tag{3.22}$$

Thus substituting (3.22) into (3.21) leads to

$$Y'(s) \le |A(s)|\,(v(s_o) + Y(s))^2. \tag{3.23}$$

By integrating (3.23) from 0 to s, we obtain

$$\int_0^s \frac{Y'(s)}{(v(s_o) + Y(s'))^2} ds' \le \int_0^s |A(s')|\, ds' = V(s).$$

Therefore,

$$\frac{1}{v(s_o)} - \frac{1}{v(s_o) + Y(s)} \le V(s).$$

Thus if $v(s_o)V(s) < 1$, equation (3.20) gives

$$|E(s)| \le Y(s) \le \frac{v^2(s_o)V(s)}{1 - v(s_o)V(s)}, \quad 0 \le s \le s_o.$$

This concludes the lemma's proof. □

We now apply these results to our Riccati equation. For brevity the argument T will be omitted in what follows. We apply the lemma above to the reflection kernel's problem (3.13)–(3.14) by letting $y(s) = \widehat{K}(x(s), \omega)$, $A(s) = \bar{\zeta}(x(s), \omega)$, $B(s) = \Gamma(x(s), \omega)$, and $s = x_o + X - x$, $s_o = X$. We deduce that

$$\widehat{K}(x, \omega) = -\int_x^{x+X} e^{2i\omega \int_{x'}^{x} \Gamma(s,\omega)\, ds} \bar{\zeta}(x', \omega) dx' + E(x, \omega), \tag{3.24}$$

with an error term given by

$$E(x, \omega) = \int_x^{x+X} e^{2i\omega \int_{x'}^{x} \Gamma(s,\omega)\, ds} \zeta(x', \omega) \widehat{K}^2(x', \omega) dx'.$$

Also

$$v(x_o, \omega) = \sup_{x_o \le x \le x_o + X} \left| \int_x^{x_o+X} e^{2i\omega \int_{x'}^{x_o+X} \Gamma(s,\omega)\, ds} \bar{\zeta}(x', \omega) dx' \right|$$

and

$$V(x_o, \omega) = \int_{x_o}^{x_o+X} |\zeta(x', \omega)|\, dx'.$$

Then, in the cases where $v(x_o, \omega)V(x_o, \omega) < 1$, the error bound follows.

Now we solve for $\widehat{R}(x_o,\omega)$ at an arbitrary point x_o by using (3.24) and (3.12) in (3.8). It results that

$$\widehat{R}_x(x,\omega)=\left(\zeta(x,\omega)\left(-\int_x^{x+X} e^{2i\omega\int_{x'}^{x}\Gamma(s,\omega)\,ds}\bar{\zeta}(x',\omega)dx'+E(x,\omega)\right)+\gamma\right)\ \widehat{R}(x,\omega).$$

Solving this initial value problem with $\widehat{R}(0,\omega)=\widehat{f}(\omega)$, then

$$\widehat{R}(x_o,\omega)=\widehat{f}(\omega)\ \exp\left[-\int_0^{x_o}\int_x^{x+X}\zeta(x,\omega)e^{2i\omega\ \int_{x'}^{x}\Gamma(s,\omega)\,ds}\ \bar{\zeta}(x',\omega)dx'+\gamma(x,\omega)\ dx\right]$$

$$\cdot\exp\left[\int_0^{x_o}\zeta(x,\omega)\ E(x,\omega)\,dx\right]. \tag{3.25}$$

For general rapidly varying orographies and $\beta\ll 1$ the error function $E(x,\omega)$ is small (see Appendix C), even though the error estimate (3.17) is not useful in this case. Estimate (3.17) is sharp, for instance, when the function describing the medium properties (the metric coefficient $M(x)$ in system (2.5)) is taken to be piecewise constant on travel-time intervals with equal length (called a *Goupillaud medium*) [21].

Therefore, (3.25) leads to the *generalized ODA approximation*

$$\widehat{R}(x_o,\omega)\approx\widehat{f}(\omega)e^{-x_o(a_\beta(x_o,\omega)+b(x_o,\omega))}, \tag{3.26}$$

where

$$a_\beta(x_o,\omega)=\frac{1}{x_o}\int_0^{x_o}\int_x^{x+X}\zeta(x,\omega)e^{2i\omega\int_{x'}^{x}\Gamma(s,\omega)\,ds}\bar{\zeta}(x',\omega)dx'\,dx, \tag{3.27}$$

$$b(x_o,\omega)=-\frac{1}{x_o}\int_0^{x_o}\gamma(x,\omega)\,dx. \tag{3.28}$$

Using approximation (3.26), we can obtain an expression for $R(x_o,\tau)$ (the transmitted field in our applications) by using the Fourier inversion formula. Thus for $0\le\tau\le T$

$$R(x_o,\tau)\approx\frac{1}{2\pi}\int_{-\infty}^{\infty}\widehat{f}(\omega)e^{-x_o(a_\beta(x_o,\omega)+b(x_o,\omega))}e^{i\omega\tau}d\omega. \tag{3.29}$$

Note that the exponential factor in (3.29) accounts for the wave attenuation at travel time x_o. As mentioned above, the error estimate given above is not always sharp. For this reason the error term $E(x_o,\omega)$ is calculated numerically in Appendix C. Furthermore, for dispersion parameter β small enough and constant depth ($r(x)\equiv 0$) we have that $a_\beta=O(\beta^2)$ and $b=O(\beta)$. Hence for variable depths it is reasonable to approximate the medium's dispersive correlation function $a_\beta(x_o,\omega)$ by the medium's hyperbolic correlation function $a_0(x_o,\omega)$; that is,

$$a_\beta(x_o,\omega)\approx\frac{1}{x_o}\int_0^{x_o}\int_x^{x+X}r(x)e^{2i\omega(x-x')}r(x')dx'\ dx\equiv a_0(x_o,\omega). \tag{3.30}$$

This approximation, valid for small values of β, is not a limitation of the theory but enables numerical efficiency in the evaluation of the Fourier integrals. Note that this hyperbolic version is easier to compute than the dispersive formula (3.27). In the

latter the function ζ depends on the frequency ω. Therefore the Fourier-type integral in a_β cannot be computed rapidly by using the FFT (fast Fourier transform). On the other hand, the integral in (3.30) can be easily evaluated by the FFT algorithm because in this case the coefficient $r(x)$ is frequency-independent. In Appendix B we detail the numerical computation of the Fourier-type integral in coefficient $a_\beta(x_o, \omega)$. The accuracy of approximation (3.30) will be verified in section 4 showing that the leading-order dispersive effects are controlled by coefficient $b(x_o, \omega)$.

The numerical experiments in section 4 will show that approximation (3.29) captures very well not only the wave front but also part of the forward scattering radiation. The parameter X in (3.27) (or (3.30)) regulates to what extent the incoherent signal is recovered.

If $\beta = 0$, then $b(x_o, \omega) \equiv 0$, and the approximation given by (3.26) reduces to

$$\widehat{R}(x_o, \omega) \approx \widehat{f}(\omega) e^{-x_o a_0(x_o, \omega)}, \tag{3.31}$$

where

$$a_0(x_o, \omega) = \frac{1}{x_o} \int_0^{x_o} \int_x^{x+X} r(x) e^{2i\omega(x-x')} r(x') dx' \, dx,$$

as in Berlyand and Burridge's work [3]. We recall that when $\beta \neq 0$, the hyperbolic medium's correlation function $a_0(x_o, \omega)$ (which controls the attenuation mechanism) is altered and an extra attenuation term $b(x_o, \omega)$ appears due to the model's dispersion.

4. Numerical validation of the generalized ODA theory. Consider the nonlinear terrain-following Boussinesq system deduced by Nachbin [23],

$$\begin{aligned} &M(\xi)\eta_t + \left(\left(1 + \frac{\alpha\eta}{M(\xi)} \right) u \right)_\xi = 0, \\ &u_t + \eta_\xi + \frac{\alpha}{2} \left(\frac{u^2}{M(\xi)^2} \right)_\xi - \frac{\beta}{3} u_{\xi\xi t} = 0. \end{aligned} \tag{4.1}$$

We recall that when $\alpha = 0$, system (4.1) reduces to (2.5). Nevertheless, for our numerical validation experiments we will use the nonlinear Boussinesq system in a small α regime. Two types of initial data are considered. We study the propagation of Gaussian-shaped disturbances of the form

$$u(\xi, 0) = \eta(\xi, 0) = e^{-(\xi - \xi_o)^2/\epsilon},$$

where the parameter ξ_o controls the pulse's initial position and $\epsilon > 0$ its effective width. Furthermore, we also consider solitary waves of the form

$$\begin{aligned} \eta(\xi, 0) &= A_1 \, \mathrm{sech}^2(B(\xi - \xi_o)) + A_2 \, \mathrm{sech}^4(B(\xi - \xi_o)), \\ u(\xi, 0) &= A \, \mathrm{sech}^2(B(\xi - \xi_o)), \end{aligned}$$

with A_1, A_2, A, and B constants (to be defined in Experiment 2). These are approximate solutions to system (4.1) with $M \equiv 1$ (see [28]).

Except in some special cases (for instance, when $M \equiv 1, \alpha = 0$, or $\beta = 0$), finding the solution of system (4.1) is a nontrivial problem. To solve system (4.1) numerically we will use a finite difference solver introduced by Wei and Kirby [28]. This scheme will be used to perform numerical experiments in order to validate the theory developed in the previous sections.

First we rewrite system (4.1) in a more convenient way, as

$$\begin{aligned} \eta_t &= E(u,\eta), \\ V_t &= F(u,\eta), \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} E(u,\eta) &= -\frac{1}{M(\xi)}\left(\left(1+\frac{\alpha\eta}{M(\xi)}\right)u\right)_\xi, \\ F(u,\eta) &= -\eta_\xi - \frac{\alpha}{2}\left(\frac{u^2}{M(\xi)^2}\right)_\xi, \end{aligned} \tag{4.3}$$

and V is an intermediate variable defined by

$$V = u - \frac{\beta}{3}u_{\xi\xi}.$$

To evaluate the boundary values we use the radiation conditions by Engquist and Majda (see [8])

$$\begin{aligned} u_t - u_\xi &= 0 \quad \text{at} \quad \xi = \xi_{\min}, \\ u_t + u_\xi &= 0 \quad \text{at} \quad \xi = \xi_{\max}, \end{aligned} \tag{4.4}$$

where $\xi_{\min}$ and $\xi_{\max}$ denote the left and right ends, respectively, of the computational domain.

The evolution system above is solved by using an efficient predictor-corrector scheme. The $V \to u$ change of variables is done with an efficient (tridiagonal) ODE numerical scheme. Details, numerical properties, and further numerical experiments are presented elsewhere [20, 21].

We now describe several experiments validating the dispersive pulse-shaping ODA theory.

Experiment 1 (Flat channel and effectively linear regime). In these experiments the pulse is assumed to propagate over a flat bottom ($M \equiv 1$). In Appendix A (cf. (A.6) and (A.7)) we show that if, in addition, $\alpha = 0$, we can explicitly solve system (4.1) by using the Fourier transform technique. Flat channel solutions are employed to verify the numerical method's accuracy regarding dispersive and stability properties.

In Figure 4.1 we see that the exact solution η (for $\alpha = 0$, $\beta = 0.03$) and the numerical solution (for $\alpha = 0.001$, $\beta = 0.03$) are in very good agreement at $t = 40$. After 40 length-units into the flat channel, the right-propagating Gaussian has developed an Airy-like oscillatory tail (cf. Appendix A). The dispersive properties of the numerical scheme are very good.

For this experiment the numerical parameters are $J = 4000$ (spatial mesh points; $\Delta\xi = 0.0125$) and $N = 5000$ (time mesh points; $\Delta t = 0.008$). As mentioned above, this test shows that the code is capturing the (effectively linear) dispersive regime very well when α is small enough.

We repeat the experiment above for $\alpha = 0.001$, $\beta = 0.0005$, and $t = 40$. Dispersion has been decreased substantially. Now the effective hyperbolic regime is clearly observed in Figure 4.2.

Radiation conditions (4.4) also proved to absorb appropriately waves leaving the computational domain. It was observed that reflected waves produced by the computational boundaries were negligible.

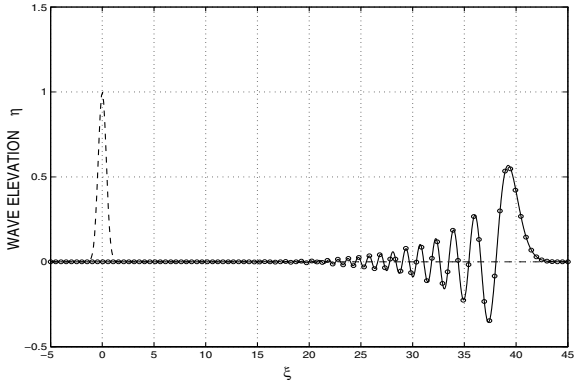


FIG. 4.1. *Dashed line: initial pulse,* $\eta(\xi,0) = u(\xi,0) = e^{-\xi^2/0.3}$. *Solid line: numerical solution for* $\alpha = 0.001$, $\beta = 0.03$, *and* $t = 40$. *Open circles: exact solution.*

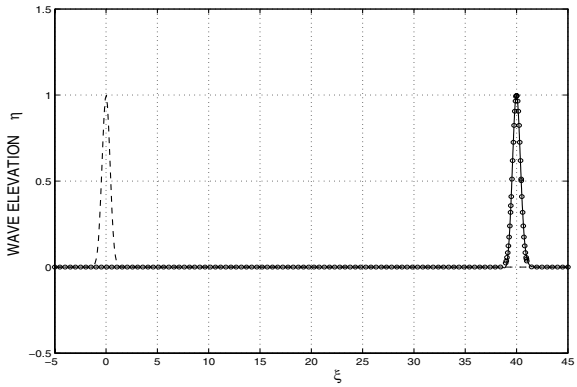


FIG. 4.2. *Dashed line: initial disturbance,* $\eta(\xi,0) = u(\xi,0) = e^{-\xi^2/0.3}$. *Solid line: numerical solution for* $\alpha = 0.001$, $\beta = 0.0005$, *and* $t = 40$. *Open circles: initial Gaussian pulse translated to* $\xi = 40$.

Experiment 2 (Flat channel and weakly nonlinear regime/solitary wave). We now study system (4.1) in the case that $M \equiv 1$, $0 < \alpha \ll 1$ (weakly nonlinear regime), and $0 < \beta \ll 1$ (weakly dispersive regime). Under these hypotheses it is possible to obtain an approximate solution of system (4.1) which has the analytical form

$$\begin{aligned} \eta(\xi,t) &= A_1 \operatorname{sech}^2(B(\xi - Ct - \xi_o)) + A_2 \operatorname{sech}^4(B(\xi - Ct - \xi_o)), \\ u(\xi,t) &= A \operatorname{sech}^2(B(\xi - Ct - \xi_o)), \end{aligned} \tag{4.5}$$

where

$$A_1 = \frac{C^2-1}{\alpha C^2} = \frac{1}{C^2}, \quad A_2 = \frac{(C^2-1)^2}{\alpha C^2} = \frac{\alpha}{C^2}, \quad C = \sqrt{1+\alpha}, \quad A = \frac{C^2-1}{\alpha C} = \frac{1}{C},$$

and

$$B = \left\{ \frac{C^2-1}{(4/3)\beta C^2} \right\}^{1/2} = \left\{ \frac{\alpha}{(4/3)\beta C^2} \right\}^{1/2}.$$

Note that $A_1 + A_2 = 1$. See Wei and Kirby [28] for details.

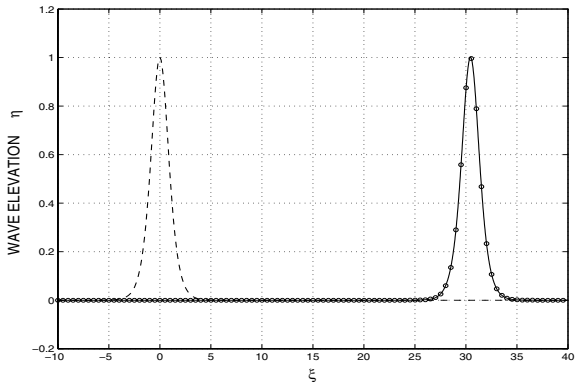


FIG. 4.3. *Solitary wave propagating over a flat bottom. Dashed line: initial disturbance* $\eta(\xi,0) = A_1\ sech^2(B\xi) + A_2\ sech^4(B\xi)$. *Solid line: numerical solution for* $\alpha = 0.03$, $\beta = 0.03$, *and* $t = 30$. *Open circles: analytical solution* (4.5).

In Figure 4.3 we show the analytical solution η given by (4.5) for $\alpha = \beta = 0.03$ and $t = 30$. The initial soliton position is set to be $\xi_o = 0$. The numerical solution is also included in this plot. The numerical parameters are $J = 2000$, $N = 2500$, $\Delta\xi = 0.025$, $\Delta t = 0.012$. Note that we have the same dispersion level as in the first simulation presented in Experiment 1. Now weak nonlinearity prevents the formation of an oscillatory tail.

Observe that the pulse's propagation velocity is $C \approx 1.0149$, in agreement with solution (4.5). The code reproduces very well the weakly dispersive, weakly nonlinear evolution of the soliton.

Experiment 3 (Disordered orography/hyperbolic regime). In Experiments 1 and 2, in which the channel's bottom was assumed to be flat, the orography-dependent coefficient $M(\xi)$ was taken to be $M \equiv 1$. For variable depths the computation of function $M(\xi)$ involves the solution of a change-of-variables problem (conformal mapping), which is not an easy task. For this reason, in next experiments, the smooth orography coefficient $M(\xi)$ will be synthesized directly as a piecewise linear function, ignoring (for the time being) its dependence on the original orography. In [20] we describe in detail how a numerical conformal mapping tool [6] is used in order to obtain a "nonsynthetic" $M(\xi)$. Nevertheless, synthetic $M(\xi)$ proves to be useful (i.e., efficient) for observing the phenomena we are interested in and for validating the theory. The synthesized orography coefficient is conceived as

$$M(\xi) = 1 + \delta\mu(\xi/\ell),$$

where μ is a mean-zero coefficient constructed using a random number generator. The fluctuation level is indicated by δ and its correlation length by ℓ. In the following experiments we use $\delta = 0.5$ and $\ell = 0.1$.

The numerical experiments are performed over a channel defined in the interval $[-15, 70]$. The fluctuations of the synthetic coefficient $M(\xi)$ cover the interval [5, 45]. The data for the right-propagating Gaussian is such that $\xi_o = -5$ and $\epsilon = 0.05$. The numerical solution is plotted as a function of the time-delay variable τ after propagating over 20 units of length and is presented in Figure 4.4. Note the wave

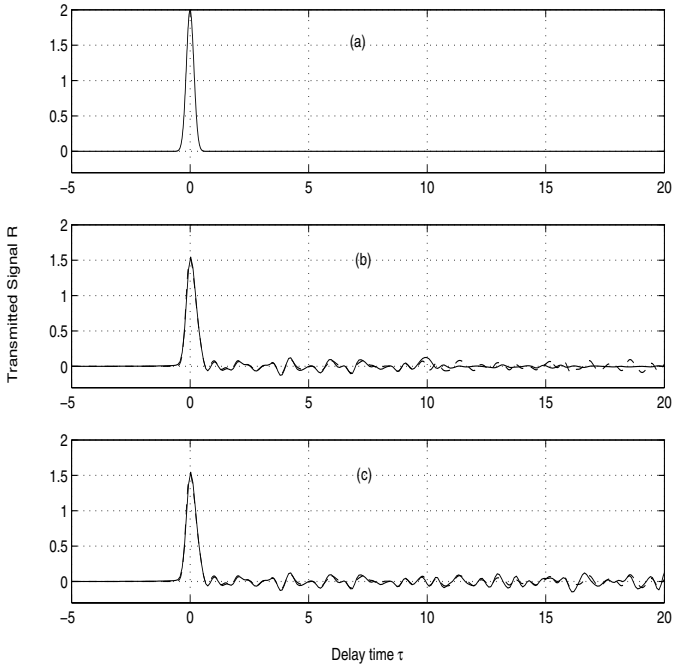


FIG. 4.4. *Pulse propagating over a synthetic disordered rapidly varying topography.* (a) *Initial disturbance* $R(0,\tau) = 2e^{-\tau^2/0.05}$. (b) *Solid line: ODA approximation* (3.31) *for* $X = 5$*; dashed line: numerical solution for* $\alpha = 0.001$, $\beta = 0$ *at* $\xi = 20$. (c) *Solid line: ODA approximation* (3.31) *for* $X = 20$*; dashed line: numerical solution at* $\xi = 20$ *and* α, β *as in* (b).

attenuation due to the orographic forcing. The transmitted signal has amplitude 1.5 (smaller than 2.0, the initial amplitude). The agreement between the numerical solution of the full nonlinear equations (with small α) and the linear theory is very good (Figure 4.4(b)). We also point out an outstanding feature of the theory, not noticed in previous work [3]. The linear hyperbolic ODA approximation is able to capture the forward scattering radiation, which is the incoherent coda behind the transmitted Gaussian. Theory and numerical experiment agree over the delay time interval up to approximately $\tau = 10$.

To verify the robustness of the theory we increase the size of the disordered medium's slab used in the invariant imbedding theory (namely, the variable X). In Figure 4.4(c) we plot the same numerical solution as above but compared with a theoretical result using an increased slab size (up to $X = 20$). The approximate theory captures an even larger segment of the forward scattering radiation beyond $\tau = 15$.

Experiment 4 (Disordered orography/dispersive regime). This set of experiments is important for two reasons. (A) It shows that we are able to properly compute the interaction of dispersive water waves with rapidly varying orographies. With previously known Boussinesq models (such as [26]) this was not possible. The classic Boussinesq equation [26] is not valid for orographies with large slopes. Moreover, its variable coefficient multiplies the highest derivative term, and this generates numerical noise as the orography's slope increases. This has been shown for a periodic topography in [20]. The same experiment was performed for the terrain-following

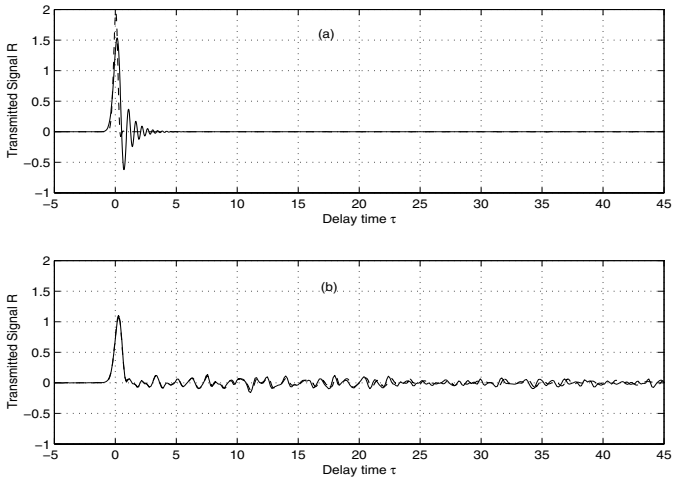


FIG. 4.5. (a) *Dashed line: initial disturbance. Solid line: solution for* $M \equiv 1$ *(flat bottom) at* $\xi = 40$. (b) *Pulse propagating over a synthetic disordered rapidly varying orography. Dashed line: numerical solution at* $\xi = 40$. *Solid line: generalized ODA approximation* (3.29) *for* $X = 20$. *In all experiments* $\alpha = 0.001$, $\beta = 0.0005$.

Boussinesq system where the metric term is positioned away from the third order (dispersive) term. No numerical noise was observed. (B) We illustrate the theoretical results in the regime for which they were deduced. Hence these linear experiments validate the nonlinear numerical model for the terrain-following Boussinesq equation in the presence of a random orography. This is important also since the code will be used (as a scientific computing tool) beyond the regime of validity of the linear theory.

In the first experiment we consider the dispersion to be very weak ($\beta = 0.0005$). Figure 4.5(a) clearly shows that a very short oscillatory tail develops when the pulse propagates in a flat channel. The final amplitude of the transmitted pulse is about 1.5. The amplitude decay in this case is entirely due to dispersion, as discussed in Appendix A, through the Airy kernel. However, in the presence of orographic forcing an additional attenuation is observed (Figure 4.5(b)). In this case the final amplitude is about 1. Note that no Airy-like oscillatory tail develops. This was systematically observed in several experiments and can be explained through the concept of *localization* [24]. The localization length of a Fourier mode is a *characteristic propagation distance* after which the transmission coefficient is negligible. The bulk of the energy is in the reflected signal. Moreover, high frequency components have small localization lengths. This means they are quickly filtered (back) by the disordered medium. In the context of the ODA theory this was phrased in a slightly different way by Berlyand and Burridge [3]. They called a layered random medium a *stratigraphic Gaussian filter*. As presented here, the transmitted pulse can be written as the convolution of its initial Fourier content with a Gaussian kernel. The Gaussian kernel is the leading-order approximation to the kernel in (3.29) with $\beta = 0$ (see [20, 21]). Applying this notion to our current problem, we have that disorder filters the higher part of the Fourier content of the incoming pulse. Hence the oscillatory tail (which is of high frequency content) has been converted into the incoherent part of the wave. Again the agreement between the numerical solution and the ODA theory is very good.

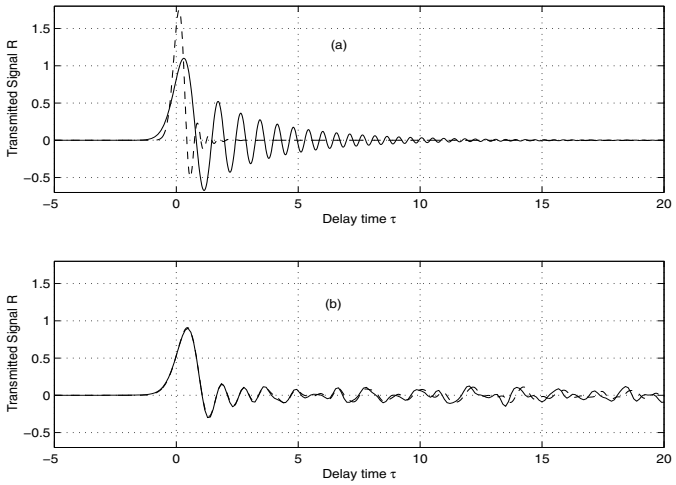


FIG. 4.6. (a) *Dashed line: initial disturbance. Solid line: solution for* $M \equiv 1$ *(flat bottom) at* $\xi = 40$. (b) *Pulse propagating over a synthetic disordered rapidly varying orography. Dashed line: numerical solution at* $\xi = 40$. *Solid line: generalized ODA approximation* (3.29) *for* $X = 20$. *In all experiments* $\alpha = 0.001$, $\beta = 0.002$.

In the next validation experiment the level of dispersion has been increased four times ($\beta = 0.002$). In Figure 4.6(a) we have the initial pulse profile and the numerical solution after propagation over 40 units of a flat channel. We observe a long oscillatory tail due to the higher dispersion level. Note that dispersion is not as small as the value of β might indicate at first sight. After large propagation distances the (small) phase lag (at higher frequencies) has accumulated in a nontrivial fashion.

To compute the theoretical ODA approximation we need the incoming pulse in time at the origin. Actually we need its Fourier content $\hat{f}(\omega)$. To be consistent with our mathematical theory we position the initial Gaussian profile (in space) to the left of the origin at time t_o, allow it to propagate over a flat portion of the channel, and record it in time at the origin. The starting time t_o is chosen so that the resulting pulse $f(\tau)$ will be centered at $\tau = 0$. This gives us the correct incoming pulse (in time) for the theoretical formula to be used. Hence the incoming pulse displays a mild oscillatory tail as displayed by Figure 4.6(a). In Figure 4.6(b) we compare the numerical solution with the generalized ODA approximation. The dispersive wave attenuation can again be observed, in particular if we examine the envelope of the Airy-like solution.

Experiment 5 (Disordered orography/solitary wave). We are now in a position to explore the (linear) generalized ODA theory beyond its regime of validity. We consider a weakly nonlinear weakly dispersive wave, namely a soliton. Using (4.5), it is easy to see that we have

$$f(\tau) = \left(\frac{1}{C^2} + \frac{1}{C}\right) \operatorname{sech}^2(B\tau) + \frac{\alpha}{C^2} \operatorname{sech}^4(B\tau)$$

in order to evaluate the ODA approximation (3.29).

In order to slowly push away from the regime of validity of our theory we choose small values for the respective parameters $\alpha = \beta = 0.001$. This amount of dispersion

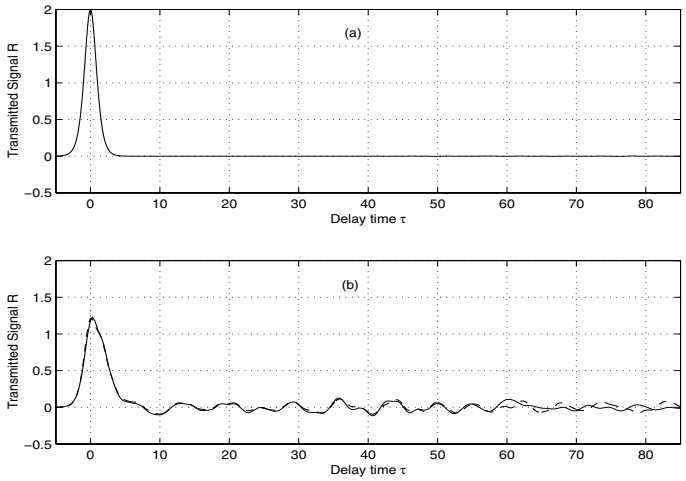


FIG. 4.7. *Soliton propagating over a synthetic disordered rapidly varying orography.* (a) *Initial soliton* $f(\tau)$. (b) *Dashed line: numerical solution for* $\alpha = 0.001$, $\beta = 0.001$ *at* $\xi = 150$. *Solid line: generalized ODA approximation* (3.29) *for* $X = 30$.

is enough to produce an oscillatory coda, as was observed in previous experiments. But now, for the particular data considered, this coda will not appear due to the perfect balance between the α and β terms. The oscillatory coda seen in Figure 4.7(b) is due entirely to forward scattering of energy generated by the interaction of the soliton with the disordered medium. Because the soliton is wider than the Gaussian (used before) we adopt $\ell = 0.6$. This keeps the wave/inhomogeneities ratio equal to approximately 10 ($\gamma = 0.1$) as in all other experiments. The orography coefficient covers the [5, 245] interval, and the amplitude of fluctuations is $\delta = 0.5$. In Figure 4.7(b) we present the excellent agreement between the theory and the numerical solution.

It is worthwhile recalling that the solitary wave (4.5) is an approximate solution to the Boussinesq equations as presented by Wei and Kirby [28] in their appendix. In [28, p. 255] they discuss solitary-wave propagation over a flat bottom and analyze the effects, under the corresponding approximation, of increasing the soliton's amplitude (namely, the nonlinearity parameter α, denoted there by δ). They observed that for $\alpha = 0.1$ the initial profile specified by (4.5) undergoes a rapid adjustment to a slightly higher solitary wave with a very small dispersive tail. This dispersive tail is not noticeable in their experiment [28, Figure 2(a)] after the soliton has propagated over 55 pulsewidths (450 length units). Nevertheless the amplitudes of the tail and of the rapid deviation from the initial solitary wave height both increase with increasing α. As explained in [28], this is partially because the fourth-order ODE used to develop the analytical solution is only asymptotically equivalent to the Boussinesq model used in the computations. In particular, for $\alpha = 0.3$ Wei and Kirby show that the corresponding evolution for (4.5) is far from a traveling wave solution.

As pointed out before, the ODA theory developed is linear, while these experiments are performed beyond the linear regime. Hence in our experiments we will gradually increase the values of α, but we will be far from the "problematic regime" indicated by Wei and Kirby [28]. In our second experiment with solitons, and to

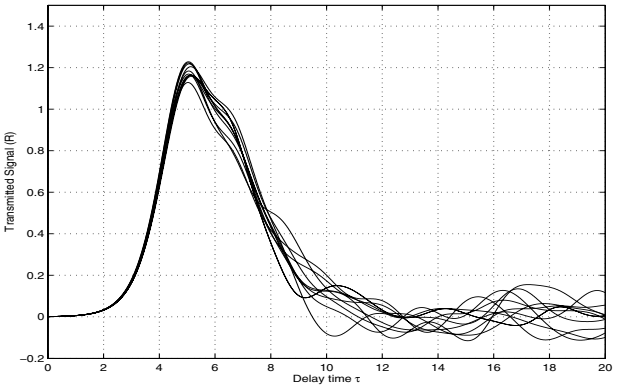


FIG. 4.8. *Transmitted pulse stabilization observed at a fixed medium's station ($\xi = 150$; 25 pulse widths). The same initial soliton ($\alpha = \beta = 0.002$) propagated over ten different realizations of the topography. The transmitted pulse shape is effectively deterministic, while the coda is random.*

further move away from the linear regime, we double the nonlinearity and dispersive parameters accordingly. We will now investigate the effect of different realizations of the medium. In Figure 4.8 we present the results for ten different realizations of the disordered topography. We observe the stabilization of the transmitted pulse: the pulse shaping of the front is independent of the specific realization. This has been proved for the linear hyperbolic case [15]. Stabilization in the linear dispersive regime has been recently proved in [10]. The present framework has been extended, through a stochastic analysis, to include stabilization [10] for the time-reversed refocusing of dispersive waves. No stabilization theory is yet available for solitons though.

As already mentioned, the solitary wave profile (4.5) is not an exact traveling wave solution to the corresponding, constant coefficient, Boussinesq system. Nevertheless the balance between weak nonlinearity and weak dispersion is maintained for large time intervals. If dispersion were not present, a Burgers-type nonlinearity ($\eta_t + \alpha\eta\eta_x = 0$) would force the solitary wave profile (4.5) to eventually break. For an initial profile denoted as $\eta(x,0) = f(x)$ the critical time t_c is known to be $t_c = -1/(\alpha f')$ for the maximum value of the negative slope of $f(x)$. For $f(x)$ given by (4.5), with $\alpha = \beta = \varepsilon$, the maximum value of the negative slope is at $\tilde{x}$ such that $\tanh(B\tilde{x}) = -z$, where

$$z = \left(\frac{1 + C + 2\alpha}{3 + 3C + 10\alpha}\right)^{1/2},$$

$$f'(\tilde{x}) = -\left(\frac{3}{(1+\varepsilon)^3}\right)^{1/2}\left[\frac{1-z^2}{z^2}\left((1+C)z + 2\varepsilon\left(\frac{1-z^2}{z}\right)\right)\right], \quad \text{with} \quad C = \sqrt{1+\varepsilon},$$

and $t_c \approx 1/(4\varepsilon)$. Hence if dispersion were switched off, the solitary wave would break after 50 length units (approximately 8.33 pulse widths), when $\varepsilon = \alpha = 0.005$, as will be used in the following experiment.

In our last experiment we further increase nonlinearity and dispersion to $\alpha = \beta = 0.005$. The result is presented in Figure 4.9. As observed in Experiment 4, disorder attenuates the effect of dispersion in this weakly nonlinear experiment. Note the soliton steepening at the wave front due to the attenuation of the dispersive mechanism. The attenuated wave front predicted by the linear theory does not match

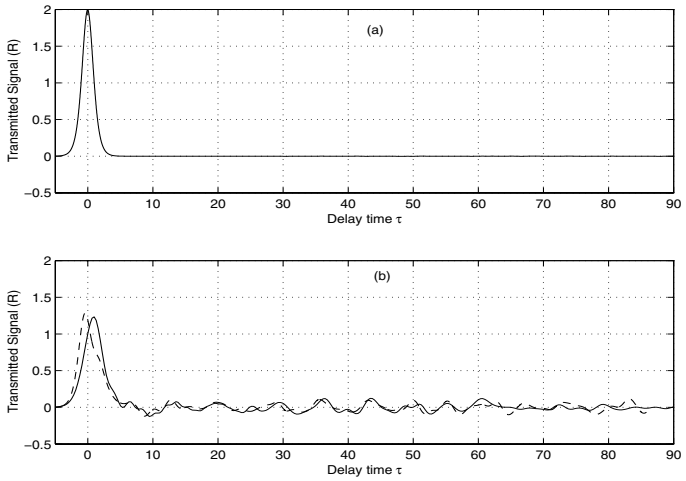


FIG. 4.9. *Soliton propagating over a synthetic disordered rapidly varying orography.* (a) *Initial soliton* $f(\tau)$. (b) *Dashed line: numerical solution for* $\alpha = 0.005$, $\beta = 0.005$ *at* $\xi = 150$. *Solid line: generalized ODA approximation* (3.29) *for* $X = 30$.

the nonlinear numerical front as before. Note also that dispersion has not been fully switched off or else a shock would have formed in finite time according to the discussion above. More experiments with solitary waves, including time-reversal and refocusing, are presented in [20].

The study of solitary waves moving over disordered topography is of great interest. In this work we have presented only scientific computing results. A complete theoretical understanding is of interest and will be considered in the future. Recently progress has been made in this direction [10, 11]. Theoretical results for nonlinear localization and soliton propagation in random media are recent and more focused on the nonlinear Schrödinger (NLS) equation. A very good source of references can be found through the work of Garnier [12, 13]. We are not aware of any results regarding solitons for the Boussinesq system (or equation).

5. Conclusion. We have formulated a generalization of the ODA theory for linear weakly dispersive waves. The theory has been validated numerically and pushed beyond its linear regime of validity. In particular, both the theoretical expressions and the numerical experiments have been able to capture the apparent diffusion of small amplitude solitary waves. This is a theme of great interest: soliton propagation in disordered media. Further mathematical analysis is needed to fully understand this problem.

Nevertheless this work has stimulated new theoretical results in the stochastic formulation for time-reversed dispersive wave refocusing [10] and also for the ODA and time-reversed refocusing [11] of weakly nonlinear hyperbolic waves. The authors [11] have shown that, to leading order, the transmitted pulse is governed by a viscous Burgers equation. The "apparent viscosity" depends on statistics of the random medium. This important result reports on a weakly nonlinear ODA theory for nondispersive waves. Therefore the regularizing effect is entirely due to the "apparent viscosity" promoted by the disordered orography. But it still does not apply to solitons. Details of the "apparently viscous" theory, including additional nonlinear experiments, are presented in [11].

Appendix A. Solutions for the linear KdV equation and the linear Boussinesq model.

The linear KdV equation. Consider the initial value problem

$$u_t + u_x + \gamma u_{xxx} = 0, \tag{A.1}$$

$$u(x,0) = f(x), \tag{A.2}$$

where γ is a nonzero constant. Using a Fourier transform in x, the solution for (A.1) and (A.2) is given by

$$u(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} e^{i(k(x-t-y)+\gamma k^3 t)}dk\right) f(y)dy. \tag{A.3}$$

Making a convenient change of variables, the inner integral in (A.3) can be expressed in terms of the Airy function to give

$$u(x,t) = \frac{1}{(3t\gamma)^{1/3}}\int_{-\infty}^{\infty} \mathrm{Ai}\left(\frac{x-t-y}{(3t\gamma)^{1/3}}\right) f(y)dy. \tag{A.4}$$

The Airy kernel gives the rate in time at which a pulse $f(x)$ will spread due to dispersion. This is useful information for the invariant imbedding technique used in the ODA theory.

The linear Boussinesq model. Now we study the linearization of system (4.1) for constant depth:

$$\begin{aligned} &\eta_t + u_\xi = 0,\\ &u_t + \eta_\xi - \frac{\beta}{3}u_{\xi\xi t} = 0, \end{aligned} \tag{A.5}$$

with the initial conditions

$$\eta(\xi,0) = u(\xi,0) = f(\xi).$$

Analogously to the KdV equation, we can apply the Fourier transform technique to obtain the Fourier coefficients

$$\begin{aligned} \hat{\eta}(k,t) = \frac{\hat{f}(k)}{2}\Bigg[&\left(1-\sqrt{1+(\beta/3)k^2}\right)e^{\frac{ikt}{\sqrt{1+\beta/3k^2}}}\\ &+\left(1+\sqrt{1+(\beta/3)k^2}\right)e^{\frac{-ikt}{\sqrt{1+\beta/3k^2}}}\Bigg], \end{aligned} \tag{A.6}$$

$$\begin{aligned} \hat{u}(k,t) = \frac{\hat{f}(k)}{2\sqrt{1+(\beta/3)k^2}}\Bigg[&\left(1+\sqrt{1+(\beta/3)k^2}\right)e^{\frac{-ikt}{\sqrt{1+\beta/3k^2}}}\\ &-\left(1-\sqrt{1+(\beta/3)k^2}\right)e^{\frac{ikt}{\sqrt{1+\beta/3k^2}}}\Bigg]. \end{aligned} \tag{A.7}$$

We point out that in the hyperbolic case ($\beta = 0$) the above initial data gives rise to (only) a right-propagating mode. In the dispersive case a negligible left-propagating mode is always present in this type of data. Notice that, because of (A.6), (A.7), and

$$\frac{1}{\sqrt{1+(\beta/3)k^2}} = 1 - (k^2/6)\beta + O(\beta^2),$$

we have that

$$\widehat{L}(t,k) = \hat{\eta}(t,k) - \hat{u}(t,k) = O(\beta) \approx 0,$$

provided that β is small enough. For the hyperbolic case the left-propagating mode is identically zero.

The dispersion relation for system (A.5) is

$$\omega_{\pm} = \omega_{\pm}(k) = \pm \frac{k}{\sqrt{1 + \frac{\beta}{3}k^2}}, \tag{A.8}$$

where ω_+ and ω_- represent Fourier modes propagating to the right and left, respectively. In contrast to the KdV equation, the phase velocity is

$$C_k^{\pm} = \frac{\omega_{\pm}}{k} = \pm \frac{1}{\sqrt{1 + \frac{\beta}{3}k^2}},$$

which does not switch signs and is bounded by one. Furthermore, $\omega_+ = \omega_+(k)$ coincides up to $O(k^3)$ with the dispersion relation for the KdV equation above with $\gamma = \beta/6$. Note also that, for waves generated in space, the range of possible time frequencies is bounded by $(3/\beta)^{1/2}$ for all k. As a consequence, solutions u, η of system (A.5) are band-limited functions in t. This fact justifies why frequencies ω higher than $\sqrt{3/\beta}$ are not considered in the analysis presented in section 3.

Appendix B. Numerical computation of coefficients $a_\beta(x_o, \omega)$ and $b(x_o, \omega)$. As mentioned in section 3, the numerical computation of the dispersive coefficient $a_\beta(x_o, \omega)$ is expensive. To override this difficulty, and use the FFT, we approximated $a_\beta(x_o, \omega)$ by the hyperbolic medium's correlation function $a_0(x_o, \omega)$, which corresponds to the leading-order term of a Taylor series expansion of a_β around $\beta = 0$. The numerical experiments in section 4 showed the high accuracy of this approximation.

To compute coefficient $a_0(x_o, \omega)$ as in (3.30) we rewrite it as

$$a_0(x_o, \omega) = \int_0^X \Phi(\eta) e^{-2i\omega\eta} d\eta, \tag{B.1}$$

where

$$\Phi(\eta) = \frac{1}{x_o} \int_0^{x_o} r(x) r(x + \eta) dx.$$

We know by the correlation theorem that

$$\int_{-\infty}^{\infty} r(x) r(x + \eta) dx = F^{-1}[\hat{r}\bar{\hat{r}}](\eta),$$

where the hat denotes the Fourier transform, F^{-1} the inverse Fourier transform, and the bar indicates complex conjugation. Therefore function Φ defined above can be computed by using the FFT algorithm, letting the coefficient $r(x)$ be zero outside the interval $[0, x_o]$. This is consistent with the invariant imbedding approach. We must append enough zeros to the tail of the sampled coefficient $r(x)$ (zero padding) in order

to eliminate the overlapping phenomenon that appears due to the fact that $r(x)$ is not a periodic function. The FFT assumes periodicity in both physical and frequency domains (see Brigham [4]). The cost of computing the discrete correlation function results in only three FFT evaluations, which is faster than an ordinary computation of the integral defining Φ for each value of η. The numerical code to perform the discrete correlation can be found in [4].

Once the function Φ is known, the windowed Fourier transform in (B.1) is evaluated by using only one FFT. Analogously to the discrete correlation, zero padding outside the interval $[0, X]$ is required on the sampling of function Φ.

To evaluate the generalized ODA approximation presented in section 3 we also need the dispersive coefficient $b(x_o, \omega)$. To make its computation faster we rewrite it in the more compact form

$$b(x_o,\omega) = -\frac{\beta}{6x_o}\int_0^{\xi(x_o)} \frac{-i\omega^3 M^{3/2}(\xi) + M'(\xi)\omega^2}{1-(\beta/3)M(\xi)\omega^2}\,d\xi, \tag{B.2}$$

where the upper limit $\xi(x_o)$ denotes the spatial position in the medium corresponding to the travel time x_o. Thus we need compute only once the coefficients $M(\xi)$, $M^{3/2}(\xi)$, and $M'(\xi)$, which can be stored at the beginning. The integral in (B.2) is approximated by the trapezoidal method for roughly 2^{14} frequencies in the range $|\omega| < C_o(x_o)\sqrt{3/\beta}$. Since $b(x_o, -\omega) = \overline{b(x_o, \omega)}$, only positive frequencies need to be evaluated.

Appendix C. Error estimation of the pulse shaping theory. In section 3 the reflection kernel K was decomposed according to the result of Lemma 3.1. In the derivation of the generalized ODA approximation (under the hypotheses of rapidly varying orography and weakly dispersive regime $0 < \beta \ll 1$) we retained only the first term on the right-hand side of the decomposition, leading to the approximation

$$\widehat{K}(x,\omega) \approx -\int_x^{x+X} e^{2i\omega\int_{x'}^{x}\Gamma(s,\omega)\,ds}\bar{\zeta}(x',\omega)dx'. \tag{C.1}$$

We wish to give numerical evidence showing the accuracy of this approximation. First, observe that the reflection kernel K can be computed by solving numerically the Riccati equation

$$\widehat{K}_x = \bar{\zeta}(x,\omega) + 2i\omega\Gamma(x,\omega)\widehat{K} - \zeta(x,\omega)\widehat{K}^2,$$

subject to

$$\widehat{K}(x_o + X, \omega; T) = 0.$$

(See the definitions of coefficients Γ and ζ in section 3.) Consequently, we can evaluate the error function $E(x_o, \omega)$ at any time x_o and frequency ω. The results are presented in Figure C.1, where we compare the norm of function $E(x_o, \omega)$ to that of approximation (C.1) at point $\xi = 40$, with $X = 20$ and $\beta = 0.005$. We see that the error function is negligible over all frequencies. The parameters used for the orography considered here are $L = 240$ (total length of the rough bottom), $\ell = 0.6$ (scale of variation of the orography), and $\delta = 0.5$ (amplitude of orography's fluctuations).

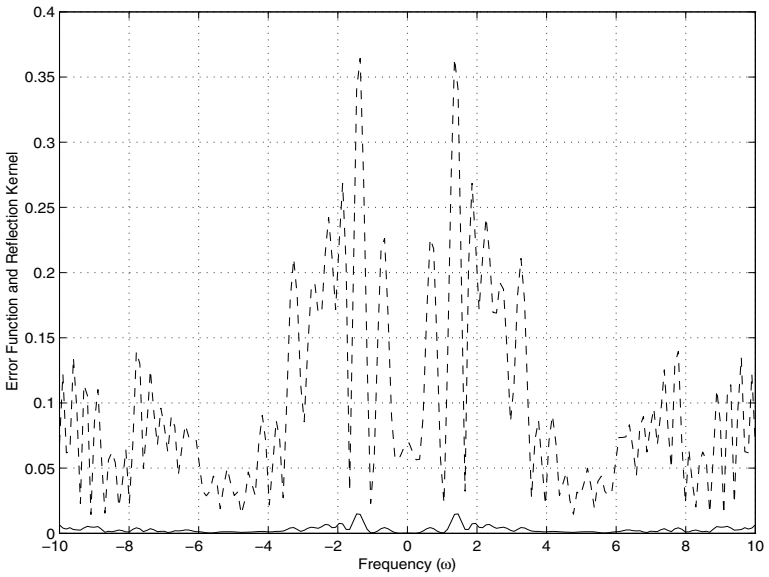


FIG. C.1. *Solid line: error norm for* $\beta = 0.005$, $X = 20$ *at* $\xi = 40$. *Dashed line: norm of approximation* (C.1) *for the same parameters.*

Acknowledgments. A. Nachbin is grateful to Prof. George Papanicolaou, as the chairman of the Mathematical Geophysics Summer School at Stanford, and for many useful discussions on this subject. He also thanks Prof. Robert Burridge for several helpful conversations regarding reference [3] during MGSS 1998.

REFERENCES

[1] M. ASCH, W. KOHLER, G. C. PAPANICOLAOU, M. POSTEL, AND B. WHITE, *Frequency content of randomly scattered signals*, SIAM Rev., 33 (1991), pp. 519–625.

[2] P. G. BAINES, *Topographic Effects in Stratified Flows*, Cambridge University Press, Cambridge, UK, 1995.

[3] L. BERLYAND AND R. BURRIDGE, *The accuracy of the O'Doherty–Anstey approximation for wave propagating in highly disordered stratified media*, Wave Motion, 21 (1995), pp. 357–373.

[4] E. O. BRIGHAM, *The Fast Fourier Transform and Its Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[5] R. G. DEAN AND R. A. DALRYMPLE, *Water Wave Mechanics for Engineers and Scientists*, 3rd ed., World Scientific, River Edge, NJ, 1993.

[6] T. DRISCOLL, http://www.math.udel.edu/~driscoll/software.

[7] EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS (ECMWF), *Orography*, Proceedings of a Workshop held at ECMWF, Shinfield Park, Reading, UK, 1998.

[8] B. ENGQUIST AND A. MAJDA, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp., 31 (1977), pp. 629–651.

[9] J. F. CLOUET AND J. P. FOUQUE, *Spreading of a pulse travelling in random media*, Ann. Appl. Probab., 4 (1994), pp. 1083–1097.

[10] J. P. FOUQUE, J. GARNIER, AND A. NACHBIN, *Time reversal for dispersive waves in random media*, SIAM J. Appl. Math., to appear.

[11] J. P. FOUQUE, J. GARNIER, AND A. NACHBIN, *Shock structure due to stochastic forcing and time reversal for nonlinear water waves in random media*, Phys. D, to appear.

[12] J. GARNIER, *Long-time dynamics of Korteweg–de Vries solitons driven by random perturbations*, J. Statist. Phys., 105 (2001), pp. 789–833.

[13] J. Garnier, *Exponential localization versus soliton propagation*, in Nonlinearity and Disorder: Theory and Applications (Proceedings of the NATO Advanced Research Workshop, Tashkent, Uzbekistan, 2001), F. Abdullaev, O. Bang, and M. P. Sørensen, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, Vol. 45, 2002, pp. 3–17.
[14] J. Hamilton, *Differential equations for long-period gravity waves on a fluid of rapidly varying depth*, J. Fluid Mech., 83 (1977), pp. 289–310.
[15] P. Lewicki, R. Burridge, and M. V. de Hoop, *Beyond effective medium theory: Pulse stabilization for multimode wave propagation in high-contrast layered media,* SIAM J. Appl. Math., 56 (1996), pp. 256–276.
[16] F. Mattioli, *Decomposition of the Boussinesq equations for shallow-water into a set of coupled Korteweg–de Vries equations*, Phys. Fluids A, 3 (1991), pp. 2355–2359.
[17] F. Mattioli, *On the Hamiltonian decomposition of the Boussinesq equations in a pair of coupled Korteweg–de Vries equations*, Wave Motion, 28 (1998), pp. 283–296.
[18] C. C. Mei, *The Applied Dynamics of Ocean Surface Waves*, John Wiley, New York, 1983.
[19] G. Meyer, *Initial Value Methods for Boundary Value Problems: Theory and Application of Invariant Imbedding*, Academic Press, New York, 1973.
[20] J. C. Muñoz Grajales and A. Nachbin, *Stiff microscale forcing and solitary wave refocusing*, Multiscale Model. Simul., submitted.
[21] J. C. Muñoz Grajales, *Dispersive Wave Attenuation and Refocusing Due to Disordered Orographic Forcing*, Ph.D. thesis (in English), Instituto de Matematica Pura e Aplicada, Rio de Janeiro, Brazil, 2002.
[22] A. Nachbin, *Modelling of Water Waves in Shallow Channels*, Computational Mechanics Publications, Southampton, UK, 1993.
[23] A. Nachbin, *A terrain-following Boussinesq system*, SIAM Appl. Math., 63 (2003), pp. 905–922.
[24] A. Nachbin, *The localization length of randomly scattered water waves,* J. Fluid Mech., 296 (1995), pp. 353–372.
[25] R. F. O'Doherty and N. A. Anstey, *Reflections on amplitudes*, Geophys. Prospecting, 19 (1971), pp. 430–458.
[26] D. H. Peregrine, *Long waves on a beach*, J. Fluid Mech., 27 (1967), pp. 815–827.
[27] G. Papanicolaou and K. Sølna, *Ray theory for a locally layered random medium*, Waves in Random Media, 10 (2000), pp. 151–198.
[28] G. Wei and J. Kirby, *Time-dependent numerical code for extended Boussinesq equations*, J. Waterway, Port, Coastal, and Ocean Engineering, 121 (1995), pp. 251–261.
[29] G. B. Whitham *Linear and Nonlinear Waves*, John Wiley, New York, 1974.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 1002–1034

HOMOGENIZED NON-NEWTONIAN VISCOELASTIC RHEOLOGY OF A SUSPENSION OF INTERACTING PARTICLES IN A VISCOUS NEWTONIAN FLUID*

L. BERLYAND† AND E. KHRUSLOV‡

Abstract. Our study is motivated by an attempt to develop a rigorous mathematical model of a suspension highly filled with a large number of small solid particles, which *interact* due to surface forces. We use asymptotic analysis in the small parameter ϵ and consider irregular (nonperiodic) geometries for which the sizes of particles and the distances between them are of order ϵ. We present conditions under which the homogenization of a Newtonian fluid with interacting particles leads to a single medium which is an anisotropic, non-Newtonian viscoelastic fluid with memory described by a relaxation term. We derive formulas for the calculation of the effective viscosity tensor and the relaxation integral kernel. For periodic arrays of particles we show how this tensor can be explicitly computed and compute the distribution of the relaxation times, which is the main quantity of interest in the rheological studies. We also show how the particles' shapes affect this distribution.

Key words. homogenization, non-Newtonian fluids, suspensions, relaxation time, viscoelasticity, interaction, surface forces

AMS subject classifications. 35B27, 35Q30, 73D25, 76T20

DOI. 10.1137/S0036139902403913

1. Introduction.

In this work we propose a rigorous mathematical model of polymer compounds highly filled with a large number of small particles, which *interact* due to the surface forces (e.g., Van der Waals or London forces). Such polymer compounds (e.g., carbon black or silica particles in a polymer matrix) are widely used in industry, and their properties have been the subject of studies in polymer science and engineering literature (see, for example, comprehensive review articles [4], [19]).

Similar questions arise in the study of colloidal suspensions of a large number of small *interacting* particles in a fluid. Such suspensions are ubiquitous, and control of their rheology is vital to the commercial success of many products (see, e.g., [5], [17] and references therein). Here we are interested in the suspensions of solid particles in fluid when particles do not clump together (do not gel).

The key common feature in both the polymer compounds and the suspensions is the *interaction* between particles due to surface forces. Our main focus is to develop a rigorous mathematical model which shows how the interaction affects the homogenized medium. We also provide an approximation for the effective (overall) viscoelastic properties of the mixture and compute, in particular, the relaxation times for several specific arrays of particles.

We start from a mathematical model of the suspension. It consists of the time dependent Stokes equations for an incompressible viscous fluid, the boundary conditions on the fluid/particle interfaces, and the balance of linear and angular momentums for

*Received by the editors March 11, 2002; accepted for publication (in revised form) October 1, 2003; published electronically April 14, 2004.

http://www.siam.org/journals/siap/64-3/40391.html

†Department of Mathematics and Materials Research Institute, Penn State University, University Park, PA 16802 (berlyand@math.psu.edu). The work of this author was supported by NSF grants DMS-9971999 and DMS-0204637.

‡Institute of Low Temperature and Engineering, Ukrainian Academy of Science, Lenin Ave 47, Kharkov 310164 (KHRUSLOV@ilt.kharkov.ua). The work of this author was partially supported by NSF grant DMS-9971999.

each particle. The latter two balance equations incorporate the interaction forces. The key ingredient is the introduction of an appropriate potential energy functional which describes particle-particle interaction. Roughly speaking, this functional takes into account the fact that for sufficiently close pairs of particles there is an attractive/repulsive force between each point on the surfaces of both particles.

This system of equations is completed by appropriate initial conditions and the boundary conditions on the external boundary, which correspond to the relaxation measurements.

Note that, while for colloidal suspensions the Stokes equations provide a natural description of the fluid phase, for polymer compounds it is natural to consider viscoelastic (non-Newtonian) equations for the fluid phase. However, under special circumstances, when the relaxation times in the fluid phase are very small (short fading memory) in comparison with other characteristic times in the system (e.g., for external forces), one can still use Stokes equations to characterize the fluid phase.

We use asymptotic analysis in the small parameter ϵ and consider the geometries for which the sizes of particles and the distances between them are of order ϵ. We show that if the strength of the interaction forces is of order of ϵ, then homogenization of a Newtonian fluid with interacting particles leads to a single medium, which is an anisotropic, non-Newtonian viscoelastic fluid with memory described by a relaxation term. We derive formulas for calculation of the effective viscosity tensor and the relaxation integral kernel. For periodic arrays of particles we show how this tensor can be explicitly computed and compute the distribution of the relaxation times, which is the main quantity of interest in the rheological studies. We also show how particles' shapes affect this distribution.

It follows from our analysis that for a weaker interaction the anisotropic effect vanishes in the homogenization limit. The case of stronger interaction is somewhat trivial; that is, it can be shown that the homogenized limit is zero, since the effective medium becomes stiff. Therefore our main focus is on the order ϵ interaction case when homogenization results in a drastic change in the constitutive equations. We restrict our attention to the case of a Newtonian fluid phase because our main focus is on showing how the interaction between the particles gives rise to viscoelastic relaxation in the effective medium. However, our approach is applicable for the non-Newtonian viscoelastic fluid filled by solid particles.

We remark here that the fact that homogenization of elastic and fluid phases can lead to effective viscoelastic behavior (with or without memory) was observed by many authors in various problems. For example, in [3] it was rigorously proved that incompressible viscous fluid in an elastic porous medium has an overall viscoelastic behavior. This work justified the formal asymptotic result obtained earlier in [2]. Also for dynamic problems of oscillations of a mixture of an elastic solid phase and viscous fluid phase the effective viscoelastic equations were rigorously derived in a series of works (see [16], [15] and references therein). A comprehensive mathematical study of elastic/viscoelastic composites that leads to a several types of effective viscoelastic behavior, including a threshold phenomenon, is presented in [14]. In all these works the viscoelastic overall behavior is obtained due to the presence of an elastic phase as well as a viscous fluid phase or viscoelastic phase.

In our work the solid phase consists of absolutely rigid particles, and the overall viscoelastic behavior is due to the *interaction* between particles and viscosity of the fluid phase. While this phenomenon was studied both theoretically and experimentally in the physics literature (see [4], [19], [17] and references therein), it was not obtained as the result of a mathematical homogenization procedure. In this paper we

present a model of interacting solid particles in a viscous fluid and derive the effective viscoelastic model as a homogenization limit. Our approach rigorously justifies this asymptotic limit for very generic nonperiodic geometries and provides new computational formulas for the viscoelastic kernel and the relaxation times. Furthermore, we analyzed this formula analytically and arrived at an interesting qualitative conclusion that the symmetry of particles affects the number of relaxation times. The issue of relaxation times is of importance for the rheological community [5] and was brought to our attention in the context of this work by R. Lakes. Thus our work provides (a) mathematical justification for qualitative observations previously made in physics literature and (b) formulas for effective properties which can be used as a tool for their numerical evaluation.

2. Formulation of the problem. Let Ω be a bounded domain in $\mathbb{R}^3$ with piecewise smooth boundary $\partial\Omega$. This domain is occupied by a composite medium, which is a suspension of a large number of small rigid particles Q_ϵ^i $(i = 1, \dots N_\epsilon)$ in a viscous incompressible fluid. The boundary $S_\epsilon^i = \partial Q_\epsilon^i$ of each particle is smooth. The small parameter $\epsilon > 0$ characterizes the array of particles so that $d_\epsilon^i = O(\epsilon)$ and $N^\epsilon = O\left(\frac{1}{\epsilon^3}\right)$, where d_ϵ^i are diameters of the particles Q_ϵ^i. Thus the average distances between neighboring particles are of order of ϵ. The location of a particle Q_ϵ^i is characterized by the position of its center of mass $\underline{x}^i$ and the vector of its three Euler angles $\underline{\alpha}^i = (\alpha_1^i, \alpha_2^i, \alpha_3^i)$.

When the fluid is at rest the system of particles Q_ϵ^i is in equilibrium, which is described by the system of vectors $\underline{x}_\epsilon^i$ and $\underline{\alpha}_\epsilon^i$ $(i = 1, \dots, N_\epsilon)$. The equilibrium is determined by the minimum of the potential energy $H_\epsilon(\underline{x}^i, \underline{\alpha}^i)$, which describes the pairwise interaction between the particles and between the particles and the boundary $\partial\Omega$:

$$H_\epsilon(\underline{x}_\epsilon^i, \underline{\alpha}_\epsilon^i) = \min H_\varepsilon(\underline{x}^i, \underline{\alpha}^i).$$

The interaction between particles is determined by the surface forces such as Van der Waals forces and London forces; see [5]. The total interaction energy H_ε is the sum of the interactions between pairs of points $\underline{x} \in S_\varepsilon^i$ and $\underline{y} \in S_\varepsilon^j$. Since the particles are rigid, the displacement (translation and rotation) $\delta\underline{x}$ of each point $\underline{x} \in S_\varepsilon^i$ is determined by the following equality:

$$\delta\underline{x} = \underline{u}^i + \underline{\theta}^i \times (\underline{x} - \underline{x}_\varepsilon^i)\,,$$

where $\underline{u}^i = \underline{x}^i - \underline{x}_\varepsilon^i$ is the displacement of the center of mass of the particle Q_ε^i and $\underline{\theta}^i$ is the vector, which determines the rotation of the particle about some axis oriented along the vector $\underline{\theta}^i$. For small displacements the vector $\underline{\theta}^i$ is related to the increments of the Euler's angles $\delta\alpha^i = \alpha^i - \alpha_\varepsilon^i$ by the standard kinematic relations $\underline{\theta}^i = \underline{\underline{R}}\ \underline{\delta\alpha}^i$, where matrix $\underline{\underline{R}}$ depends on $\underline{\alpha}_\varepsilon^i$; see [6].

Thus near the equilibrium we write the potential energy due to the interaction between the particles in the following form (a quadratic approximation):

$$H_{v\varepsilon}(\underline{x}^i, \underline{\alpha}^i) = H_{v\varepsilon}(\underline{x}_\varepsilon^i, \underline{\alpha}_\varepsilon^i) + \frac{1}{2}\sum_{\substack{i,j \\ j\neq i}} \int_{S_\varepsilon^i}\int_{S_\varepsilon^j} \langle \underline{\underline{C}}_\varepsilon^{ij}(\underline{x}, \underline{y}) \left[\underline{u}^i + \underline{\theta}^i \times (\underline{x} - \underline{x}_\varepsilon^i)\right.$$

$$\left. - \underline{u}^j - \underline{\theta}^j \times (\underline{y} - \underline{x}_\varepsilon^j)\right], \left[\underline{u}^i + \underline{\theta}^i \times (\underline{x} - \underline{x}_\varepsilon^i) - \underline{u}^j - \underline{\theta}^j \times (\underline{y} - \underline{x}_\varepsilon^j)\right]\rangle dS_x\, dS_y + \text{h.o.t.} \tag{2.1}$$

Here $\underline{u}^i = \underline{x}^i - \underline{x}_\varepsilon^i$, $\underline{\theta}^i = \underline{\underline{R}}[\underline{\alpha}^i - \underline{\alpha}_\varepsilon^i]$, $\langle\ ,\ \rangle$ stands for the dot product in $\mathbb{R}^3$, $\underline{\underline{C}}_\varepsilon^{ij}(\underline{x}, \underline{y})$ are symmetric nonnegative 3×3 matrices, and h.o.t. stands for higher order terms.

It is natural to assume that if two points $\underline{x}$ and $\underline{y}$ move as end points of a rigid rod, that is, their displacements are given by

$$\underline{u}(\underline{x}) = \underline{a} + \underline{b} \times \underline{x} \qquad \text{and} \qquad \underline{u}(\underline{y}) = \underline{a} + \underline{b} \times \underline{y}, \tag{2.2}$$

where $\underline{a}$ and $\underline{b}$ are constant vectors, then their interaction energy is zero, that is, the matrix $\underline{\underline{C}}^{ij}_{\varepsilon}(\underline{x}, \underline{y})$ satisfies the following condition:

$$\langle \underline{\underline{C}}^{ij}_{\varepsilon}(\underline{x}, \underline{y})[\underline{b} \times (\underline{x} - \underline{y})], [\underline{b} \times (\underline{x} - \underline{y})] \rangle = 0. \tag{2.3}$$

In this work we assume that up to a scalar factor $a^{ij}_{\varepsilon}(\underline{x}, \underline{y}) > 0$ the matrix $\underline{\underline{C}}^{ij}_{\varepsilon}(\underline{x}, \underline{y})$ corresponds to the operator of projection on the unit vector $\underline{\ell}(x, y) = (\underline{x} - \underline{y})/|\underline{x} - \underline{y}|$; that is, for any unit vector $\underline{v}$

$$\underline{\underline{C}}^{ij}_{\varepsilon}\, \underline{v} = a^{ij}_{\varepsilon}(\underline{x}, \underline{y})\ P_{xy}\underline{v} := a^{ij}_{\varepsilon}(\underline{x}, \underline{y}) \langle \underline{\ell}(\underline{x}, \underline{y}), \underline{v} \rangle \underline{\ell}(\underline{x}, \underline{y}), \tag{2.3$'$}$$

where P_{xy} is the projection operator and $a^{ij}_{\varepsilon}(\underline{x}, \underline{y})$ is a nonnegative bounded function $a^{ij}_{\varepsilon}(\underline{x}, \underline{y}) \in L^{\infty}(S^i_{\varepsilon} \times S^j_{\varepsilon})$. Clearly (2.3′) implies (2.3). We consider the short-range interactions (see condition (a_3) below) and assume that the matrices $\underline{\underline{C}}^{ij}_{\varepsilon}$ satisfy the following conditions:

$$\underline{\underline{C}}^{ij}_{\varepsilon}(\underline{x}, \underline{y}) = \underline{\underline{C}}^{ji}_{\varepsilon}(\underline{y}, \underline{x}). \tag{2.4}$$

Observe that the interaction energy (2.1) is invariant under translations and rotations of the system of particles as a whole. In other words, the system of particles connected by "virtual springs" which represent the interaction has a continuum of nonlocalized equilibriums if not clamped to some external boundary.

In order to obtain a localized equilibrium $\{\underline{x}^i_{\varepsilon}, \underline{\theta}^i_{\varepsilon}, i = 1, \dots N_{\varepsilon}\}$ we take into account interaction between some particles and the external boundary $\partial\Omega$ (e.g., particles in a boundary layer of thickness ε and "fully visible" from the boundary). Figure 2.1 illustrates the interaction between a particle Q^i_{ε} and a piece of the boundary $\partial\Omega^i_{\varepsilon}$ (a quasiparticle), which "sees" this particle.

Then, in addition to the (volume) energy (2.1), we need to take into account the energy due to the interaction with $\partial\Omega$, for which in the quadratic approximation can be written in the following form:

$$\begin{aligned} H_{b\varepsilon}(\underline{u}^i, \underline{\theta}^i, \underline{u}) = \sum_{\substack{i,j \\ i \neq j}} \int_{S^i_{\varepsilon}} \int_{\partial\Omega^j_{\varepsilon}} < \underline{\underline{C}}^{ij}(\underline{x}, \underline{y})[\underline{u}^i_{\varepsilon} + \underline{\theta}^i_{\varepsilon} \times (\underline{x} - \underline{x}^i_{\varepsilon}) - \underline{u}(\underline{y})], \\ [\underline{u}^i_{\varepsilon} + \underline{\theta}^i_{\varepsilon} \times (\underline{x} - \underline{x}^i_{\varepsilon}) - \underline{u}(y)] > dS_x dS_y, \end{aligned} \tag{2.5}$$

where $\partial\Omega^i_{\varepsilon}$ are "pieces" of the external boundary $\partial\Omega$ (size of $\partial\Omega^i_{\varepsilon} \sim \varepsilon$), $\underline{u}(y, t) = \int_0^t \underline{v}(y, t)dt$ is the displacement vector, $\underline{v}(y, t)$ is the velocity vector on $\partial\Omega$, and summation in i is taken over particles close to the boundary $\partial\Omega$ (see Figure 2.1).

Thus the total interaction energy $H_{\varepsilon} = H_{\varepsilon}(\underline{u}^i, \underline{\theta}^i, \underline{u})$ is the sum of the volume and the boundary parts:

$$H_{\varepsilon}(u^i, \theta^i, u) = H_{v\varepsilon}(\underline{u}^i, \underline{\theta}^i) + H_{b\varepsilon}(\underline{u}^i, \underline{\theta}^i, \underline{u}). \tag{2.6}$$

Note that the pieces $\partial\Omega^i_{\varepsilon}$ can be viewed as "flat, nonrigid particles" (quasiparticles) which are glued to the boundary $\partial\Omega$. The total number of quasiparticles and

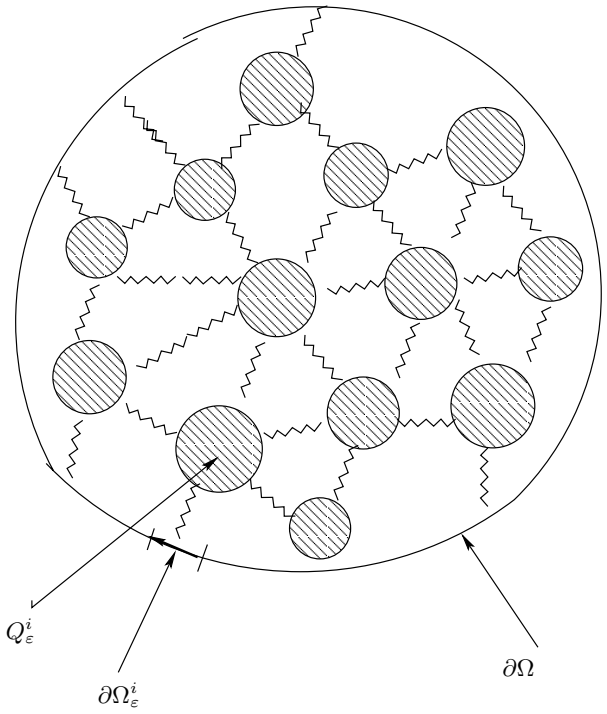


FIG. 2.1. *Domain Ω and its boundary.*

the particles which interact with them is of order $O(\varepsilon^{-2}) \ll N_\varepsilon \sim O(\varepsilon^{-3})$. Unless otherwise is specified, we will use the same notation Q_ε^j and $S_\varepsilon^j = \partial Q_\varepsilon^j$ for both particles and quasiparticles. Note that the energy (2.1) in the variables $\underline{u}^i, \underline{\theta}^i$ can also be rewritten in the following form, which is convenient for further considerations:

$$
\begin{aligned}
H_{v\varepsilon}(\underline{u}^i, \underline{\theta}^i) &= H_{v\varepsilon}(0,0) + \frac{1}{2}\sum_{\substack{i,j\\ i\neq j}} \langle \underline{\underline{C}}_{1\varepsilon}^{ij}[\underline{u}^i - \underline{u}^j], [\underline{u}^i - \underline{u}^j]\rangle \\
&+ \sum_{\substack{i,j\\ i\neq j}} \langle \underline{\underline{C}}_{2\varepsilon}^{ij}[\underline{u}^i - \underline{u}^j], \underline{\theta}^i\rangle + \sum_{\substack{i,j\\ i\neq j}} \langle \underline{\underline{C}}_{3\varepsilon}^{ij}\underline{\theta}^i, \underline{\theta}^i\rangle + \sum_{\substack{i,j\\ i\neq j}} \langle \underline{\underline{C}}_{4\varepsilon}^{ij}\underline{\theta}^i, \underline{\theta}^j\rangle,
\end{aligned} \tag{2.7}
$$

where

$$
\begin{cases}
\underline{\underline{C}}_{1\varepsilon}^{ij} = \displaystyle\int_{S_\varepsilon^i}\int_{S_\varepsilon^j} \underline{\underline{C}}_{\varepsilon}^{ij}(\underline{x},\underline{y})\,dS_x dS_y, \\
\underline{\underline{C}}_{2\varepsilon}^{ij} = \displaystyle\int_{S_\varepsilon^i}\int_{S_\varepsilon^j} \underline{\underline{A}}^T(\underline{x} - \underline{x}_\varepsilon^i)\underline{\underline{C}}_{\varepsilon}^{ij}(\underline{x},\underline{y})\,dS_x dS_y, \\
\underline{\underline{C}}_{3\varepsilon}^{ij} = \displaystyle\int_{S_\varepsilon^i}\int_{S_\varepsilon^j} \underline{\underline{A}}^T(\underline{x} - \underline{x}_\varepsilon^i)\underline{\underline{C}}^{ij}(\underline{x},\underline{y})\underline{\underline{A}}(\underline{x} - \underline{x}_\varepsilon^i)\,dS_x dS_y, \\
\underline{\underline{C}}_{4\varepsilon}^{ij} = -\displaystyle\int_{S_\varepsilon^i}\int_{S_\varepsilon^j} \underline{\underline{A}}^T(\underline{y} - \underline{x}^j)\underline{\underline{C}}^{ij}(\underline{x},\underline{y})\underline{\underline{A}}(\underline{x} - \underline{x}^i)\,dS_x dS_y.
\end{cases} \tag{2.8}
$$

Here we introduced a skew symmetric matrix $\underline{\underline{A}}(x)$ defined by

$$\underline{\theta} \times \underline{x} = \underline{\underline{A}}(\underline{x})\underline{\theta} \tag{2.9}$$

and pulled the constant translation and rotation vectors outside the integral.

We introduce the following notation: $\Omega_\varepsilon = \Omega \backslash \cup_{i=1}^{N_\varepsilon} Q_\varepsilon^i$ is the domain occupied by the fluid; ρ_s, ρ_f are the specific mass density of the solid particles and the fluid, respectively; μ is the dynamic viscosity of the fluid; $m_\varepsilon^i = \rho_s |Q_\varepsilon^i|$ is the mass of a particle Q_ε^i; and $\underline{\underline{I}}_\varepsilon$ is the tensor of inertia of a particle Q_ε^i.

Then we write a linearized system of equations that describes the dynamics of the suspension (viscous fluid filled by the interacting particles):

$$\rho_f \frac{\partial \underline{v}_\varepsilon}{\partial t} - \mu \Delta \underline{v}_\varepsilon = \nabla p_\varepsilon \quad \text{in } \Omega_\varepsilon, \tag{2.10}$$

$$\operatorname{div} \underline{v}_\varepsilon = 0 \quad \text{in } \Omega_\varepsilon, \tag{2.11}$$

$$\underline{v}_\varepsilon = \underline{\dot{u}}_\varepsilon^i + \underline{\dot{\theta}}_\varepsilon^i \times (\underline{x} - \underline{x}_\varepsilon^i), \quad \underline{x} \in Q_\varepsilon^i, \quad i = 1, \dots N_\varepsilon, \tag{2.12}$$

$$m_\varepsilon^i \underline{\ddot{u}}_\varepsilon^i = -\int_{S_\varepsilon^i} \underline{\underline{\sigma}}[\underline{v}_\varepsilon]\underline{\nu}\, dS - \nabla_{u^i} H_\varepsilon, \tag{2.13}$$

$$\underline{\underline{I}}_\varepsilon^i \underline{\ddot{\theta}}_\varepsilon^i = -\int_{S_\varepsilon^i} (\underline{x} - \underline{x}_\varepsilon^i) \times \underline{\underline{\sigma}}[\underline{v}_\varepsilon]\underline{\nu}\, dS_x - \nabla_{\theta^i} H_\varepsilon. \tag{2.14}$$

Conditions (2.13)–(2.14) hold for all particles Q_ε^i located inside the domain Ω but not for the quasiparticles $\partial\Omega_\varepsilon^i$ (see Figure 2.1). Here $\underline{v}_\varepsilon = \underline{v}_\varepsilon(\underline{x}, t)$ is the velocity of the fluid, $p_\varepsilon = p_\varepsilon(\underline{x}, t)$ is the pressure, $\underline{u}_\varepsilon^i$ is the displacement of the center of mass of a particle Q_ε^i, and $\underline{\theta}_\varepsilon^i$ is the rotation vector of Q_ε^i. We also use the following notation: $\dot{u}_\varepsilon^i = du_\varepsilon^i/dt$ and $\ddot{u}_\varepsilon^i = d^2u_\varepsilon^i/dt^2$ for the velocity and the acceleration, respectively, of the center of mass of Q_ε^i; $\dot{\theta}_\varepsilon^i$ for the instant angular velocity of Q_ε^i; and $\underline{\nu}$ for the unit inner normal vector to the surface $S_\varepsilon^i = \partial Q_\varepsilon^i$. The stress tensor in the fluid $\underline{\underline{\sigma}}[v_\varepsilon]$ is a symmetric second rank tensor defined as follows (see [7]):

$$\sigma_{ik} = \mu \left[\frac{\partial v_{\varepsilon i}}{\partial x_k} + \frac{\partial v_{\varepsilon k}}{\partial x_i} \right] - p_\varepsilon \delta_{ik} \quad (i, k = 1, 2, 3). \tag{2.15}$$

The equations (2.12) represent the nonslip condition at the fluid-particle interfaces $S_\varepsilon^i = \partial Q_\varepsilon^i$, $i = 1, \dots N_\varepsilon$. Using this condition, one can naturally extend the velocity field $\underline{v}_\varepsilon$ into the particles Q_ε^i. Equations (2.13) and (2.14) represent the balance of linear and angular momentums. In (2.14) we assume that the tensor of inertia $\underline{\underline{I}}_\varepsilon^i$ is constant, since we consider linearization for small displacements. The first (integral) terms in the right-hand side (RHS) of (2.13)–(2.14) are due to the forces exerted on the particles by the fluid, and the second terms are due to the interaction between the particles (surface or Van der Waals forces).

The system (2.10)–(2.14) is supplemented by the initial conditions

$$\begin{cases} \underline{v}_\varepsilon(\underline{x}, 0) = \underline{v}_{\varepsilon 0}(\underline{x}), \quad \underline{x} \in \Omega_\varepsilon, \\ \underline{u}_\varepsilon^i(0) = 0,\ \underline{\dot{u}}_\varepsilon^i(0) = \underline{u}_{\varepsilon 1}^i,\ \underline{\theta}_\varepsilon^i(0) = 0,\ \underline{\dot{\theta}}_\varepsilon^i(0) = \underline{\theta}_{\varepsilon 1},\ \underline{x} \in Q_\varepsilon^i \quad \text{(on the particles)}, \end{cases} \tag{2.16}$$

and the boundary conditions on the external boundary

$$\underline{v}_\varepsilon(\underline{x},t) = \underline{f}(\underline{x},t), \quad \underline{x} \in \partial\Omega, \quad t \geq 0. \tag{2.17}$$

We consider the boundary function $\underline{f}(\underline{x},t)$, which is sufficiently smooth (e.g., $\underline{f}(\underline{x},t) \in C^2(\partial\Omega)$) and decays sufficiently fast as $t \to \infty$. Since the fluid is incompressible, we have

$$\int_{\partial\Omega} \langle \underline{f}(\underline{x},t), \underline{\nu}(x)\rangle \, dS_x = 0, \tag{2.18}$$

where $\underline{\nu}(x)$ is the unit normal to $\partial\Omega$ at a point $x \in \partial\Omega$.

The boundary condition (2.17) corresponds to experimental measurements in which the macroscopic rheological properties of the compound are determined when the rate of change of the displacement at the external boundary is prescribed. Then one measures the normal stresses (the response) on the boundary (relaxation measurements).

Another type of experiment, when the load forces (stresses) are prescribed at the external boundary and the displacements are measured (creep measurements [7]), corresponds to the boundary conditions

$$\underline{\underline{\sigma}}[\underline{v}_\varepsilon] \cdot \underline{\nu}(\underline{x}) = \underline{f}_\varepsilon(\underline{x}), \quad \underline{x} \in \partial\Omega, \tag{2.19}$$

for the system (2.10)–(2.16).

The main goal of our work is to obtain the homogenized problem for the initial boundary value problem (2.10)–(2.18) in the limit as $\varepsilon \to 0$ and to establish the convergence of its solutions $\underline{u}_\varepsilon$ to the solution $\underline{u}$ of the homogenized problem, which is a single macroscopic medium with new effective (rheological) properties. We also show how to compute these properties for particular geometries if the interaction matrix $\underline{\underline{C}}^{ij}(\underline{x},\underline{y})$ is known.

Both initial boundary value problems (2.10)–(2.18) and (2.10)–(2.16), (2.19) can be studied using the homogenization approach developed in this work for the boundary condition (2.17).

3. Mesocharacteristic and formulation of the main result. In order to formulate the main result we first formulate two conditions, which describe the geometry of the particles.

(a_1) Each particle Q^i_ε is obtained by ε-rescaling of a body from a collection $\mathcal{M}$, where $\mathcal{M} = \{Q^{(m)}, m = 1, 2, \dots M\}$ is a finite collection of convex bodies in $\mathbb{R}^3$, with smooth boundaries $S^{(m)} = \partial Q^{(m)}$; that is, $Q^i_\varepsilon = \underline{\underline{R}}^i \underline{\underline{T}}^i_\varepsilon Q^{(m_i)}$, $Q^{(m_i)} \in \mathcal{M}$, and R^i and T^i_ε are the rotation and the translation operators, respectively. Thus, the diameters of the particles $d^i_\varepsilon = \varepsilon$ diam $(Q^{(m_i)})$, $1 \leq m_i \leq M$, and the area $|S^i_\varepsilon|$ of a particle Q^i_ε can be bounded as follows: $|S^i_\varepsilon| < C\varepsilon^2$, $C > 0$, is independent on ε.

(a_2) Let $B(Q^i_\varepsilon)$ be a ball of minimal radius such that $Q^i_\varepsilon \subseteq B(Q^i_\varepsilon)$, and let r^i_ε be the distance from $B(Q^i_\varepsilon)$ to other minimal balls and the boundary $\partial\Omega$, $r^i_\varepsilon =$ dist $\{B(Q^i_\varepsilon), \bigcup_{j\neq i} B(Q^j_\varepsilon) \bigcup \partial\Omega\}$.

We assume that the following inequalities hold,

$$C_1\varepsilon \leq r^i_\varepsilon \leq C_2\varepsilon\,, \tag{3.1}$$

where the constants C_1 and C_2 do not depend on ε ($0 < C_1 < C_2 < \infty$), and that as $\varepsilon \to 0$ the particles densely fill the domain Ω (get into any finite subdomain for

sufficiently small ε). The condition (3.1) means that the particles do not form clusters and do not come too close to the boundary $\partial\Omega$.

We next present two more conditions which describe the nature and the magnitude of the interaction between the particles. First, we consider only short-range interactions so that the matrix-functions $\underline{\underline{C}}^{ij}_{\varepsilon}(\underline{x},\underline{y})$ (see (2.1)) satisfy the following condition:

(a_3)

$$\underline{\underline{C}}^{ij}_{\varepsilon}(\underline{x},\underline{y}) \equiv 0 \quad \text{if dist } (Q^i_\varepsilon, Q^j_\varepsilon) \geq C\varepsilon, \tag{3.2}$$

where $0 < C < \infty$ does not depend on ε. Here the constant C is chosen in such a way that each particle interacts with its nearest neighbors only, for an appropriate definition of the nearest neighbor.

Second, we consider the case when the entries of the matrix-functions $\underline{\underline{C}}^{ij}_{\varepsilon}(\underline{x},\underline{y})$ are of order $O(\varepsilon^{-3})$. More precisely, the scalar function $a^{ij}_{\varepsilon}(\underline{x},\underline{y})$ in the condition (2.3′) can be written as follows:

(a_4)

$$a^{ij}_{\varepsilon}(\underline{x},\underline{y}) = \varepsilon^{-3} a^{ij}_0(\underline{x},\underline{y}), \tag{3.3}$$

where $a^{ij}_0(\underline{x},\underline{y})$ is a bounded nonnegative function.

We remark here that our proof applies also for the weaker interactions $a^{ij}_{\varepsilon} = o(\varepsilon^{-3})a^{ij}_0$; however, the interactions of order ε^{-3} lead to the most interesting rheological properties of the effective medium (viscoelasticity and memory), whereas weaker interactions lead to the effective viscosity with no memory.

We introduce a mesoscopic characteristic of the suspension (the fluid-particle compound). This *mesocharacteristic* plays the key role in our consideration. Roughly speaking, it allows us to compute the energy of the compound in some mesoscopic cube of size $h : \varepsilon \ll h \ll 1$, which is a so-called representative volume element. In other words, the size h is much larger than the microscale ε and much smaller than the macroscopic size of the system so that effective properties in this volume element provide correct rheological properties of the compound locally. The mesocharacteristic is motivated by experimental techniques in which one cuts out a certain piece of a compound and measures its properties in order to evaluate the overall properties of the compound.

We now define the mesocharacteristic rigorously using the variational formulation presented below in (4.8) (parameter λ is introduced in (4.1)–(4.6)).

Let $K^{\underline{\xi}}_h = K(\underline{\xi}, h)$ be a cube of side length $h > 0$ such that $\varepsilon \ll h \ll 1$ centered at a point $\underline{\xi} \in \Omega$. The orientation of the cube is arbitrary but independent of $\underline{\xi}$ and h. For the sake of definiteness we assume that the edges of this cube are parallel to the coordinate axis.

Introduce the following class of functions

$$\mathcal{J}_\varepsilon[K^{\underline{\xi}}_h] = \{\underline{w}_\varepsilon \in H^1(K^{\underline{\xi}}_h), \quad \underline{w}_\varepsilon = \underline{a}^i_\varepsilon + \underline{b}^i_\varepsilon \times (\underline{x} - \underline{x}^i_\varepsilon) \quad \text{on } Q^i_\varepsilon \cap K^{\underline{\xi}}_h\}, \tag{3.4}$$

where $\underline{a}^i_\varepsilon$ and $\underline{b}^i_\varepsilon$ are arbitrary constant vectors, and consider a minimization problem in the class $\mathcal{J}_\varepsilon(K^{\underline{\xi}}_h)$ for the following functional:

$$\begin{aligned} &A_{\varepsilon h}[\underline{u}_\varepsilon, \underline{\underline{T}}, \underline{\xi}, \lambda, \tau] \\ &\quad = E_{K^{\underline{\xi}}_h}[\underline{u}_\varepsilon, \underline{u}_\varepsilon] + h^{-2-\tau} \int_{K_h{}^{\underline{\xi}}} \left| \underline{u}_\varepsilon - \sum_{p,q=1}^{3} \underline{\psi}^{pq}(\underline{x} - \underline{\xi}) T_{pq} \right|^2 dx + \frac{1}{\lambda} I^{\varepsilon}_{K^{\underline{\xi}}_h}[\underline{u}_\varepsilon, \underline{u}_\varepsilon], \end{aligned} \tag{3.5}$$

where $e_{k\ell}[\underline{u}] = \frac{1}{2}[\frac{\partial u_k}{\partial x_\ell} + \frac{\partial u_\ell}{\partial x_k}]$ is the deformation rate of the fluid moving with velocity $\underline{u}$, $\underline{\underline{T}} = \{T_{pq}\}$ is an arbitrary symmetric second rank tensor with constant components T_{pq}, λ and τ are arbitrary positive numbers, and the matrix-functions $\underline{\underline{C}}^{ij}_{\varepsilon}(\underline{x}, \underline{y})$ are defined in (2.1).

Hereafter we use the following notation:

$$E_G[\underline{u}_\varepsilon, \underline{v}_\varepsilon] = 2\mu \int_G \int \sum_{k,l=1}^{3} e_{kl}[\underline{u}_\varepsilon] e_{kl}[\underline{v}_\varepsilon] d\underline{x}, \tag{3.6}$$

$$I^{\varepsilon}_G[\underline{u}_\varepsilon, \underline{v}_\varepsilon] = \frac{1}{2} \sum_{ij} \int_{S^i_\varepsilon} \int_{S^j_\varepsilon} \langle \underline{\underline{C}}^{ij}_{\varepsilon}(\underline{u}(\underline{x}) - \underline{u}(\underline{y})), \underline{v}(\underline{x}) - \underline{v}(\underline{y}) \rangle dS^i_x dS^j_y, \tag{3.7}$$

$$\underline{\psi}^{pq}(\underline{x}) = \frac{1}{2}(x_p \underline{e}^q + x_q \, \underline{e}^p) - \frac{\delta_{pq}}{3} \sum_{n=1}^{3} x_n \underline{e}^n \,, \tag{3.8}$$

where $p, q = 1, 2, 3$; $\underline{e}^q$ are the basis vectors so that $\underline{x} = \sum_{q=1}^{3} x_q \underline{e}^q$; and δ_{pq} stands for the Kronecker delta symbol. It is easy to check that $\operatorname{div} \underline{\psi}^{pq} = 0$. The sum $\sum_{ij}$ in (3.7) is taken over all particles located inside the domain G.

LEMMA 3.1. *For any constant tensor $\underline{\underline{T}}$ there exists the unique minimizer $\underline{u}(\underline{x}) = \underline{u}(\underline{x}, \underline{\underline{T}})$ (dependence on the parameters $\varepsilon, h, \underline{\xi}, \lambda$, and τ is suppressed) of the functional* (3.5) *in the class $\mathcal{J}_\varepsilon[K^{\underline{\xi}}_h]$. The minimum value of this functional is a quadratic function of the tensor $\underline{\underline{T}} = \{T_{pq}\}$, and the following representation holds:*

$$\min_{\underline{u}_\varepsilon \in \mathcal{J}_\varepsilon[K^{\underline{\xi}}_h]} A_{\varepsilon h}[\underline{u}_\varepsilon, \underline{\underline{T}}, \underline{\xi}, \lambda, \tau] = \sum_{npqr=1}^{3} a_{npqr}(\underline{\xi}, \lambda, \varepsilon, h, \tau) T_{np} \, T_{qr} \,, \tag{3.9}$$

where $a_{npqr}(\xi, \lambda, \varepsilon, h, \tau)$ are the components of a 4th rank tensor, defined as

$$\begin{aligned} a_{npqr}(\underline{\xi}, \lambda; \varepsilon, h, \tau) &= E_{K^{\underline{\xi}}_h}[\underline{u}^{np}, \underline{u}^{qr}] \\ &+ \int_{K^{\underline{\xi}}_h} h^{-2-\tau} \langle \underline{u}^{np}(\underline{x}) - \underline{\psi}^{np}(\underline{x} - \underline{\xi}), \underline{u}^{qr}(\underline{\xi}) - \underline{\psi}^{qr}(\underline{x} - \xi) \rangle d\underline{x} + \frac{1}{\lambda} I_{K^{\underline{\xi}}_h}[\underline{u}^{np}, \underline{u}^{qr}]. \end{aligned} \tag{3.10}$$

Here $\underline{u}^{np}(\underline{x})$ is the minimizer of the functional $A_{\varepsilon h}[\underline{u}, \underline{\underline{T}}, \lambda, \underline{\xi}, \tau]$ when $\underline{\underline{T}} = \frac{1}{2}(\underline{e}^n \otimes \underline{e}^p + \underline{e}^p \otimes \underline{e}^n)$.

This lemma is proved in [1]. We now briefly outline the proof. First we show that the class $\mathcal{J}_\varepsilon[K^{\underline{\xi}}_h]$ is not empty by an explicit construction of a function $\underline{\psi}^{np}_{\varepsilon}(\underline{x})$ from this class. This is done by an appropriate modification of the function $\underline{\psi}^{np}(\underline{x} - \underline{\xi})$. Next we generalize this construction to obtain a dense in $\mathcal{J}_\varepsilon[K^{\underline{\xi}}_h]$ set of functions starting from an arbitrary function $w(\underline{x}) \in C^1(K^{\underline{\xi}}_h)$. Using the function $\underline{\psi}^{np}_{\varepsilon}(\underline{x})$, we construct a comparison function $\hat{\underline{u}}_\varepsilon(\underline{x}, \underline{\underline{T}}) = \sum_{np} \underline{\psi}^{np}_{\varepsilon}(\underline{x}) T_{np} \in \mathcal{J}_\varepsilon[K^{\underline{\xi}}_h]$ for an arbitrary constant tensor $\underline{\underline{T}} = \{T_{np}\}$. Substitution of this function into the functional (3.9) provides a uniform in $\varepsilon \le \hat{\varepsilon}(h)$ upper bound:

$$0 < A_{\varepsilon h}[\hat{\underline{u}}_\varepsilon, \underline{\underline{T}}, \underline{\xi}, \lambda, \tau] < Ch^3. \tag{3.11}$$

Now we can use standard techniques to establish existence and compactness of a minimizing sequence, which implies existence of a minimizer $\underline{u}_\varepsilon$ of the problem (3.9); the lemma follows.

From (3.10) we obtain that the tensor $\{a_{npqr}(x,\lambda;\varepsilon,h,\tau)\}$ is invariant with respect to the permutation of pairs of indexes as well as permutation of indexes in each pair; that is,

$$a_{npqr} = a_{qrnp} = a_{pnqr} = \cdots .$$

This tensor describes macroscopic rheological properties of the suspension (compound). It depends on the space variable $\underline{x} \in \Omega$ if the suspension is inhomogeneous on the macroscale (e.g., for periodic structures there is no dependence on $\underline{x}$; see section 7), and on the spectral parameter λ.

We provide a heuristic basis for the introduction of the mesocharacteristic (3.5).

First, as we have mentioned earlier, if a compound can be described within the effective single medium approach, then the rheological properties of the compound can be determined by calculation or measurements in some representative volume element of an intermediate mesoscale h, which is why we choose cube $K_h^{\underline{\xi}}$.

Next, observe that the sum of the first and the third term in (3.5) represents the energy of the compound (suspension). The minimizer u_ε of (3.5) is "close," up to an additive constant, to the true global minimizer $\underline{v}_\varepsilon$ of the variational problem, which corresponds to (2.10)–(2.18) if the tensor $\underline{\underline{T}}$ is chosen appropriately. Now one should choose $\underline{\underline{T}}$. If the single medium homogenized description is possible, then $\underline{v}_\varepsilon(\underline{x})$ is "close" to some smooth (homogenized) function $\underline{v}(\underline{x})$, which depends only on macroscopic variable $\underline{x}$ and does not depend on ε, so that it does not vary on the microscale ε. We then minimize the energy of the compound, adding the constraint that the minimizer $\underline{u}_\varepsilon$ is "close" to the linear part $L(\underline{x})$ (the differential) of the global minimizer $\underline{v}(\underline{x})$, so that $|\underline{u}_\varepsilon - \underline{v}(\underline{x})| = \circ(h) \sim h^{1+\tau/2}$, $h \to 0$ for some $\tau > 0$. This condition is imposed by introducing the penalty term (the second term in (3.5)).

We consider arrays of the particles Q_ε^i such that for all $\underline{x} \in \Omega$, each $\lambda > 0$, and some real number $\tau > 0$ the following limits exist:

$$(\mathrm{b}_1) \quad \lim_{h\to 0} \underline{\lim_{\varepsilon\to 0}} \frac{a_{npqr}(\underline{x},\lambda;\varepsilon,h,\tau)}{h^3} = \lim_{h\to 0} \overline{\lim_{\varepsilon\to 0}} \frac{a_{npqr}(\underline{x},\lambda;\varepsilon,h,\tau)}{h^3} = a_{npqr}(\underline{x},\lambda),$$

where $a_{npqr}(\underline{x},\lambda)$ is a continuous function of $\underline{x} \in \Omega$ and $\lambda > 0$ is a parameter.

LEMMA 3.2. *If the condition* (b_1) *holds for some* $\tau > 0$, *then it also holds for any* $\tau > 0$ *and the limit* (b_1) *does not depend on* τ. *The functions* $a_{npqr}(\underline{x},\lambda)$ *are defined for* $\lambda > 0$ *and can be analytically extended into the complex plane with a cut along the negative semiaxis* $\lambda \le 0$. *The extended functions can be represented in the form*

$$a_{npqr}(\underline{x},\lambda) = a^0_{npqr}(\underline{x}) + a^1_{npqr}(\underline{x},\lambda)$$

so that in the domain $\Phi_\delta = \{\lambda \in \mathbf{C} : |arg\lambda - \pi| \ge \delta > 0\}$ *for any* $\delta > 0$ *the following bound holds:*

$$|a^1_{npqr}(\underline{x},\lambda)| < C\left(\frac{1}{|\lambda|^{1/2}}\right) \quad as\ \lambda \to \infty, \tag{3.12}$$

where $C > 0$ *does not depend on* λ.

This lemma is proved in [1]. The key idea of the proof is to present the Euler–Lagrange boundary value problem, which corresponds to the minimization problem

(3.9), in abstract operator form (a Friedrichs extension) in the appropriate Hilbert space $\tilde{\mathcal{J}}_\varepsilon[K_h^\xi]$. Roughly speaking, the space $\tilde{\mathcal{J}}_\varepsilon[K_h^\xi]$ consists of functions from $W_2^2(K_h^\xi)$ which satisfy the boundary conditions of the Euler–Lagrange problem and the divergence-free conditions. Due to the presence of the factor $1/\lambda$ in (3.5), the obtained operator equation in the Hilbert space $\tilde{\mathcal{J}}_\varepsilon[K_h^\xi]$ is an operator pencil. Then analyticity and estimate (3.12) follow from known results for operator pencils in an abstract Hilbert space [9, Chapter 7].

It follows from Lemma 3.2 that the functions $a_{npqr}(\underline{x},\lambda)$ are inverse Laplace transforms

$$a_{npqr}(\underline{x},\lambda)=\int_0^\infty e^{-\lambda t}\widehat{a}_{npqr}(x,t)dt \tag{3.13}$$

of the functions

$$\widehat{a}_{npqr}(\underline{x},t)=a^0_{npqr}(\underline{x})\delta(t)+\widehat{a}^1_{npqr}(\underline{x},t), \tag{3.14}$$

where $\delta(t)$ is the Dirac delta-function and $\widehat{a}^1_{npqr}(\underline{x},t)$ are locally summable in t for $t\geq 0$.

We introduce one more characteristic which describes the distribution of the masses in the suspension.

Denote by $\chi_\varepsilon(\underline{x})$ the characteristic function of the domain Ω_ε, occupied by the fluid, and by $\chi^i_\varepsilon(\underline{x})$-the characteristic function of a particle Q^i_ε, and introduce

$$\rho_\varepsilon(\underline{x})=\rho_f\chi_\varepsilon(\underline{x})+\rho_s\sum_{i=1}^{N_\varepsilon}\chi^i_\varepsilon(\underline{x}). \tag{3.15}$$

We consider the arrays of particles such that for any $\underline{x}\in\Omega$ the following limit exists:

$$\lim_{h\to\infty}\lim_{\varepsilon\to 0}\frac{1}{h^3}\int_{K^{\underline{x}}_h}\rho_\varepsilon(\underline{\xi})d\xi=\rho(\underline{x}), \tag{b_2}$$

where $\rho(\underline{x})>0$ is a continuous function which describes the limiting density of the suspension. Note that the condition (b_2) holds for very generic geometries. In particular, for periodic structures (section 7) $\rho(\underline{x})=$ const. and is easy to compute.

We are now in a position to formulate the main mathematical result of this paper and discuss its physical consequences. We construct the vector function

$$\underline{\tilde{v}_\varepsilon}(\underline{x},t)=\chi_\varepsilon(\underline{x})\underline{v}_\varepsilon(\underline{x},t)+\sum_i\chi^i_\varepsilon(\underline{x})\left[\underline{\dot{u}}^i_\varepsilon+\underline{\dot{\theta}}^i_\varepsilon(t)\times(\underline{x}-\underline{x}^i_\varepsilon)\right] \tag{3.16}$$

using the solutions $\{\underline{v}_\varepsilon(\underline{x},t),\underline{u}^i_\varepsilon(t),\theta^i_\varepsilon(t),\ i=1,\dots N_\varepsilon\}$ of the problem (3.1)–(3.9).

THEOREM 3.3. *Let the conditions* (a_1)–(a_4) *and* (b_1)–(b_2) *hold. Suppose that the initial functions in the problem* (2.10)–(2.17) *converge as* $\varepsilon\to 0$ *to a vector function* $\underline{v}_0(\underline{x})\in L_2(\Omega)$,

$$\underline{\tilde{v}_\varepsilon}(\underline{x},0)\to\underline{v}_0(\underline{x})\quad \textit{in } L_2(\Omega)\quad \textit{as } \varepsilon\to 0, \tag{3.17}$$

where $\underline{\tilde{v}_\varepsilon}(\underline{x},0)=\underline{\tilde{v}_\varepsilon}(\underline{x},t)|_{t=0}$ *and* $\underline{\tilde{v}_\varepsilon}(\underline{x},t)$ *is defined by* (3.16).

Then the functions $\underline{v}_\varepsilon(x,t)$ *converge in* $L_2(\Omega)$ *for all* $t<\infty$ *to a vector function* $\underline{v}(\underline{x},t)$*, which is a solution of the following homogenized problem:*

$$\rho(\underline{x})\frac{\partial \underline{v}}{\partial t}-\left\{\sum_{n,p,q,r}\frac{\partial}{\partial x_p}[a^0_{npqr}(\underline{x})e_{qr}[\underline{v}]] + \int_0^t \widehat{a}^1_{npqr}(x,t-\tau)e_{qr}[\underline{v}(\underline{x},\tau)]d\tau\right\}\underline{e}^n=\nabla p,\quad x\in\Omega,\ t>0, \tag{3.18}$$

$$\operatorname{div}\underline{v}=0,\quad \underline{x}\in\Omega,\quad t>0, \tag{3.19}$$

$$\underline{v}(\underline{x},t)=f(\underline{x},t),\quad \underline{x}\in\partial\Omega,\ t>0, \tag{3.20}$$

$$\underline{v}(\underline{x},0)=\underline{v}_0(\underline{x}),\quad \underline{x}\in\Omega, \tag{3.21}$$

where the components of the tensors $\{a^0_{npqr}(\underline{x})\}$ *and* $\{\widehat{a}^1_{npqr}(\underline{x},t)\}$ *are defined by* (3.9)–(3.14) *and the condition* (b_1).

In order to explain the physical meaning of the homogenized problem it is convenient to rewrite (3.18)–(3.19) in terms of the displacements $\underline{u}(\underline{x},t)=\int_0^t \underline{v}(\underline{x},t)d\tau$. Then

$$\rho(x)\frac{\partial^2 \underline{u}}{\partial t^2}-\left\{\sum_{n,p,q,r=1}^{3}\frac{\partial}{\partial x_p}[a^0_{npqr}(\underline{x})\dot{e}_{qr}[\underline{u}]] + \int_0^t \widehat{a}_{npqr}(\underline{x},t-\tau)\dot{e}_{qr}[\underline{u}]d\tau\right\}\underline{e}^n=\nabla p, \tag{3.18$'$}$$

$$\operatorname{div}\underline{u}=0, \tag{3.19$'$}$$

where $\dot{e}_{qr}[\underline{u}]=\frac{\partial}{\partial t}e_{qr}$.

Recall (see [8]) that the strain-stress relation in linear viscoelasticity has the form

$$\underline{\underline{\sigma}}(t)=\int_{-\infty}^{t}C(t-\tau)\underline{\underline{\dot{e}}}(\tau)d\tau, \tag{3.22}$$

where $C(t)$ is the fourth rank relaxation tensor. If $C(t)$ is a delta-function $k\delta(t)$, then $\underline{\underline{\sigma}}=k\underline{\underline{\dot{e}}}$ and we obtain the constitutive relation for a Newtonian fluid. If $C(t)$ does not depend on t, then (3.22) reduces to Hooke's law for an elastic solid. Thus the second term (the sum) in (3.18′) or (3.18) describes the effective Newtonian fluid with the effective viscosity tensor a^0_{npqr}, while the third term (the integral) describes the effective viscoelastic behavior with the effective relaxation tensor $\hat{a}^1_{npqr}$. Now we see that the homogenized equations (3.18′) or (3.18) suggest the following qualitative picture for the effective single phase medium. On a short time scale the integral terms in (3.18′) or (3.18) are small, and therefore the homogenized problem describes an incompressible fluid with anisotropic viscosity tensor a^0_{npqr}. On a longer time scale (intermediate scale) both the viscosity term and the relaxation term become significant, and the homogenized medium is an isotropic viscoelastic fluid with memory

(relaxation). Finally, on a very large time scale, provided that $f(x,t)$ has finite support in time, $\underline{u}(\underline{x},t)$ and $\underline{v}(\underline{x},t)$ become time independent, and therefore the $\dot{e}_{qr}(\underline{u})$ term is negligible. In addition, the kernel $\widehat{a}_{npqr}(\underline{x},t-\tau)$ also becomes time independent, and integration in (3.18′) or (3.18) yields to the equations for an incompressible elastic medium.

The tensors $\{a^0_{npqr}(\underline{x})\}$ and $\{\widehat{a}^1_{npqr}(\underline{x},t)\}$ are called the *effective viscosity* and the *effective relaxation* tensors, respectively [7]. Both tensors incorporate the information about the geometric array of the particles, the strength of the interparticle interactions, and the viscosity of the fluid phase.

We now provide a heuristic explanation of the choice of scaling in the condition (a_4). If the entries of the pairwise interaction matrices are such that the pairwise interaction energy is $\sup_{\|u\|=1}\langle \underline{\underline{C}}^{ij}_{1\varepsilon}\underline{u},\underline{u}\rangle \sim \varepsilon$, then both tensors $\{a^0_{npqr}(\underline{x})\}$ and $\{\widehat{a}^1_{npqr}(\underline{x},t)\}$ are positive and finite, which implies that the "fluid-particle" suspension in the homogenization limit behaves as a viscoelastic incompressible medium.

If these matrices have large entries $\sup_{\|u\|=1}\langle \underline{\underline{C}}^{ij}_{1}\underline{u},\underline{u}\rangle\varepsilon^{-1} \to \infty$ as $\varepsilon \to 0$, then the homogenized medium becomes absolutely rigid, so that if the velocity at the external boundary $\partial\Omega$ is zero, then the displacements and velocities of all points in Ω are zero (in the homogenized limit $\varepsilon \to 0$). Finally, if $\sup_{\|u\|=1}\langle \underline{\underline{C}}^{ij}_{1\varepsilon}\underline{u},\underline{u}\rangle \ll \varepsilon$, then the interaction does not affect the homogenized medium.

In sections 4–5 we prove Theorem 3.3. To this end we use the Laplace transform to obtain a time independent analogue of the problem (2.10)–(2.17) (with the spectral parameter $\lambda > 0$) and then analyze a variational formulation of this problem. Next we study the asymptotic behavior of the variational problem as $\varepsilon \to 0$, obtain the limiting (homogenized) functional, and write down the corresponding Euler–Lagrange equations. In this presentation we formulate techinical lemmas and estimates, whose detailed proofs can be found in [1].

In section 6 we discuss analytical dependence of the solutions of these equations in the parameter λ, and, by taking the inverse Laplace transform, we obtain the homogenized time dependent problem.

In section 7 we consider a periodic structure, prove existence of the limits in (b_2), and show that the computation of the viscosity and the relaxation tensors amounts to solving a cell problem. The cell problem can be solved using standard numerical techniques. We also analyze this problem analytically and obtain the distribution of the relaxation times, which is the main quantity of interest in rheological studies of filled polymers and suspensions.

4. Variational formulation of the homogenization problem. Take the Laplace transform $t \to \lambda$ in the problem (2.10)–(2.17). For simplicity we keep the same notation: $\underline{v}_\varepsilon(\underline{x},t) \to \underline{v}_\varepsilon(\underline{x},\lambda)$, $p_\varepsilon(\underline{x},t) \to p_\varepsilon(\underline{x},\lambda)$, $\underline{u}^i_\varepsilon(t) \to \underline{u}^i_\varepsilon(\lambda)$, $\underline{\theta}^i_\varepsilon(t) \to \underline{\theta}^i_\varepsilon(\lambda)$, $\underline{f}(\underline{x},t) \to \underline{f}(\underline{x},\lambda)$. Then, taking into account (2.8), (2.1), (2.5), (2.7), and (2.16), we obtain the following problem:

$$-\mu\Delta\underline{v}_\varepsilon + \lambda\rho_f\underline{v}_\varepsilon - \nabla p_\varepsilon = \rho_f\underline{v}_{\varepsilon 0}(\underline{x}), \quad \underline{x}\in\Omega_\varepsilon, \tag{4.1}$$

$$\operatorname{div}\underline{v}_\varepsilon = 0, \quad \underline{x}\in\Omega, \tag{4.2}$$

$$\underline{v}_\varepsilon = \lambda[\underline{u}^i_\varepsilon + \underline{\theta}^i_\varepsilon \times (\underline{x}-\underline{x}^i_\varepsilon)], \quad \underline{x}\in Q^i_\varepsilon, \tag{4.3}$$

$$\lambda^2 m_\varepsilon^i \underline{u}_\varepsilon^i = -\int_{S_\varepsilon^i} \underline{\underline{\sigma}}[\underline{v}_\varepsilon]\cdot\underline{\nu} dS - \frac{1}{\lambda}\sum_{j,j\neq i}\int_{S_\varepsilon^i}\int_{S_\varepsilon^j} \underline{\underline{C}}_\varepsilon^{ij}(\underline{x},\underline{y})[\underline{v}_\varepsilon(\underline{x})-\underline{v}_\varepsilon(\underline{y})]dS_x dS_y + m_\varepsilon^i \underline{u}_{\varepsilon 1}^i, \quad i=1,2,\dots N_\varepsilon, \tag{4.4}$$

$$\lambda^2 I_\varepsilon^i \underline{\theta}_\varepsilon^i = -\int_{S_\varepsilon^i} (\underline{x}-\underline{x}_\varepsilon^i)\times\underline{\underline{\sigma}}[\underline{v}_\varepsilon]\cdot\underline{\nu} dS - \frac{1}{\lambda}\sum_{j,j\neq i}\int_{S_\varepsilon^i}\int_{S_\varepsilon^i} \underline{\underline{C}}_\varepsilon^{ij}(\underline{x},\underline{y})[\underline{v}_\varepsilon(\underline{x})-\underline{v}_\varepsilon(\underline{y})]dS_x dS_y + I_\varepsilon^i \underline{\theta}_\varepsilon^i, \quad i=1,2,\dots N_\varepsilon, \tag{4.5}$$

$$\underline{v}_\varepsilon = \underline{f}(\underline{x},\lambda) \quad \text{on } \partial\Omega. \tag{4.6}$$

We extend the velocity function $\underline{v}_\varepsilon(\underline{x},\lambda)$ onto the particles according to (4.3) and keep the same notation $\underline{v}_\varepsilon = \underline{v}_\varepsilon(\underline{x},\lambda)$ for the extended functions. Then $\underline{v}_\varepsilon \in H^1(\Omega)$ and $\operatorname{div}\underline{v}_\varepsilon = 0$ in Ω. Denote by $\mathcal{J}_\varepsilon^f(\Omega)$ the class of divergence-free vector functions from $H^1(\Omega)$, which satisfy the rigid displacement conditions (2.2) on the particles Q_ε^i and take the prescribed values $\underline{f}(\underline{x},\lambda)$ on $\partial\Omega$.

Consider the minimization problem

$$\Phi_\varepsilon[\underline{v}_\varepsilon] = \min_{\underline{v}'_\varepsilon\in\mathcal{J}_\varepsilon^f(\Omega)} \Phi_\varepsilon[\underline{v}'_\varepsilon] \tag{4.7}$$

for the following functional:

$$\Phi_\varepsilon[\underline{v}_\varepsilon] = E_\Omega[\underline{v}_\varepsilon,\underline{v}_\varepsilon] + \int_\Omega [\lambda\langle\rho_\varepsilon\underline{v}_\varepsilon,\underline{v}_\varepsilon\rangle - 2\langle\rho_\varepsilon\underline{v}_{\varepsilon 0},\underline{v}_\varepsilon\rangle]d\underline{x} + \frac{1}{\lambda} I_\Omega^\varepsilon[\underline{v}_\varepsilon,\underline{v}_\varepsilon], \tag{4.8}$$

where $\lambda>0$, $\tilde{\underline{v}}_{\varepsilon 0} = \tilde{\underline{v}}_\varepsilon(\underline{x},0)$, $\rho_\varepsilon(\underline{x})$, and $\underline{v}_\varepsilon(\underline{x},t)$ are defined in (3.16), (3.17).

LEMMA 4.1. *There exists a unique minimizer $\underline{v}_\varepsilon = \underline{v}_\varepsilon(\underline{x},\lambda)$ of the functional* (4.8) *in the class $\mathcal{J}_\varepsilon^f(\Omega)$. This minimizer provides the solution $\{\underline{v}_\varepsilon(\underline{x},\lambda)\chi_\varepsilon(\underline{x}), \underline{v}_\varepsilon(\underline{x},\lambda)\chi_\varepsilon^i(\underline{x})\}$ of the boundary value problem* (4.1)–(4.6).

This lemma can be proved by standard techniques of calculus of variations [10], [18].

Introduce the homogenized functional

$$\Phi_0[\underline{v}] = \int_\Omega \left\{ \sum_{k,\ell=1}^3 a_{npqr}(\underline{x},\lambda) e_{np}[\underline{v}] e_{qr}[\underline{v}] + \lambda\langle\rho\underline{v},\underline{v}\rangle - 2\langle\rho\underline{v}_0,\underline{v}\rangle \right\} dx, \tag{4.9}$$

where the tensor $a_{npqr}(\underline{x},\lambda)$ and the function $\rho=\rho(\underline{x})$ are defined by the conditions (b_1)–(b_2), and the vector-function $\underline{v}_0(\underline{x})$ is defined by the condition (3.17).

Denote by $\mathcal{J}^f(\Omega)$ a class of the divergence-free functions from $W_2^1(\Omega)$ that are equal to $\underline{f}(\underline{x},\lambda)$ on $\partial\Omega$, and introduce the variational problem for the functional (4.9) in this class:

$$\Phi_0[\underline{v}] = \min_{\underline{v}'_\varepsilon\in\mathcal{J}^f(\Omega)} \Phi[\underline{v}']. \tag{4.10}$$

THEOREM 4.2. *Suppose that the conditions* (a_1)–(a_4), (b_1)–(b_2), *and* (3.17) *hold. Then the minimizers* $\underline{v}_\varepsilon(\underline{x},\lambda)$ *of the problem* (4.7) *converge weakly in* $H^1(\Omega)$ *(strongly in* $L_2(\Omega)$*) as* $\varepsilon \to 0$ *to the minimizer* $\underline{v}(\underline{x},\lambda)$ *of the homogenized problem* (4.10).

This theorem is the key result of this paper, and its proof is outlined in section 5. It follows from standard techniques of calculus of variations that if the functions $a_{npqr}(\underline{x},\lambda)$ are sufficiently smooth (e.g., in $C^1(\Omega)$), then the minimizer $\underline{v}(\underline{x},\lambda)$ of the problem (4.10) is the solution of the following boundary value problem:

$$-\sum_{n,p,q,r} \frac{\partial}{\partial x_r}\left[a_{npqr}(\underline{x},\lambda)e_{np}[\underline{v}]\right]\underline{e}^n + \lambda\rho\underline{v} - \nabla p = \rho\underline{v}_0\,, \quad x \in \Omega, \tag{4.11}$$

$$\operatorname{div}\underline{v} = 0, \tag{4.12}$$

$$\underline{v}(\underline{x},\lambda) = \underline{f}(\underline{x},\lambda), \qquad x \in \partial\Omega. \tag{4.13}$$

Remark 4.1. If $a_{npqr}(\underline{x},\lambda)$ is not sufficiently smooth in $\underline{x}$, then $\underline{v}(\underline{x},\lambda)$ is a generalized (weak) solution of the problem (4.11)–(4.13).

Applying the inverse Laplace transform, one can see that the main homogenization Theorem 3.3 for the time-dependent problem is obtained from this theorem. The key step here is to establish necessary analytical properties in λ of the functions $a_{npqr}(\underline{x},\lambda)$ and $\underline{v}(\underline{x},\lambda)$ for complex λ, which is done in [1] and briefly outlined in section 6 below.

5. Convergence theorem for stationary problems.

5.1. Compactness of the solutions $\{v_\epsilon,\ \epsilon > 0\}$ of the problem (4.7) in $H^1(\Omega)$. We use the following lemma to obtain the boundedness of the solutions $\{v_\epsilon,\ \epsilon > 0\}$.

LEMMA 5.1. *Suppose that the domain* $\Omega_\epsilon = \Omega \setminus \cup_i Q^i_\epsilon$ *satisfies the conditions* (a_1)–(a_2). *Then for any function* $\underline{f}(x) \in C^2(\partial\Omega)$ *satisfying the condition* (2.18) *there exists a function* $\underline{f}_\epsilon(\underline{x}) \in H^1(\Omega)$ *such that* $\operatorname{div}\underline{f}_\epsilon(\underline{x}) = 0$ *in* Ω, $\underline{f}_\epsilon(\underline{x}) = \underline{f}(\underline{x})$ *on* $\partial\Omega$, $\underline{f}_\varepsilon(\underline{x}) = \underline{f}^i_\varepsilon$ *on the minimal balls* $B(Q^i_\varepsilon)$, *and*

$$\|\underline{f}_\varepsilon\|_{H^1(\Omega)} < C, \qquad I^\varepsilon_\Omega[\underline{f}_\varepsilon, \underline{f}_\varepsilon] < C, \tag{5.1}$$

where $C > 0$ *does not depend on* ε.

We apply this lemma to the function $f(\underline{x},\lambda)$, which is Laplace transform of the boundary function $\underline{f}(\underline{x},t)$ (from (2.17)) in time. Then, we find a function $\underline{f}_\varepsilon(\underline{x}) \in \mathcal{J}^f_\varepsilon(\Omega)$ (admissible for the problem (4.7)) and satisfying (5.1).

Let $\underline{v}_\varepsilon = \underline{v}_\varepsilon(\underline{x},\lambda)$ be a solution of the problem (4.7); i.e., the functional Φ_ε in the class $\mathcal{J}^f_\varepsilon(\Omega)$ attains its minimum on $\underline{v}_\varepsilon$. Since $\underline{f}_\varepsilon \in \mathcal{J}^f_\varepsilon(\Omega)$, the following inequality holds:

$$\Phi_\varepsilon[\underline{v}_\varepsilon] \le \Phi_\varepsilon[\underline{f}_\varepsilon].$$

With this inequality, taking into account (4.8) and the nonnegativity of the matrices $\underline{\underline{C}}^{ij}_\varepsilon(\underline{x})$, we obtain

$$E_\Omega[\underline{v}_\varepsilon, \underline{v}_\varepsilon] + 2\mu\int_\Omega \lambda\langle\rho_\varepsilon\underline{v}_\varepsilon, \underline{v}_\varepsilon\rangle d\underline{x} \le |\Phi_\varepsilon[\underline{f}_\varepsilon]| + 2\|\rho_\varepsilon\tilde{\underline{v}}_{\varepsilon 0}\|_{L_2(\Omega)}\|\underline{v}_\varepsilon\|_{L_2(\Omega)}. \tag{5.2}$$

Next we use the second Korn inequality (see [13])

$$\|\underline{v}_\varepsilon\|^2_{H^1(\Omega)} \le \mathcal{C}\left(E_\Omega[\underline{v}_\varepsilon, \underline{v}_\varepsilon] + \|\underline{v}_\varepsilon\|_{L_2(\Omega)}\right), \tag{5.3}$$

where $\mathcal{C}$ depends on domain Ω only. Then, since $\mu > 0$, $\lambda > 0$, $\rho_\varepsilon(\underline{x}) \ge \min\{\rho_f, \rho_s\} > 0$, we find from (5.2)

$$\|\underline{v}_\varepsilon\|^2_{H^1(\Omega)} \le \mathcal{C}_1\left(|\Phi_\varepsilon[\underline{f}_\varepsilon]| + \|\rho_\varepsilon \tilde{\underline{v}}_{\varepsilon 0}\|_{L_2(\Omega)}\|\underline{v}_\varepsilon\|_{H^1(\Omega)}\right),$$

from which it follows that

$$\|\underline{v}_\varepsilon\|_{H^1(\Omega)} \le \mathcal{C}_2\left(|\Phi_\varepsilon(\underline{f}_\varepsilon)|^{\frac{1}{2}} + \|\rho_\varepsilon \tilde{\underline{v}}_{\varepsilon 0}\|_{L_2(\Omega)}\right), \tag{5.4}$$

where the constants $\mathcal{C}_1$ and $\mathcal{C}_2$ do not depend on ε, $\tilde{\underline{v}}_{\varepsilon 0} = \tilde{\underline{v}}_\varepsilon(\underline{x}, 0)$ and we have used the notation (3.6).

Now we use the second inequality in (5.1), the convergence (3.17) of $\tilde{\underline{v}}_{\varepsilon 0}$, and the uniform boundedness of $\rho_\varepsilon(\underline{x})$ in ε to obtain $\Phi_\varepsilon[\underline{f}_\varepsilon] < C$. Combing the latter inequality with (5.4), we get

$$\|\underline{v}_\varepsilon\|_{H^1(\Omega)} < C.$$

Therefore the set of the functions $\{\underline{v}_\varepsilon(\underline{x}), \varepsilon > 0\}$ is weakly compact in $H^1(\Omega)$.

Let us select a sequence $\{\underline{v}_{\varepsilon_k}(\underline{x}), \varepsilon_k \to 0\}$ weakly convergent in $H^1(\Omega)$ to the function $\underline{v}(\underline{x}) \in H^1(\Omega)$.

Due to embedding theorem, $\underline{v}(\underline{x}) = \underline{f}(\underline{x})$ at $\underline{x} \in \partial\Omega$ and $\underline{v}_{\varepsilon_k}(\underline{x})$ converges to $\underline{v}(\underline{x})$ strongly in $L_2(\Omega)$. In subsections 5.2–5.3 we show that $\underline{v}(\underline{x})$ is a solution of the minimization problem (4.10). Since this problem, due to nonnegativity of the tensor $\{a_{npqr}(\underline{x})\}$ and the condition $\lambda > 0$, has the unique solution; then, using the weak compactness of $\{\underline{v}_\varepsilon(\underline{x}, \lambda),\ \varepsilon > 0\}$, we conclude that $\underline{v}_\varepsilon \to v$ weakly in $H^1(\Omega)$ and strongly in $L_2(\Omega)$ as $\varepsilon \to 0$.

5.2. The upper bound. Denote by $\mathcal{J}^f_\varepsilon(\Omega)$ and $\mathcal{J}^f(\Omega)$ the spaces of admissible functions for the original and the homogenized variational problems (4.8)–(4.7) and (4.9)–(4.10), respectively. The goal of this subsection is to show that for any $\underline{w} \in \mathcal{J}^f(\Omega) \cap C^2$

$$\lim_{\varepsilon\to 0} \Phi_\varepsilon[\underline{v}_\varepsilon] \le \Phi_0[\underline{w}] \quad \forall\, \underline{w} \in \mathcal{J}^t(\Omega). \tag{5.5}$$

This is done by constructing a quasiminimizer $\underline{w}_{\varepsilon h} \in \mathcal{J}^f_\varepsilon(\Omega)$, $\underline{w}_{\varepsilon h} \to \underline{w}$ in $L_2(\Omega)$ as $\varepsilon \ll h \to 0$ so that

$$\Phi_\varepsilon[\underline{v}_\varepsilon] \le \Phi_\varepsilon[\underline{w}_{\varepsilon h}]. \tag{5.6}$$

Due to explicit construction of the quasiminimizer $\underline{w}_{\varepsilon h}$, it is possible to obtain the limiting inequality

$$\overline{\lim_{h\to 0}\lim_{\varepsilon\to 0}}\Phi_\varepsilon[\underline{w}_{\varepsilon h}] \le \Phi_0[\underline{w}]. \tag{5.7}$$

In particular, we will choose $\underline{w} = \underline{v}$, where $\underline{v}$ is the minimizer of the limiting functional Φ_0 and $\underline{v}_{\varepsilon_k} \rightharpoonup \underline{v}$. Note that (5.6) and (5.7) imply (5.5).

The key point in this step is the construction of the quasiminimizer $\underline{w}_{\varepsilon h}$. We now describe main ideas behind this construction. Choose a mesoscale parameter h such that $\varepsilon \ll h \ll 1$ and cover the domain Ω by a family of cubes K^α_h centered at the points $\underline{x}^\alpha$ of a cubic lattice with period $h - \delta$, $\delta = o(h)$, $h \to 0$, so that the cubes overlap. In each cube consider the mesocharacteristic (3.5). Minimizer $u^\alpha_{\varepsilon h}$ of the mesocharacteristic functional (see (3.9) with $\underline{\xi} = \underline{x}^\alpha$) is "close" (up to an additive constant) to the true minimizer $\underline{v}_\varepsilon$ (restricted to the cube K^α_h) if the tensor

$\underline{\underline{T}}$ in (3.5) is chosen so that $\underline{u}^\alpha_{\varepsilon h}$ is "close" to the linear symmetric part of the global homogenized minimizer $\underline{v}(\underline{x})$. In other words, we minimize the energy of the original problem in the perforated domain with an additional condition that the minimizer is smooth, i.e., close up to quadratic terms in h, to the linear function mentioned above. This condition is enforced by the penalty in term (3.5), and thus we obtain the quasiminimizer in each cube K^α_h. Now we need to "glue" them together, in order to obtain a globally smooth quasiminimizer. For this purpose we use a nested partition $K^\alpha_{h'} \subset K^\alpha_h$, $h' = h - 2\delta$, so that the cubes $K^\alpha_{h'}$ do not intersect each other. We retain the functions $\underline{u}^\alpha_{\varepsilon h}$ inside the cubes $K^\alpha_{h'}$ and modify them in the "thin" layers $K^\alpha_n \backslash K^\alpha_{h'}$ so that the obtained function $\underline{z}_{\varepsilon h}(\underline{x})$ is globally smooth (from $H^1(\Omega)$ and satisfies the rigid conditions on the particles Q^i_ε). However, $\operatorname{div} \underline{z}_{\varepsilon h}$ is no longer zero in the layers $K^\alpha_h \backslash K^\alpha_{h'}$, where the minimizers have been modified. To fix this, we introduce an additional function $\zeta_{\varepsilon h}(x)$ so that $\operatorname{div} \zeta_{\varepsilon h} = \operatorname{div} \underline{z}_{\varepsilon h}$ in the layers $K^\alpha_h \backslash K^\alpha_{h'}$, $\operatorname{div} \zeta_{\varepsilon h} = 0$ in $\cup_\alpha K^\alpha_{h'}$, and $\|\zeta_{\varepsilon h}\|_{H^1(\Omega)}$ is small so that its contribution to the energy of the global quasiminimizer is negligible, since $\operatorname{vol}(\cup_\alpha K^\alpha_h \backslash K^\alpha_{h'}) \to 0$ as $h \to 0$. Thus we obtain a quasiminimizer $\underline{w}_{\varepsilon h} = \underline{z}_{\varepsilon h} + \underline{\zeta}_{\varepsilon h} \in \mathcal{J}^f_\varepsilon(\Omega)$, and (5.6) holds. Here we use a special partition of unity, which preserves the rigid body conditions on the particles.

We now derive the inequality (5.6). Consider a cover of the domain Ω by cubes $K_\alpha = K(\underline{x}^\alpha, h)$ centered at the points $\underline{x}^\alpha$ and with the side length h, oriented along the coordinate axes. The centers $\underline{x}^\alpha$ form cubic lattice of the period $h - \delta$ ($0 < \varepsilon \le \delta \le h$). Let $K'_\alpha(\underline{x}^\alpha, h')$ be a concentric with K_α cube of side length $h' = h - 2\delta$; then $K'_\alpha = K_\alpha \backslash \cup_{\beta \ne \alpha} K_\beta$.

We will need three technical lemmas for constructing the function $\underline{w}_{\varepsilon h}(\underline{x}) \in \mathcal{J}^f_\varepsilon(\underline{x})$.

LEMMA 5.2 (special partition of unity). *Suppose the condition* (a_2) *holds. Then it is possible to construct the special partition of unity which corresponds to the cover* $\cup_\alpha K(\underline{x}^\alpha, h)$*, i.e., to construct the set of functions* $\{\varphi^\alpha_{\varepsilon h}(\underline{x}),\ \alpha = 1, 2, \dots\}$ *which satisfies the following conditions:*

1. $\varphi^\alpha_{\varepsilon h}(\underline{x}) \in C^2(\mathbb{R}^3)$,
2. $0 \le \varphi^\alpha_{\varepsilon h}(\underline{x}) \le 1$ *everywhere and*

$$\varphi^\alpha_{\varepsilon h}(\underline{x}) = \begin{cases} 1 \text{ inside } K'_\alpha, \\ 0 \text{ outside } K_\alpha, \end{cases}$$

3. $\sum_\alpha \varphi^\alpha_{\varepsilon h}(\underline{x}) = 1$ *in* $\mathbb{R}^3$,
4. $|\nabla \varphi^\alpha_{\varepsilon h}(\underline{x})| < \frac{C}{\delta}$ *in* $\mathbb{R}^3$,
5. $\varphi^\alpha_{\varepsilon h}(\underline{x}) = C^i_\varepsilon$ *when* $\underline{x} \in B(Q^i_\varepsilon)$,

where C^i_ε *are constants* ($0 \le C^i_\varepsilon \le 1$) *and* C *does not depend on* ε, δ, *or* h.

LEMMA 5.3. *Let the domain* $\Omega_\varepsilon = \Omega \backslash \cup_i Q^i_\varepsilon$ *satisfy conditions* (a_1)–(a_2). *Then for any function* $\underline{F}_\varepsilon(\underline{x}) \in L_2(\Omega)$ *which satisfies the conditions*

1. $\underline{F}_\varepsilon(\underline{x}) = 0$ *at* $x \in \cup_i B(Q^i_\varepsilon)$,
2. $\int_\Omega \underline{F}_\varepsilon(\underline{x}) dx = 0$,

there exists a function $\underline{\zeta}_\varepsilon(\underline{x}) \in H^1_0(\Omega)$ *such that* $\operatorname{div} \underline{\zeta}_\varepsilon(\underline{x}) = \underline{F}_\varepsilon$, $x \in \Omega$, $\underline{\zeta}_\varepsilon(\underline{x}) = \underline{\zeta}^i_\varepsilon$ *for* $\underline{x} \in B(Q^i_\varepsilon)$, *and* $\|\underline{\zeta}_\varepsilon\|_{H^1(\Omega)} \le C \|\underline{F}_\varepsilon\|_{L_2(\Omega)}$, *where* $\underline{\zeta}^i_\varepsilon$ *are constant vectors and* C *does not depend on* ε.

LEMMA 5.4. *Let the domain* $\Omega_\varepsilon = \Omega \backslash \cup_i Q^i_\varepsilon$ *satisfy conditions* (a_1), (a_2). *Then for any divergence-free function* $\underline{w}(\underline{x}) \in C^2(\Omega)$ *there exists such a function* $\underline{w}_\varepsilon(\underline{x}) \in H^2(\Omega)$ *that* $\operatorname{div} \underline{w}_\varepsilon = 0$ *in* Ω, $\underline{w}_\varepsilon(\underline{x}) = \underline{w}(\underline{x})$ *on* $\partial\Omega$, *and* $\underline{w}_\varepsilon(\underline{x})$ *is equal to the constant vectors* $\underline{w}^i_\varepsilon$ *on the balls* $B(Q^i_\varepsilon)$ ($\underline{w}^i_\varepsilon$ *is equal the mean value of* $\underline{w}(\underline{x})$ *on* $B(Q^i_\varepsilon)$). *In addition,*

the following inequalities are valid:

$$\|\underline{w}_\varepsilon - \underline{w}\|_{L_2(\Omega)} < C\varepsilon \quad \textit{and} \quad \|\underline{w}_\varepsilon\|_{H^1(G)} \le C\|\underline{w}\|_{H^1(G)}, \tag{5.8}$$

where the constants C do not depend on ε and $G \subset \Omega$ is any subdomain in Ω.

We now construct the function $\underline{w}_{\varepsilon h}(\underline{x})$. For any divergence-free in Ω function $\underline{w}(\underline{x}) \in C^2(\Omega)$, which is equal to $\underline{f}(\underline{x})$ on $\partial\Omega$, we wish to construct a function $\underline{w}_{\varepsilon h}(\underline{x}) \in \mathcal{J}^f_\varepsilon(\Omega)$, which is close to $\underline{w}(\underline{x}) \in \mathcal{J}^f(\underline{x})$ (in metric $L_2(\Omega)$) when ε and h are small enough ($\varepsilon \ll h \ll 1$).

A straightforward calculation shows that any function $\underline{w}(\underline{x}) \in C^2(K^\alpha)$ can be represented in the form

$$\begin{aligned} \underline{w}(\underline{x}) = \underline{w}(\underline{x}^\alpha) &+ \sum_{p,q=1} e_{pq}[\underline{w}(\underline{x}^\alpha)]\psi^{pq}(\underline{x} - \underline{x}^\alpha) \\ &+ \sum_{p,q} \omega_{pq}[\underline{w}(\underline{x}^\alpha)]\varphi^{pq}(\underline{x} - \underline{x}^\alpha) + \sigma^\alpha, \end{aligned} \tag{5.9}$$

where $\sigma^\alpha = O(|x - \underline{x}^\alpha|^2)$, and we introduce the notations for the symmetric and antisymmetric parts of $\nabla\underline{w}(\underline{x})$, respectively:

$$e_{pq}[\underline{w}(\underline{x}^\alpha)] = \frac{1}{2}\left[\frac{\partial w_p}{\partial x_q}(\underline{x}^\alpha) + \frac{\partial w_q}{\partial x_p}(\underline{x}^\alpha)\right], \tag{5.10}$$

$$\omega_{pq}[\underline{w}(\underline{x}^\alpha)] = \frac{1}{2}\left[\frac{\partial w_p}{\partial x_q}(\underline{x}^\alpha) - \frac{\partial w_q}{\partial x_p}(\underline{x}^\alpha)\right]. \tag{5.11}$$

Linear function $\underline{\psi}^{pq}(\underline{x})$ is defined by the equalities (3.8) and

$$\underline{\varphi}^{pq}(\underline{x}) = \frac{1}{2}[x_p e^q - x_q e^p], \tag{5.12}$$

and they both are divergence-free.

The first and the second sums in (5.9) are usually called the deformational and the rotational parts, respectively, of the velocity field $\underline{w}(\underline{x})$.

We need to construct a divergence-free function $\underline{w}_{\varepsilon h}(\underline{x})$, which satisfies the condition (2.12) on Q^i_ε and approximates $\underline{w}(\underline{x})$ as $\varepsilon \ll h \to 0$. Note that the second sum in the RHS of (5.9) satisfies (2.12), but the first sum does not, because the functions $\underline{\psi}^{pq}(\underline{x} - \underline{x}^\alpha)$ do not satisfy (2.12). However, the function $\underline{u}^{\alpha,pq}_{\varepsilon h}(\underline{x})$, which minimizes functional (3.5) in the class $\mathcal{J}_\varepsilon(K(\underline{x}^\alpha, h))$ when $T = \frac{1}{2}(e^p \otimes e^q + e^q \otimes e^p)$, and according to Lemma 5.4 (see below) is "close" to $\underline{\psi}^{pq}(\underline{x} - \underline{x}^\alpha)$ when $\varepsilon \ll h \ll 1$, does not satisfy (2.12). That is why, when constructing $\underline{w}_{\varepsilon h}(\underline{x})$ in cubes $K_\alpha = K(\underline{x}^\alpha, h) \subsetneq \Omega$, we replace $\underline{\psi}^{pq}(\underline{x} - \underline{x}^\alpha)$ by $\underline{u}^{\alpha,pq}_{\varepsilon h}(\underline{x})$ in the formula (5.9).

Further, we glue together the functions $\underline{w}_{\varepsilon h}(\underline{x})$ defined in each cube K_α by a special partition of unity $\{\varphi^\alpha_{\varepsilon h}(\underline{x})\}$ constructed in Lemma 5.2 which guaranties the rigid displacement conditions on Q^ε_i.

However, this construction does not ensure that $\underline{w}_{\varepsilon h}(\underline{x}) = \underline{w}(\underline{x}) = \underline{f}(\underline{x})$ on $\partial\Omega$. Also, since we used the partition of unity, the divergence-free condition is not guaranteed. Therefore, we use a simpler construction on the cubes K_α that intersect $\partial\Omega$. To ensure that the construction is divergence-free (without violation of all other requirements), we add the function $\underline{\zeta}_{\varepsilon h}(\underline{x})$, whose existence is established in Lemma 5.3.

Thus, we construct the quasiminimizer $\underline{w}_{\varepsilon h}(\underline{x})$ as follows:

$$\underline{w}_{\varepsilon h}(\underline{x}) = \sum{}' \left\{ \underline{w}(\underline{x}^\alpha) + \sum_{n,p=1}^{3} (e_{np}[\underline{w}(\underline{x}^\alpha)]\underline{u}_{\varepsilon h}^{\alpha,np}(\underline{x}) + \omega_{np}[\underline{w}(\underline{x}^\alpha)]\underline{\varphi}^{pq}(\underline{x}-\underline{x}^\alpha)) \right\} \varphi_{\varepsilon h}^{\alpha}(\underline{x})$$

$$+ \sum{}'' \underline{w}_\varepsilon(\underline{x})\varphi_{\varepsilon h}^{\alpha}(\underline{x}) + \underline{\zeta}_{\varepsilon h}(\underline{x}) \equiv \underline{z}_{\varepsilon h}(\underline{x}) + \underline{\zeta}_{\varepsilon h}(\underline{x}), \tag{5.13}$$

where the sum $\sum'_\alpha$ is taken over the interior cubes (which lie inside the domain Ω), sum $\sum''_\alpha$ is taken over the cubes,which intersect the boundary $\partial\Omega$, and in the partition of unity $\varphi_{\varepsilon h}^{\alpha}(\underline{x})$ we choose $\delta = h^{1+\tau/2}$, $0 < \tau < 2$. The function $\underline{w}_\varepsilon(\underline{x}) \in H^1(\Omega)$ is constructed according to Lemma 5.4.

We now construct the function $\underline{\zeta}_{\varepsilon h}(\underline{x})$. Due to (5.13), the function $\underline{z}_{\varepsilon h}(\underline{x}) \in H^1(\Omega)$ is equal $\underline{w}(\underline{x})$ on the boundary $\partial\Omega$, and since $\int_\Omega \langle \underline{w}, \underline{\nu}\rangle dS = 0$, we have $\int_\Omega \operatorname{div} \underline{z}_{\varepsilon h} dx = 0$.

Moreover, using the properties of the functions $\underline{u}_{\varepsilon h}^{\alpha,np}(\underline{x})$, $\omega_{np}[\underline{w}(\underline{x}^\alpha)]\underline{\varphi}^{pq}(\underline{x}-\underline{x}^\alpha)$, $\underline{w}_\varepsilon(\underline{x})$, and the functions $\varphi_{\varepsilon h}^{\alpha}(\underline{x})$, we obtain

$$\operatorname{div} z_\varepsilon(\underline{x}) = 0 \quad \text{if} \quad \underline{x} \in B(Q_\varepsilon^i).$$

Therefore, applying Lemma 5.3 to the function $\underline{F}_\varepsilon(\underline{x}) = -\operatorname{div} \underline{z}_{\varepsilon h}$, we construct the divergence-free function $\underline{\zeta}_{\varepsilon h}(\underline{x})$, which is equal to the constant vectors $\underline{\zeta}_{\varepsilon h}^i$ on the balls $B(Q_\varepsilon^i)$ and zero on $\partial\Omega$. Thus the construction (5.13) guarantees that $\underline{w}_{\varepsilon h}(\underline{x}) \in \mathcal{J}_\varepsilon^f(\Omega)$.

Let us calculate the functional (4.8) on the function $\underline{w}_{\varepsilon h}(\underline{x}) \in \mathcal{J}_\varepsilon^f(\Omega)$ ($\varepsilon \ll \delta = h^{1+\frac{\tau}{2}} \ll h \ll 1$).

The next lemma shows that the contribution from the function $\underline{\zeta}_{\varepsilon h}(\underline{x})$ is negligible.

LEMMA 5.5. *The following equalities hold:*

$$\lim_{h\to 0} \overline{\lim}_{\varepsilon\to 0} \|\underline{\zeta}_{\varepsilon h}\|_{H^1(\Omega)} = I_\Omega[\underline{\zeta}_{\varepsilon h}, \underline{\zeta}_{\varepsilon h}] = 0.$$

To calculate $e_{k\ell}[\underline{w}_{\varepsilon h}]$ were rewrite (5.13) as follows:

$$w_{\varepsilon h}(\underline{x}) = w(\underline{x}) + \sum_\alpha{}' \left\{ \sum_{np=1}^{3} e_{np}[\underline{w}(\underline{x}^\alpha)](u_{\varepsilon h}^{\alpha,np}(\underline{x}) - \underline{\psi}^{\alpha,np}(\underline{x})) - \underline{\sigma}^\alpha(\underline{x}) \right\} \varphi_{\varepsilon h}^{\alpha}(\underline{x})$$

$$+ \sum_\alpha{}'' (w_\varepsilon(\underline{x}) - w(\underline{x}))\varphi_{\varepsilon h}^{\alpha}(\underline{x}) + \underline{\zeta}_{\varepsilon h}(\underline{x}). \tag{5.14}$$

Using Lemma 5.5, we distinguish the leading term in $e_{k\ell}[\underline{w}_{\varepsilon h}(\underline{x})]$,

$$e_{k\ell}[\underline{w}_{\varepsilon h}(\underline{x})] = \sum_\alpha{}' \sum_{n,p=1}^{3} e_{np}[\underline{w}(\underline{x}^\alpha)]e_{k\ell}[u_{\varepsilon h}^{\alpha,np}]\varphi_{\varepsilon h}^{\alpha}(\underline{x}) + L(\varepsilon,h), \tag{5.15}$$

where $\lim_{h\to 0} \overline{\lim}_{\varepsilon\to 0} \|L(\varepsilon,h)\|_{L_2(\Omega)} = 0$, and compute the bulk and the interaction energies:

(5.16)

$$E_\Omega[\underline{w}_{\varepsilon h}, \underline{w}_{\varepsilon h}] = \sum_\alpha{}' \sum_{n,p,q,r=1}^{3} e_{np}[\underline{w}(\underline{x}^\alpha)]e_{qr}[\underline{w}(\underline{x}^\alpha)]E_{K'_\alpha}[\underline{u}_{\varepsilon h}^{\alpha,pn}, \underline{u}_{\varepsilon h}^{\alpha,qr}] + L_1(\varepsilon,h),$$

$$I_\Omega^\varepsilon[\underline{w}_{\varepsilon h}, \underline{w}_{\varepsilon h}] \le \sum_\alpha{}' \sum_{n,p,q,r} e_{np}[\underline{w}(\underline{x}^\alpha)]e_{qr}[\underline{w}(\underline{x}^\alpha)] \cdot I_{K'_\alpha}^\varepsilon[\underline{u}_{\varepsilon h}^{\alpha,np}, \underline{u}_{\varepsilon h}^{\alpha,qr}] + L_2(\varepsilon,h). \tag{5.17}$$

Here $\lim_{h\to 0} \overline{\lim}_{\varepsilon\to 0} L_i(\varepsilon,h) = 0$, $i = 1,2$, and we have used the notation (3.6).

Combining (5.16) and (5.17), we estimate the bulk and the interaction energy in $\Phi_\varepsilon[\underline{w}_{\varepsilon h}]$:

$$
\begin{aligned}
&E_\Omega[\underline{w}_{\varepsilon h},\underline{w}_{\varepsilon h}]+\frac{1}{\lambda}I^\varepsilon_\Omega[\underline{w}_{\varepsilon h},\underline{w}_{\varepsilon h}]\\
&\le \sum_\alpha{}' \sum_{n,p,q,r} e_{np}[\underline{w}(\underline{x}^\alpha)]e_{qr}[\underline{w}(\underline{x}^\alpha)]\Bigg\{2\mu\int_{K_\alpha}\sum_{k,\ell}e_{k\ell}[(\underline{u}^{\alpha,np}_{\varepsilon h}(\underline{x})]e_{n\ell}[\underline{u}^{\alpha,qr}_{\varepsilon h}(\underline{x})]dx\\
&\qquad\qquad +\frac{1}{\lambda}I^\varepsilon_{K_\alpha}[\underline{u}^{\alpha,np}_{\varepsilon h},\underline{u}^{\alpha,qr}_{\varepsilon h}]\Bigg\}\\
&\le \sum_\alpha{}' \sum_{n,p,q,r} e_{np}[\underline{w}(\underline{x}^\alpha)]e_{qr}[\underline{w}(\underline{x}^\alpha)]\Bigg\{2\mu\int_{K_\alpha}\Bigg[\sum_{k,\ell}e_{k\ell}[\underline{u}^{\alpha,np}_{\varepsilon h}(\underline{x})]e_{k\ell}[\underline{u}^{\alpha,qr}_{\varepsilon h}(\underline{x})]\\
&\quad +h^{-2-\tau}\langle\underline{u}^{\alpha,np}_{\varepsilon h}(\underline{x})-\underline{\psi}^{\alpha,np}(\underline{x}),\underline{u}^{\alpha qr}_{\varepsilon h}(\underline{x})-\underline{\psi}^{\alpha,qr}(\underline{x})\rangle\Bigg]dx+\frac{1}{\lambda}I^\varepsilon_{K_\alpha}[\underline{u}^{\alpha np}_{\varepsilon h},\underline{u}^{\alpha qr}_{\varepsilon h}]\Bigg\}+o(1)\\
&= \sum_\alpha{}' \sum_{n,p,q,r} a_{npqr}(\underline{x}^\alpha,\lambda,\varepsilon,h)e_{np}[\underline{w}(\underline{x}^\alpha)]e_{qr}[\underline{w}(\underline{x}^\alpha)]+o(1)\quad(\varepsilon\ll h\ll 1).
\end{aligned}
\tag{5.18}
$$

Here we use the definition of $a_{npqr}(\underline{x},\lambda,\varepsilon,h)$ (see Lemma 3.1) and add to the RHS of the first inequality in (5.18) a positive term which corresponds to the penalty term in the mesocharacteristic.

Finally we use (5.18) to estimate the functional (4.8):

$$
\begin{aligned}
\Phi_\varepsilon[\underline{w}_{\varepsilon h}]\le{}&\sum_\alpha{}' h^3\sum_{n,p,q,r}\frac{a_{npqr}(\underline{x}^\alpha,\lambda,\varepsilon,h)}{h^3}\,e_{np}[\underline{w}(\underline{x}^\alpha)]e_{qr}[\underline{w}(\underline{x}^\alpha)]\\
&+\lambda\int_\Omega\langle\rho_\varepsilon w_{\varepsilon h},w_{\varepsilon h}\rangle dx-2\int_\Omega\langle\rho_\varepsilon v_{\varepsilon 0},w_{\varepsilon h}\rangle dx+\Delta(\varepsilon,h),
\end{aligned}
\tag{5.19}
$$

where the remainder $\Delta(\varepsilon,h)$ satisfies the equality

$$
\lim_{h\to 0}\overline{\lim}_{\varepsilon\to 0}\Delta(\varepsilon,h)=0.
$$

We now pass to the limit in (5.19) first as $\varepsilon\to 0$ and then as $h\to 0$. Taking into account that $\underline{w}(\underline{x})\in C^2(\overline{\Omega})$ and using the conditions (b_1), (b_2), (3.17), and $\lim_{h\to 0}\overline{\lim}_{\varepsilon\to 0}\|w_{\varepsilon h}(\underline{x})-\underline{w}(\underline{x})\|^2_{L_2(\Omega)}=0$ (verified by a direct calculation), we obtain

$$
\lim_{h\to 0}\overline{\lim}_{\varepsilon\to 0}\Phi_\varepsilon[\underline{w}_{\varepsilon h}]\le\Phi_0[\underline{w}]\,,
$$

where the functional $\Phi_0[\underline{w}]$ is defined by the equality (4.9). Since $\underline{w}_{\varepsilon h}\in\mathcal{J}^f_\varepsilon(\Omega)$ and $\underline{v}_\varepsilon(\underline{x})$ is the solution of the minimization problem (4.7), inequality (5.6) is established for any function $\underline{w}(\underline{x})\in\mathcal{J}^f(\Omega)\cap C^2(\Omega)$. Since the class $\mathcal{J}^f(\Omega)\cap C^2(\Omega)$ is dense in $\mathcal{J}^f(\Omega)$ in the metric $H^1(\Omega)$, this inequality holds for any function $\underline{w}\in\mathcal{J}^f(\Omega)$.

5.3. The lower bound. We begin with a short overview and then present actual technical constructions which are quite lengthy. The goal of this step is to establish the following lower bound:

$$
\underline{\lim}_{\varepsilon=\varepsilon_k\to 0}\Phi_\varepsilon[\underline{v}_\varepsilon]\ge\Phi_0[\underline{v}],
\tag{5.20}
$$

where $\underline{v}$ is the weak limit of $\underline{v}_\varepsilon$ as $\varepsilon = \varepsilon_k \to 0$. To this end we partition the domain Ω into cubes K_h^α (as opposed to the covering by intersecting cubes in subsection 5.2). Then the energy of the global minimizer $\underline{v}_\varepsilon$ in each cube K_h^α is bounded below by the mesocharacteristic (3.5) with the proper choice of the tensor $\underline{\underline{T}}$ (from the linear symmetric part of the limiting function $\underline{v}(x)$).

More precisely, we modify the function $\underline{v}_\varepsilon(\underline{x})$ in K_h^α so that the modified function $\underline{u}_\varepsilon^\alpha$ satisfies the following conditions:

(i) $\underline{u}_\varepsilon^\alpha \in \mathcal{J}_\varepsilon^f(K_h^\alpha)$ ($\underline{u}_\varepsilon^\alpha$ is an admissible function for the variational problem (3.5)).

(ii) $\left|A_{\varepsilon h}[\underline{u}_\varepsilon^\alpha, \underline{\underline{T}}, \underline{x}^\alpha, \tau] - (\text{total energy of } \underline{v}_\varepsilon^\alpha \text{ in } K_h^\alpha)\right| = o(h^3)$, $h \to 0$.

(iii) $\underline{u}_\varepsilon^\alpha$ is closed to the symmetric part (5.9) of $\nabla \underline{v}$ defined by (5.9) when $\underline{w} = \underline{v}$.

Then (i) and (3.5)–(3.10) imply

$$A_{\varepsilon h}[\underline{u}_\varepsilon^\alpha, \underline{\underline{T}}, \underline{x}^\alpha, \tau] \geq \sum_{npqr} a_{npqr}(\underline{x}^\alpha, \lambda, \varepsilon, h, \tau) T_{np}\, T_{qn}. \tag{5.21}$$

We now describe the modification of $\underline{u}_\varepsilon^\alpha$ and the corresponding choice of the tensor $\underline{\underline{T}}$ in the penalty term of (3.5).

Represent the limiting function $\underline{v}(\underline{x})$ in the form ($x \in K_h^\alpha$)

$$\begin{aligned} \underline{v}(\underline{x}) &= \underline{v}(\underline{x}^\alpha) + \sum_{k\ell} e_{k\ell}[\underline{v}(\underline{x}^\alpha)]\underline{\psi}^{k\ell} \\ &+ \sum_{k\ell} \omega_{k\ell}(v(\underline{x}^\alpha))\underline{\varphi}^{k\ell} + O(h^2) := \underline{v}(\underline{x}^\alpha) + S + A + O(h^2), \end{aligned} \tag{5.22}$$

where the linear part in (5.22) in decomposed into symmetric and antisymmetric parts and both are divergence-free (see the definitions of $\psi^{k\ell}$ in (3.8) and $\phi^{k\ell}$ in Lemma 5.2). Note that in the standard decomposition $\nabla \underline{u} = \overline{\frac{1}{2}(\nabla \underline{u} + \nabla \underline{u}^T)} + \frac{1}{2}(\nabla \underline{u} - \nabla \underline{u}^T)$ the symmetric part is not necessarily divergence-free but $\operatorname{div} \underline{v} = 0$ in (5.22) (weakly). We now describe the main idea of the construction of $\underline{u}_\varepsilon^\alpha$. Choose $\underline{u}_\varepsilon^\alpha$ in the form

$$\underline{u}_\varepsilon^\alpha = \underline{v}_\varepsilon(\underline{x}) - \underline{v}(\underline{x}^\alpha) - A\,, \tag{5.23}$$

where A is defined in (5.22). Then the RHS of (5.23) is close to the symmetric part S of the $\nabla \underline{v}(\underline{x})$. This is sufficient to make the penalty term in (3.5) small. Indeed, only the symmetric part S of the deformation enters the penalty term in the mesocharacteristic (3.5) since the antisymmetric part A (the rotational part) does not contribute to the energy. Due to (5.23), $e_{k\ell}[\underline{u}_\varepsilon^\alpha] = e_{k\ell}[\underline{v}_\varepsilon]$ in K^α and $I^\varepsilon_{K^\alpha}[\underline{v}_\varepsilon, \underline{v}_\varepsilon] = I^\varepsilon_{K^\alpha}[\underline{u}_\varepsilon^\alpha, \underline{u}_\varepsilon^\alpha]$. We choose $T_{np} = e_{np}[\underline{v}(x^\alpha)]$. Then the penalty term in the LHS of (5.21) becomes small, the interaction term in the LHS of (5.21) is simplified due to the condition (2.3), and the inequality (5.21) follows. Summing up over all cubes in the partition and passing to the limit $\varepsilon \ll h \to 0$ yields (5.20).

We now present the detailed derivation of the lower bound (5.20). We prove (5.20) when the function $\underline{v}(\underline{x})$ is the weak limit in $H^1(\Omega)$ of solutions $\underline{v}_\varepsilon(\underline{x})$ of the problem (4.7) in some subsequence $\{\varepsilon_k \to 0,\ k = 1, 2, \dots\}$. First, we assume for simplicity that the limiting function $\underline{v}(\underline{x})$ is smooth enough, namely, $\underline{v}(\underline{x}) \in C^2(\Omega) \cap \mathcal{J}^f(\Omega)$.

Then we partition the domain Ω by the nonintersecting cubes $K_\alpha = K(\underline{x}^\alpha, h)$ centered at the points $\underline{x}^\alpha$ aligned along the coordinate axes. In each internal with respect to the Ω cube (which does not intersect $\partial\Omega$) consider a function

$$\underline{u}_\varepsilon^\alpha(\underline{x}) = \underline{v}_\varepsilon(\underline{x}) - \left\{ \underline{v}(\underline{x}^\alpha) + \sum_{p,q} \omega_{pq}[\underline{w}(\underline{x}^\alpha)]\varphi^{pq}(\underline{x} - \underline{x}^\alpha) \right\}, \tag{5.24}$$

where we have used the notation (5.11) for the rigid rotation part.

Clearly $\underline{u}_\varepsilon(\underline{x}) \in \mathcal{J}_\varepsilon(K_\alpha)$ (admissible for the functional (3.9)). Therefore, (3.10) implies (5.21) for any tensor $T = \{T_{np}\}_{n,p=1}^3$. Set $T_{np} = e_{np}[\underline{v}(\underline{x}^\alpha)]$. Then, taking into account the form (3.5) of the functional $A_{\varepsilon h}$, the form (5.24) of the function $\underline{u}_\varepsilon^\alpha(\underline{x})$, and the condition (2.3), we obtain

$$\int_{K_\alpha} \left\{ 2\mu \sum_{k,\ell} e_{k\ell}^2[\underline{v}_\varepsilon] + h^{-2-\tau} \left| \underline{u}_\varepsilon^\alpha(\underline{x}) - \sum_{n,p=1}^3 e_{np}[v(\underline{x}^\alpha)] \underline{\psi}^{\alpha,np}(\underline{x}) \right|^2 \right\} dx$$

$$+ \frac{1}{2\lambda} I_{K_\alpha^\varepsilon}^\varepsilon[\underline{v}_\varepsilon, \underline{v}_\varepsilon] \geq \sum_{n,p,q,r} a_{npqr}(\underline{x}^\alpha, \lambda; \varepsilon, h, \tau) e_{np}[v(\underline{x}^\alpha)] e_{qr}[v(\underline{x}^\alpha)] \,. \tag{5.25}$$

Next we note that the penalty term is small, namely (5.24), and the convergence of $\underline{v}_{\varepsilon_k} \to \underline{v}$ implies

$$\lim_{\varepsilon = \varepsilon_k \to 0} \int_{K_\alpha} \left| \underline{u}_\varepsilon^\alpha(\underline{x}) - \sum_{n,p=1}^3 e_{np}[v(\underline{x}^\alpha)] \underline{\psi}^{\alpha,np}(\underline{x}) \right|^2 dx = O(h^7) \,. \tag{5.26}$$

Now we sum over all cubes and take into account (4.8), (5.25)–(5.26) to conclude that if $\varepsilon_k \leq \widehat{\epsilon}(h)$, then

$$\Phi_\varepsilon[\underline{v}_\varepsilon] \geq {\sum_\alpha}' \sum_{n,p,q,r=1}^3 a_{npqr}(\underline{x}^\alpha, \lambda; \varepsilon, h, \tau) e_{np}[v(\underline{x}^\alpha)] e_{qr}[v(\underline{x}^\alpha)]$$

$$+ \int_\Omega \{ \lambda \langle \underline{v}_\varepsilon, \underline{v}_\varepsilon \rangle - 2 \langle \rho_\varepsilon \underline{v}_{\varepsilon 0}, \underline{v}_\varepsilon \rangle \} \, dx + O(h^{2-\tau}) \,. \tag{5.27}$$

Note that, according to our assumption, $\tau < 2$. Then we pass to the limit in (5.27) first in $\varepsilon_k \to 0$ and then in $h \to 0$. Taking into account that $\underline{v}(\underline{x}) \in C^2(\Omega)$ and using the conditions (b_1), (b_2), (3.17), and the convergence $\underline{v}_\varepsilon(\underline{x}) \to \underline{v}(\underline{x})$ in $L_2(\Omega)$ as $\varepsilon = \varepsilon_k \to 0$, we obtain

$$\varliminf_{\varepsilon = \varepsilon_k \to 0} \Phi_\varepsilon[\underline{v}_\varepsilon] \geq \int_\Omega \left\{ \sum_{n,p,q,r} a_{npqr}(\underline{x}, \lambda) e_{np}[\underline{v}(\underline{x})] e_{qr}[\underline{v}(\underline{x})] \right.$$

$$\left. + \lambda \langle \underline{v}(\underline{x}), \underline{v}(\underline{x}) \rangle - 2 \langle \rho(\underline{x}) \underline{v}_0(\underline{x}), \underline{v}(\underline{x}) \rangle \right\} dx = \Phi_0[\underline{v}]. \tag{5.28}$$

Thus, we have obtained the required inequality (5.20) under the assumption that $\underline{v} \in C^1(\overline{\Omega})$. The proof for a nonsmooth case $\underline{v} \in H^1$ is more technical, although its scheme is the same: it is necessary to construct smooth approximations $\underline{v} \approx \underline{v}^\sigma$ and $\underline{v}_\varepsilon \approx \underline{v}_\varepsilon^\sigma$ for some small $\sigma > 0$. The above scheme is applied to the approximations when passing to the limit first as $\varepsilon \ll h \to 0$ and then as $\sigma \to 0$. The details of this construction are presented in [1].

6. Analyticity in the spectral parameter of the mesocharacteristic and the solutions. In Theorem 4.2 we established the convergence of the solutions $\underline{v}_\varepsilon(\underline{x}, \lambda)$ to the homogenized solution $\underline{v}(\underline{x}, \lambda)$ as $\varepsilon \to \infty$ for real $\lambda > 0$. To prove the main Theorem 3.3 we need to apply the inverse Laplace transform to get the convergence of $\underline{v}_\varepsilon(\underline{x}, t)$ to $\underline{v}(\underline{x}, \lambda)$. Thus we need to show that $\underline{v}_\varepsilon(\underline{x}, \lambda)$ can be analytically extended into the complex plane (more precisely, into $\mathbf{C} \setminus \mathbf{R}^-$).

The extension procedure is quite technical, and here we only outline it. The details are presented in [1]. First we formulate the problem (2.10)–(2.14) in an abstract operator form in an appropriate Hilbert space. For a cell problem, the analogous abstract form is obtained below in Lemmas 7.5 and 7.6. The operator formulation of the problem (2.10)–(2.14) leads to an operator bundle of the form $(\mathbf{I}+\lambda A_{1\varepsilon}+\frac{1}{\lambda}A_{2\varepsilon}\underline{v}_\varepsilon = f_\varepsilon)$, where $\mathbf{I}$ is the unity operator and $A_{1\varepsilon}$ and $A_{2\varepsilon}$ are compact operators. Using well-known results from the operator bundle theory [9, Chapter 7], we show that $\underline{v}_\varepsilon(\underline{x},\lambda)$ is analytic in $\mathbf{C}\setminus\mathbf{R}^-$ and

$$\|\underline{v}_\varepsilon\|^2_{L_2(\Omega)} < \frac{C}{|\lambda|}. \tag{6.1}$$

Using analogous abstract operator theory considerations and the Vitaly theorem [11], we prove Lemma 3.2. Furthermore, we establish a bound analogous to (6.1) and show analyticity in λ of the homogenized solution $\underline{v}(\underline{x},\lambda)$ in $\mathbf{C}\setminus\mathbf{R}^-$. The Vitaly theorem yields that the convergence results from Theorem 4.2 hold for complex λ in $\mathbf{C}\setminus\mathbf{R}^-$. Finally, we apply the inverse Laplace transform (with an appropriate choice of the integration contour specified in [1]) to show that Theorem 4.2 implies Theorem 3.3.

7. Periodic structures. We now present a case when all the conditions (a_1)–(a_4), (b_1), (b_2) in Theorem 3.3 are satisfied and the Laplace transform $\{a_{npqr}(\underline{x},\lambda)\}$ of the effective tensor $\{a_{npqr}(\underline{x},t)\}$ does not depend on x and can be computed by solving the so-called cell problems.

We consider a periodic array when particles Q^i_ε are obtained by a homothetic compression by the factor ε of a fixed body Q with diameter $d \le 1/2$ and center of mass at the origin. All particles are identically oriented in space, and their centers of mass x^i_ε form a cubic lattice L_ε with period ε.

We are interested in the case of short-range interactions when each particle interacts only with neighboring 3^n-1, $n=2,3$, particles (a schematic image for $n=2$ is given in Figure 7.1) and the corresponding interaction matrix $\underline{\underline{C}}^{ij}_{k\varepsilon}$ does not change under translations of the lattice L_ε, i.e., $\underline{\underline{C}}^{ij}_{k\varepsilon} = \underline{\underline{C}}^{i+l,j+l}_{k\varepsilon}$. We also assume that $\underline{\underline{C}}^{ij}_{1\varepsilon} = \varepsilon\underline{\underline{C}}^{ij}_1$, $\underline{\underline{C}}^{ij}_{2\varepsilon} = \varepsilon^2\underline{\underline{C}}^{ij}_2$, $\underline{\underline{C}}^{ij}_{3\varepsilon} = \varepsilon^3\underline{\underline{C}}^{ij}_3$, and $\underline{\underline{C}}^{ij}_{4\varepsilon} = \varepsilon^3\underline{\underline{C}}^{ij}_4$. Thus both the geometry and the interactions are ε-periodic.

Consider a particle Q^i_ε placed inside a cube K^i_ε of side length ε so that both the particle and the cube are centered at the point $\underline{x}^i_\varepsilon$. Then $D^i_\varepsilon = K^i_\varepsilon \setminus Q^i_\varepsilon$ (see Figure 7.1) is a periodicity cell filled with the fluid. To obtain the standard unit cell we rescale D^i_ε by the factor ε^{-1} and shift its center to the origin. Then the domain $D = K\setminus Q$ is a unit periodicity cell, where K is the cube of side length 1 centered at the origin and Q is a domain in K with diameter $d<1/2$ and boundary $\partial Q \in C^2$.

Let us consider the boundary value problem in $D = K\setminus Q$, the so-called cell problem:

$$-\mu\Delta\underline{u}^{qr} + \nabla p^{qr} = 0, \qquad \underline{x}\in K\setminus Q, \tag{7.1}$$

$$\operatorname{div}\underline{u}^{qr} = 0, \qquad \underline{x}\in K\setminus Q, \tag{7.2}$$

$$\underline{u}^{qr} = -\underline{\psi}^{qr}(\underline{x}) + \underline{b}^{qr}\times\underline{x}, \quad x\in\partial Q, \tag{7.3}$$

$$\int_{\partial Q}\underline{x}\times(\underline{\underline{\sigma}}[\underline{u}^{qr}]\cdot\underline{\nu})dS = -\frac{1}{\lambda}\underline{\underline{C}}\,\underline{b}^{qr}, \tag{7.4}$$

$$\underline{u}^{qr}(\underline{x})\big|_{\Gamma_i^-} = \underline{u}^{qr}(\underline{x})\big|_{\Gamma_i^+}, \quad \underline{\underline{\sigma}}[\underline{u}^{qr}(\underline{x})]\big|_{\Gamma_i^-} = \underline{\underline{\sigma}}[\underline{u}^{qr}(\underline{x})]\big|_{\Gamma_i^+}. \tag{7.5}$$

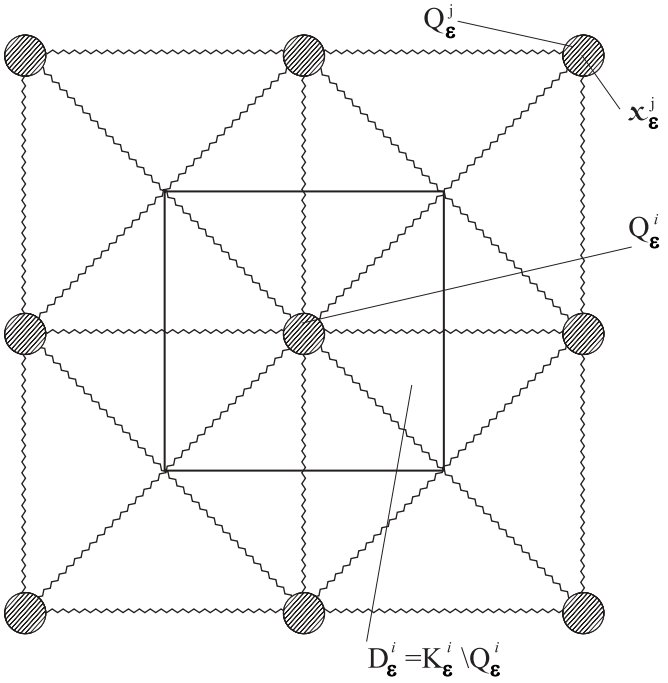


FIG. 7.1. *Two-dimensional periodic structure.*

Here $\underline{\underline{\sigma}}[\underline{u}]$ is the stress tensor defined by (2.15), $\underline{\psi}^{qr}(\underline{x})$ are functions defined by (3.8), $\underline{\nu}$ is the unit outward normal to ∂D, $\underline{b}^{qr}$ are arbitrary constant vectors, and $\underline{\underline{C}}$ is the following matrix:

$$\underline{\underline{C}} = \sum_j{}' \frac{\underline{\underline{C}}_{3\varepsilon}^{ij} - \underline{\underline{C}}_{4\varepsilon}^{ij}}{\varepsilon^3} = \sum_j{}' \underline{\underline{C}}_3^{ij} - \underline{\underline{C}}_4^{ij}, \tag{7.4$'$}$$

where $C_{3\varepsilon}^{ij}$, $C_{4\varepsilon}^{ij}$ are interaction matrices defined in (2.8), the sum $\sum_j'$ is taken over all neighbors of Q_ε^i, and due to periodicity $\underline{\underline{C}}$ does not depend on i. Γ_i^- and Γ_i^+ are opposite faces of the cube K ($i = 1, 2, 3$).

LEMMA 7.1. *The problem* (7.1)–(7.5) *for any* $\lambda > 0$ *has the unique solution (up to an additive constant in the pressure* p^{qr}*)* $\{\underline{u}^{qr}(\underline{x}), \underline{b}^{qr}, p^{qr}(\underline{x})\}$*, such that* $\underline{u}^{qr}(\underline{x}) \in C^{2+\alpha}\overline{(K \setminus Q)}$*,* $p^{qr} \in C^{1+\alpha}\overline{(K \setminus Q)}$*,* $\alpha > 0$*, and it admits a periodic extension on* $\mathbb{R}^3$*.*

The proof of this lemma uses standard variational techniques and is presented in [1].

The following theorem allows us to compute the effective properties of the suspension.

THEOREM 7.2. *For the periodic structure described above there exist the limits in the condition* (b_1)*, the components of limiting tensor* $\{a_{npqr}(\underline{x}, \lambda)\}$ *are constants with respect to* $\underline{x}$*, and they can be calculated as follows:*

$$a_{npqr}(\lambda) = \mu \mathbf{I}_{npqr} + \frac{1}{\lambda} \sum_j{}' \langle \underline{\underline{C}}_1^{ij} \underline{\psi}^{np}(\underline{x}^j), \underline{\psi}^{qr}(\underline{x}^j) \rangle$$

$$+ 2\mu \int_K \sum_{k,l=1}^{3} e_{kl}[\underline{u}^{np}(\underline{x}, \lambda)] e_{kl}[\underline{u}^{qr}(\underline{x}, \lambda)] dx + \frac{1}{\lambda} \langle \underline{\underline{C}} \underline{b}^{np}, \underline{b}^{qr} \rangle. \tag{7.6}$$

Here $\mathbf{I}_{npqr}$ *are components of the fourth rank tensor defined by*

$$\mathbf{I}_{npqr} = \frac{1}{2}(\delta_{nq}\delta_{pr} + \delta_{nr}\delta_{pq}) - \frac{1}{3}\delta_{np}\delta_{qr}. \tag{7.7}$$

The constant matrices $\underline{\underline{C}}_1^{ij} = \frac{1}{2\varepsilon}\underline{\underline{C}}_{1\varepsilon}^{ij}$ *and* $\underline{\underline{C}}$ *(see* (7.4′) *and* (2.8)*) do not depend on* ε*;* $\underline{x}^j$ *are the centers of mass of the neighbors to* Q^i *particles, i.e.,* $\underline{x}^j = \sum_{k=1}^3 n_k^j e^k$*,* $n_k^j = 0, \pm 1$*; and* $\underline{u}^{np}(\underline{x}, \lambda)$ *is the solution of the cell problem* (7.1)–(7.5) *extended onto* Q *by* (7.3).

Remark 7.1. Clearly in the periodic case the limits in condition (b_2) exist, and the limiting density ρ is the following constant:

$$\rho = \rho_s|Q| + \rho_f(1 - |Q|),$$

where $|Q|$ is the volume of the particle Q.

Outline of the proof of Theorem 7.2. Let K_h^ξ be a cube of the side length $h \gg \varepsilon$ centered at the point $\xi \in \Omega$. Consider a function in K_h^ξ

$$\underline{U}_\varepsilon^{qr}(\underline{x}) = \underline{\psi}^{qr}(\underline{x} - \underline{\xi}_\varepsilon) + \varepsilon \underline{\widetilde{u}}^{qr}\left(\frac{\underline{x} - \underline{\xi}_\varepsilon}{\varepsilon}, \lambda\right), \tag{7.8}$$

where $\underline{\widetilde{u}}^{qr}(\underline{x}, \lambda)$ is a periodic extension of the solution $\underline{u}^{qr}(\underline{x}, \lambda)$ of the cell problem (7.1)–(7.5) (see Lemma 7.1); $\underline{\xi}_\varepsilon = \underline{x}_\varepsilon^i$ is the nearest to $\underline{\xi}$ center of mass of particles Q_ε^i. Using the properties of the functions $\underline{\psi}^{qr}(\underline{x})$ and $\underline{u}^{qr}(\underline{x}, \lambda)$, we have

$$\operatorname{div} \underline{\widetilde{u}}_\varepsilon^{qr}(\underline{x}, \lambda) = 0$$

in $K_{h\varepsilon}^\xi = K_h^\xi \setminus \cup_i Q_\varepsilon^i$. On the boundaries of the particles Q_ε^j,

$$\underline{U}_\varepsilon^{qr}(\underline{x}, \lambda) = \underline{\psi}^{qr}(\underline{x}_\varepsilon^j - \underline{\xi}_\varepsilon) + \underline{b}(\lambda)^{qr} \times (\underline{x} - \underline{x}_\varepsilon^j), \quad \underline{x} \in S_\varepsilon^j, \tag{7.9}$$

where $\underline{b}(\lambda)^{qr}$ is a constant vector from (7.3). We extend $\underline{U}_\varepsilon^{qr}(\underline{x}, \lambda)$ on the particles Q_ε^j, using (7.9). Then we have

$$\operatorname{div} \underline{U}_\varepsilon^{qr}(\underline{x}) = 0 \quad \text{in } K_h^\xi. \tag{7.10}$$

We seek a function $\underline{\hat{u}}_\varepsilon^{qr}(\underline{x}, \lambda)$ that minimizes the functional (3.5) for $T = \frac{1}{2}(\underline{e}^q \otimes \underline{e}^r + \underline{e}^r \otimes \underline{e}^q)$ in the form

$$\underline{\hat{u}}_\varepsilon^{qr}(\underline{x}, \lambda) = \underline{U}_\varepsilon^{qr}(\underline{x}, \lambda) + \underline{v}_\varepsilon^{qr}(\underline{x}, \lambda), \tag{7.8′}$$

where $\underline{U}_\varepsilon^{qr}(\underline{x}, \lambda)$ is defined by (7.8). Next we obtain a variational problem for the corrector $\underline{v}_\varepsilon^{qr}(\underline{x}, \lambda)$. Analysis of this problem shows that $E_{K_h^\xi}[\underline{v}_\varepsilon^{qr}(\underline{x}, \lambda), \underline{v}_\varepsilon^{qr}(\underline{x}, \lambda)]$ and $I_{K_h^\xi}[\underline{v}_\varepsilon^{qr}(\underline{x}, \lambda), \underline{v}_\varepsilon^{qr}(\underline{x}, \lambda)]$ vanish in the limit $\varepsilon \to 0$ and $h \to 0$ ($\varepsilon \ll h$), and therefore the first term on the RHS of (7.8′) is the leading one and the second is a small corrector.

Substituting (7.8′) into (3.10), we obtain

$$\begin{aligned}
\frac{1}{h^3} a_{npqr}(\xi, \lambda, \varepsilon, h) &= E_{K_h^\xi}[\underline{U}_\varepsilon^{np}, \underline{U}_\varepsilon^{qr}] + \frac{1}{\lambda} I_{K_h^\xi}[\underline{U}_\varepsilon^{np}, \underline{U}_\varepsilon^{qr}] \\
&+ H[\underline{U}_\varepsilon^{np} - \underline{\psi}^{np}, \underline{U}_\varepsilon^{qr} - \underline{\psi}^{qr}] + E_{K_h^\xi}[\underline{U}_\varepsilon^{np}, \underline{v}_\varepsilon^{qr}] + E_{K_h^\xi}[\underline{U}_\varepsilon^{qr}, \underline{v}_\varepsilon^{np}] \\
&+ \frac{1}{\lambda} I_{K_h^\xi}[\underline{U}_\varepsilon^{np}, \underline{v}_\varepsilon^{qr}] + \frac{1}{\lambda} I_{K_h^\xi}[\underline{U}_\varepsilon^{qr}, \underline{v}_\varepsilon^{np}] + H[\underline{U}_\varepsilon^{np} - \underline{\psi}^{np}, \underline{v}^{qr}] \\
&+ H[\underline{U}_\varepsilon^{qr} - \underline{\psi}^{qr}, \underline{v}^{np}] + E_{K_h^\xi}[\underline{v}_\varepsilon^{np}, \underline{v}_\varepsilon^{qr}] + \frac{1}{\lambda} I_{K_h^\xi}[\underline{v}_\varepsilon^{np}, \underline{v}_\varepsilon^{qr}] + H[\underline{v}_\varepsilon^{np}, \underline{v}_\varepsilon^{qr}].
\end{aligned} \tag{7.11}$$

Here we have used the notation of (3.6)–(3.7) and introduced the notation $H[\underline{u},\underline{v}] = h^{-2-\tau}\int_{K_h^\xi}\langle \underline{u},\underline{v}\rangle dx$.

On the RHS of (7.11) only the first two terms provide a nonzero contribution in the limit $\varepsilon \to 0$ and $h \to 0$ ($\varepsilon \ll h$). The remaining terms vanish in this limit, which can be shown by a direct calculation. We now calculate these nonvanishing terms. Using (7.8), we can write

$$E_{K_h^\xi}[\underline{U}_\varepsilon^{np}, \underline{U}_\varepsilon^{qr}] = E_{K_h^\xi}[\underline{\psi}_\varepsilon^{np}, \underline{\psi}_\varepsilon^{qr}] + E_{K_h^\xi}[\widetilde{\underline{u}}_\varepsilon^{np}, \widetilde{\underline{u}}_\varepsilon^{qr}]$$

$$+ E_{K_h^\xi}[\underline{\psi}_\varepsilon^{np}, \widetilde{\underline{u}}_\varepsilon^{qr}] + E_{K_h^\xi}[\underline{\psi}_\varepsilon^{qr}, \widetilde{\underline{u}}_\varepsilon^{np}], \tag{7.12}$$

where $\underline{\psi}_\varepsilon^{np} = \underline{\psi}^{np}(\underline{x} - \underline{\xi}_\varepsilon)$, $\widetilde{\underline{u}}_\varepsilon^{np} = \varepsilon\widetilde{\underline{u}}(\frac{\underline{x}-\underline{\xi}_\varepsilon}{\varepsilon}, \lambda)$.

Integration by parts shows that the third and fourth terms on the RHS of (7.12) are equal to zero, and we calculate the first two terms.

To simplify the calculation we assume that the cube K_h^ξ can be partitioned into an integer number of elementary cubes (cells) K_ε^j. Then, taking into account the linearity of $\underline{\psi}^{np}(\underline{x})$ and periodicity of $\widetilde{\underline{u}}^{np}(\underline{x})$, we obtain

$$E_{K_h^\xi}[\underline{\psi}_\varepsilon^{np}, \underline{\psi}_\varepsilon^{qr}] = \frac{\mu}{h^3}\sum_j 2\varepsilon^2 \int_{K_\varepsilon^j} \sum_{k,l=1}^3 e_{kl}\left[\underline{\psi}^{np}\left(\frac{\underline{x} - \underline{x}_\varepsilon^i}{\varepsilon}\right)\right] e_{kl}\left[\underline{\psi}^{qr}\left(\frac{\underline{x} - \underline{x}_\varepsilon^i}{\varepsilon}\right)\right] dx$$

$$= 2\mu \int_K \sum_{k,l=1}^3 e_{kl}[\underline{\psi}^{np}] e_{kl}[\underline{\psi}^{qr}(\underline{x})] dx = \mu \mathbf{I}_{\mathbf{npqr}}, \tag{7.13}$$

$$E_{K_h^\xi}[\widetilde{\underline{u}}_\varepsilon^{np}, \widetilde{\underline{u}}_\varepsilon^{qr}] = \frac{\mu}{h^3}\sum_j 2\varepsilon^2 \int_{K_\varepsilon^j} \sum_{k,l=1}^3 e_{kl}\left[\underline{u}^{np}\left(\frac{\underline{x} - \underline{x}_\varepsilon^i}{\varepsilon}\right)\right] e_{kl}\left[\underline{u}^{qr}\left(\frac{\underline{x} - \underline{x}_\varepsilon^i}{\varepsilon}\right)\right] dx$$

$$= 2\mu \int_K \sum_{k,l=1}^3 e_{kl}[\underline{u}^{np}(\underline{x}, \lambda)] e_{kl}[\underline{u}^{qr}(\underline{x}, \lambda)] dx. \tag{7.14}$$

Here we take into account that the number of summands in $\sum_j$ is equal to $\frac{h^3}{\varepsilon^3}$ and $\widetilde{\underline{u}}^{np}(\underline{x})$ are extended on Q by (7.3), and also we use the notation (7.7).

Using (7.9), (2.8), and the linearity of $\underline{\psi}^{np}(\underline{x})$, we obtain

$$I[\underline{U}_\varepsilon^{np}, \underline{U}_\varepsilon^{qr}] = \frac{1}{2h^3}\sum_{ij}\left\langle \underline{\underline{C}}_{1\varepsilon}^{ij}\underline{\psi}^{np}(\underline{x}_\varepsilon^i - \underline{x}_\varepsilon^j), \underline{\psi}^{qr}(\underline{x}_\varepsilon^i - \underline{x}_\varepsilon^j)\right\rangle$$

$$+ \frac{1}{2h^3}\sum_{ij}\left\langle \left(\underline{\underline{C}}_{3\varepsilon}^{ij} - \underline{\underline{C}}_{4\varepsilon}^{ij}\right)\underline{b}^{np}, \underline{b}^{qr}\right\rangle = \sum_j{}' \left\langle \underline{\underline{C}}_1^{ij}\underline{\psi}^{np}(\underline{x}^j), \underline{\psi}^{qr}(\underline{x}^j)\right\rangle + \langle C\underline{b}^{np}, \underline{b}^{qr}\rangle, \tag{7.15}$$

where the matrices C_1^{ij}, C and the points $\underline{x}^j$ are defined in the formulation of Theorem 7.2. (see also Figure 7.1).

Combining (7.13)–(7.15), we obtain (7.6). The detailed proof is presented in [1].

THEOREM 7.3. *Tensor* $\{a_{npqr}(\lambda)\}$ *obtained in Theorem* 7.2 *can be represented in*

the form

$$a_{npqr}(\lambda) = \mu \mathbf{I}_{npqr} + 2\mu \int_K \sum_{k,l=1}^{3} e_{kl}[\underline{u}_0^{np}] e_{kl}[\underline{u}_0^{qr}] dx$$

$$+\frac{1}{\lambda} \left\{ \sum_j{}' \langle \underline{\underline{C}}_1^{ij} \underline{\psi}^{np}(\underline{x}^j) \underline{\psi}^{qr}(\underline{x}^j) \rangle + \langle \underline{\underline{C}} \underline{b}_0^{np}, \underline{b}_0^{qr} \rangle \right\} + a^1_{npqr}(\lambda). \tag{7.16}$$

Here

$$a^1_{npqr}(\lambda) = 2\mu \left\{ \int_K \sum_{k,l=1}^{3} e_{kl}[\underline{u}_0^{np}(\underline{x})] e_{kl}[\underline{v}^{qr}(\underline{x},\lambda)] dx \right.$$

$$\left. + \int_K \sum_{k,l=1}^{3} e_{kl}[\underline{v}^{np}(\underline{x},\lambda)] e_{kl}[\underline{u}_0^{qr}(\underline{x})] dx + \int_K \sum_{k,l=1}^{3} e_{kl}[\underline{v}^{np}(\underline{x},\lambda)] e_{kl}[\underline{v}^{qr}(\underline{x},\lambda)] dx \right\}$$

$$+\frac{1}{\lambda} \left[\langle \underline{\underline{C}} \underline{b}_0^{np}, \underline{d}^{qr}(\lambda) \rangle + \langle \underline{\underline{C}} \underline{d}^{np}(\lambda), \underline{b}_0^{qr} \rangle + \langle \underline{\underline{C}} \underline{d}^{np}(\lambda), \underline{b}_0^{qr} \rangle \right], \tag{7.17}$$

where the function $\underline{u}_0^{np}(\underline{x})$ *and the constant vector* $\underline{b}_0^{qr}$ *solve the cell problem*

$$-\mu \Delta \underline{u}_0^{np} + \nabla p_0^{np} = 0, \quad \underline{x} \in K \setminus Q,$$
$$\operatorname{div} \underline{u}_0^{np} = 0, \quad \underline{x} \in K,$$
$$\underline{u}_0^{np} = -\underline{\psi}^{np}(\underline{x}) + \underline{b}_0^{np} \times \underline{x}, \quad \underline{x} \in Q, \quad \underline{b}_0 \in \ker \underline{\underline{C}},$$
$$\underline{\underline{\sigma}}_0 := \int_{\partial Q} \underline{x} \times (\underline{\underline{\sigma}}[\underline{u}_0^{np}] \cdot \underline{\nu}) dS \perp \ker \underline{\underline{C}}, \tag{7.18}$$

function $\underline{v}^{np}(\underline{x},\lambda)$ *and vector* $\underline{d}^{np}(\lambda)$ *are the solution of the problem*

$$-\mu \Delta \underline{v}^{np} + \nabla p_0^{np} = 0, \quad \underline{x} \in K \setminus Q,$$
$$\operatorname{div} \underline{v}^{np} = 0, \quad \underline{x} \in K,$$
$$\underline{v}^{np} = \underline{d}(\lambda) \times \underline{x}, \quad \underline{x} \in Q, \quad \underline{d}(\lambda) \perp \ker \underline{\underline{C}},$$
$$\int_{\partial Q} \underline{x} \times (\underline{\underline{\sigma}}[\underline{v}^{np}] \cdot \underline{\nu}) dS + \frac{1}{\lambda} \underline{\underline{C}} \underline{d} = -\underline{\underline{\sigma}}_0 \perp \ker \underline{\underline{C}}, \tag{7.19}$$

$\underline{v}^{np}(\underline{x},\lambda)$ *and* $\underline{\underline{\sigma}}[\underline{v}^{np}]$ *as well as* $\underline{u}_0$ *and* $\underline{\underline{\sigma}}[\underline{u}_0]$ *are periodic in the cube* K*, and* $\underline{\nu}$ *is the inner normal vector to* ∂Q*.*

Proof of Theorem 7.3. Comparison of (7.18)–(7.19) and (7.1)–(7.5) shows that due to linearity the solution of problem (7.1)–(7.5) can be represented in the form

$$\underline{u}^{np}(\underline{x},\lambda) = \underline{u}_0^{np}(\underline{x}) + \underline{v}^{np}(\underline{x},\lambda), \tag{7.20}$$

$$\underline{b}^{np}(\lambda) = \underline{b}_0^{np}(\lambda) + \underline{d}^{np}(\lambda), \tag{7.21}$$

where $\{\underline{u}_0^{np}(\underline{x}), \underline{b}_0^{np}\}$ is the solution of problem (7.18), which does not depend on λ, and $\{\underline{v}^{np}(\underline{x},\lambda), \underline{d}^{np}(\lambda)\}$ is the solution of problem (7.19). Each of these problems has a unique solution. Indeed $\{\underline{u}_0^{np}, \underline{b}^{np}\}$ minimizes the functional

$$I_0 = 2\mu \int_K \sum_{k,l=1}^{3} e_{kl}^2[\underline{u}^{np}] dx$$

in the class of divergence-free in K and periodic functions which are equal to $-\psi^{np}(\underline{x}) + \underline{b} \times \underline{x}$ on Q, with $\underline{b} \in \ker \underline{\underline{C}}$ ($\underline{\underline{C}} = \sum(\underline{\underline{C}}^{ij}_{3\varepsilon} - \underline{\underline{C}}^{ij}_{4\varepsilon})/(2\varepsilon^3)$), and $\{\underline{v}^{np}(\underline{x}, \lambda), \underline{d}^{np}(\underline{x}, \lambda)\}$ minimizes the functional

$$I = 2\mu \int_K \sum_{k,l=1}^{3} e^2_{kl}[\underline{v}^{np}] + \frac{1}{\lambda}\langle \underline{\underline{C}}\underline{d}, \underline{d}\rangle + 2\langle \underline{\underline{\sigma}}_0, \underline{d}\rangle$$

in the class of divergence-free in K and periodic functions equal to $\underline{d} \times \underline{x}$ on Q, $\underline{d} \perp \ker \underline{\underline{C}}$. Note that the pressure does not appear in the above variational formulations and can be computed from the Stokes equation. To obtain (7.16) we substitute (7.20)–(7.21) into (7.6). Theorem 7.3 follows.

We now obtain a detailed characterization of the tensor $\{a^1_{npqr}(\lambda)\}$. The following Theorem 7.4 and Proposition 7.7 show how this characterization allows us to get explicit formulas for the effective viscoelastic properties. Indeed, taking the inverse Laplace transform of the representation (7.16), we observe that the first two terms do not depend on λ and together represent the homogenized tensor of the effective viscosity a^0_{npqr} (see (3.18)); the term with the factor λ^{-1} (the sum in brackets in (7.16)) is the effective elasticity tensor. Tensor $a^1_{npqr}(\lambda)$ represents the memory term, and the numbers $-\lambda_k$ are inverse relaxation times of the effective viscoelastic medium. These numbers are the eigenvalues of the spectral problem which is obtained from (7.19) by setting $\underline{\underline{\sigma}}_0 = 0$ in the last equation in (7.19).

THEOREM 7.4. *The functions $a^1_{npqr}(\lambda)$ obtained in Theorem* 7.3 *are meromorphic in $\mathbf{C} \setminus 0$ with poles at the points $\lambda_k < 0$ ($\lambda_k \to -0, k \to \infty$). The numbers $\lambda_k < 0$ are the eigenvalues of the spectral problem which corresponds to the cell problem* (7.19) *when $\underline{\underline{\sigma}}_0 = 0$.*

Proof of Theorem 7.4. In short, the proof of this theorem consists of the introduction of an operator and a functional space which correspond to the problem (7.19), followed by the standard spectral analysis.

Let us denote by $\mathcal{J}_C(K)$ the closure in $L_2^{per}(K)$ of the set of divergence-free functions from $H^1_{per}(K)$ equal to $\underline{d} \times \underline{x}$ in subdomains Q, where $\underline{d} \perp \ker \underline{\underline{C}}$. The spaces $L_2^{per}(K)$ and $H^1_{per}(K)$ of periodic on K functions is defined in [12]. Next we introduce the space $G(K)$ as a set of functions of the form $\mathrm{grad}\varphi(\underline{x}) + \chi_Q(\underline{x})\underline{\psi}(\underline{x})$, where the function $\varphi(\underline{x}) \in H^1_{per}(K)$ and $\underline{\psi}(\underline{x})$ is a function orthogonal in $L_2(Q)$ to functions of the form $\underline{d} \times \underline{x}$, $\underline{d} \perp \ker \underline{\underline{C}}$.

The next lemma is a straightforward generalization of the standard Weyl decomposition [12] and can be proved analogously. It reduces to the Weyl decomposition in the absence of the rigid inclusion Q.

LEMMA 7.5. *The following orthogonal decomposition holds:*

$$L_2^{per}(K) = \mathcal{J}_C(K) \oplus G(K).$$

Let us introduce the operator $\hat{A}$ acting from $\mathcal{J}_C(K) \cap H^2(K \setminus Q)$ into $L_2(K)$:

$$\hat{A}\underline{u} = \begin{cases} -\mu\delta\underline{u}, & \underline{x} \in K \setminus Q, \\ B_0[\underline{u}] \times \underline{x}, & \underline{x} \in Q. \end{cases}$$

Here vector $B_0[\underline{u}]$ is defined by the equation

$$B_0[\underline{u}] = (I)^{-1} \int_{\partial Q} \underline{x} \times (\underline{\underline{\sigma}}[\underline{u}] \cdot \underline{\nu}) dS,$$

where I is the tensor of inertia of the body Q with the density $\rho_s = 1$ ($\sigma^0_{ik} = \mu e_{ik}(\underline{v})$).

Let P_C be the operator of orthogonal projection on the subspace $\mathcal{J}_C(K)$. Let us define an operator

$$\widetilde{A} = P_C \hat{A}.$$

Direct calculations show that

$$\widetilde{A}\underline{u} = \begin{cases} -\mu\Delta\underline{u} + \nabla p, & \underline{x} \in K \setminus Q, \\ B_0[\underline{u}] \times \underline{x} + (I)^{-1} \int_{\partial Q} (\underline{x} \times \underline{\nu})\, p dS \times \underline{x}, & \underline{x} \in Q. \end{cases}$$

Introduce an operator A as the Friedrichs extension of the operator $\widetilde{A}$.

LEMMA 7.6. *Operator A is self-adjoint in $\mathcal{J}_C(K)$, invertible, and has a compact inverse operator.*

We now introduce an operator

$$A_1 = P_C \hat{A}_1, \tag{7.22}$$

where $\hat{A}_1$ is a bounded operator acting from $\mathcal{J}_C(K)$ into $L_2^{per}(K)$ by the formula

$$\hat{A}_1 \underline{u} = \chi_Q(\underline{x}) \left[I^{-1} \underline{\underline{C}} \underline{d} \times \underline{x} \right],$$

where $\underline{u}$ is defined on Q by the equation $\underline{u} = \underline{d} \times \underline{x}$, $\underline{d} \perp \ker \underline{\underline{C}}$. Taking into account the symmetry of the matrices $\underline{\underline{C}}^{ij}_{3\varepsilon}$, $\underline{\underline{C}}^{ij}_{4\varepsilon}$ (which follows from (2.8)), $C^T = C$, $(\underline{\underline{C}}^{ij}_{3\varepsilon})^T = \underline{\underline{C}}^{ji}_{3\varepsilon}$, $(\underline{\underline{C}}^{ij}_{4\varepsilon})^T = \underline{\underline{C}}^{ji}_{4\varepsilon}$, the periodicity property $\underline{\underline{C}}^{i+k,j+k}_{\varepsilon}(\underline{x}, \underline{y}) = \underline{\underline{C}}^{ij}_{\varepsilon}(\underline{x}, \underline{y})$, a $\underline{k}$-arbitrary integer vector, and (7.4′), we obtain

$$\langle \underline{\underline{C}} \underline{b}, \underline{d} \rangle = \langle \underline{b}, \underline{\underline{C}} \underline{d} \rangle, \tag{7.23}$$

$$\langle \underline{\underline{C}} \underline{b}, \underline{b} \rangle \geq 0 \tag{7.24}$$

for any $\underline{b}, \underline{d} \in \mathbf{R}^3$.

Then since $\underline{b} \perp \ker \underline{\underline{C}}$, we get from (7.24)

$$\langle \underline{\underline{C}} \underline{b}, \underline{b} \rangle > 0, \quad \underline{b} \neq 0. \tag{7.25}$$

Using the definitions of the operator $\hat{A}_1$ and the tensor of inertia $\underline{\underline{I}}$, we show that the operator A_1 defined by (7.22) is a self-adjoint and positive operator in $\mathcal{J}_C(K)$. Indeed,

$$\begin{aligned} (\hat{A}_1 \underline{u}, \underline{v}) &= \int_K \chi_Q(\underline{x}) \langle \underline{\underline{I}} \underline{\underline{C}} \underline{b} \times \underline{x}, \underline{d} \times \underline{x} \rangle dx \\ &= \int_Q \langle I^{-1} \underline{\underline{C}} \underline{b} \times \underline{x}, \underline{d} \times \underline{x} \rangle dx = \langle \underline{\underline{C}} \underline{b}, \underline{d} \rangle = \langle \underline{b}, \underline{\underline{C}} \underline{d} \rangle = (\underline{u}, \hat{A}_1 \underline{v}) \end{aligned} \tag{7.26}$$

and

$$(\hat{A}_1 \underline{u}, \underline{u}) > 0$$

for any $\underline{u}$, $\underline{v} \in \mathcal{J}_C(K)$ ($\underline{u}(\underline{x}) = \underline{b} \times \underline{x}$, $\underline{v} = \underline{d} \times \underline{x}$, $\underline{x} \in Q$, and $\underline{b}, \underline{d} \perp \ker \underline{\underline{C}}$); $(\hat{A}_1 \underline{u}, \underline{v})$ is the dot product in $L_2(K)$.

Similarly we can represent the projection of function $-(\underline{\underline{\sigma}}_0 \times \underline{x})\chi_Q(\underline{x})$ on the subspace $\mathcal{J}_C(K)$:

$$\underline{f}_0 = -P_C[(\underline{\underline{\sigma}}_0 \times \underline{x})\chi_Q(\underline{x})] = \begin{cases} \operatorname{grad} p, \\ -\underline{\underline{\sigma}}_0 \times \underline{x} + I^{-1} \int_{\partial Q} (\underline{x} \times \underline{\nu}) p ds \times \underline{x}. \end{cases}$$

Then the problem (7.19) can be written in the operator form in the space $\mathcal{J}_C(K)$:

$$A\underline{v} + \frac{1}{\lambda} A_1 \underline{v} = \underline{f}_0 \tag{7.27}$$

or equivalently as

$$(B - \lambda I)\underline{v} = \lambda \underline{g}_0, \qquad \underline{g}_0 = A^{-1} f_0, \tag{7.27$'$}$$

where $B = -A^{-1}A_1$ is a compact operator in $\mathcal{J}_C(K)$ whose spectrum consists of the eigenvalues $\lambda_k \neq 0$ with the only accumulation point at 0.

Operator B in bounded and self-adjoint in the energy space H_A of the operator A, that is,

$$(B\underline{u}, \underline{v})_A = (\underline{u}, B\underline{v})_A,$$

and due to the positivity of the operator A_1,

$$(B\underline{u}, \underline{u})_A = -(A_1\underline{u}, \underline{u})_{L_2(K)} < 0.$$

Thus eigenvalues λ_k are negative and the corresponding eigenfunctions are orthogonal in energy space H_A. Eigenfunctions corresponding to multiple eigenvalues can be chosen to be orthogonal in H_A. Due to the Gilbert–Schmidt theorem, the system of eigenfunctions $\{\varphi_k\}$ is complete in $\mathcal{J}_C(K)$. Taking into account this fact, we can easily find the solution $\underline{v}$ of (7.27$'$):

$$\underline{v}(\underline{x}, \lambda) = -\lambda \sum_{k=1}^{\infty} \frac{C_k}{\lambda - \lambda_k} \varphi_k(\underline{x}), \quad C_k = (A\varphi_k, \underline{g}_0). \tag{7.28}$$

Thus, the solution of problem (7.19) is a meromorphic function in the complex plane λ.

Substituting (7.28) into (7.17), we obtain an explicit representation for a^1_{npqr},

$$a^1_{npqr}(\lambda) = \sum_{i=1}^{\infty} \frac{\lambda A_i^{npqr} + C_i^{npqr}}{\lambda - \lambda_i} + \sum_{i,j=1}^{\infty} \frac{\lambda^2 A_{ij}^{npqr} + \lambda C_{ij}^{npqr}}{(\lambda - \lambda_i)(\lambda - \lambda_j)},$$

where the first sum and the second sum correspond to the linear and quadratic terms in (7.17), respectively, and

$$\begin{aligned} A_i^{npqr} &= 2\mu \int_K \sum_{k,l} \left(e_{kl}[\underline{u}_0^{np}] C_i^{qr} + e_{kl}[\underline{u}_0^{qr}] C_i^{np} \right) e_{kl}[\varphi_i] dx, \\ A_{ij}^{npqr} &= 2\mu \int_K \sum_{kl} e_{kl}^2[\varphi_i] C_i^{np} C_j^{qr} dx, \\ C_i^{npqr} &= 2\mu \left\{ \langle \underline{\underline{C}} \underline{b}_0^{np}, \underline{d}_i \rangle C_i^{qr} + \langle \underline{\underline{C}} \underline{d}_i, \underline{b}_0^{qr} \rangle C_i^{np} \right\}, \\ C_{ij}^{npqr} &= 2\mu \langle \underline{\underline{C}} \underline{d}_i, \underline{d}_j \rangle C_i^{np} C_j^{qr}. \end{aligned} \tag{7.29}$$

Finally, we distinguish the λ independent part in $a^1_{npqr}(\lambda)$,

$$a^1_{npqr}(\lambda) = a^{10}_{npqr} + a^{11}_{npqr}(\lambda),$$

where

$$\begin{cases} a^{10}_{npqr} & = \sum_i A_i^{npqr} + \sum_{i,j} A_{ij}^{npqr}, \\ a^{11}_{npqr}(\lambda) & = \sum_i \frac{\lambda_i A_i^{npqr} + C_i^{npqr}}{\lambda - \lambda_i} \\ & \quad + 2\sum_{ij} \frac{(\lambda(\lambda_i+\lambda_j) - \lambda_i\lambda_j)A_{ij}^{npqr} + \lambda C_{ij}^{npqr}}{(\lambda-\lambda_i)(\lambda-\lambda_j)}. \end{cases} \tag{7.30}$$

Then

$$a^{11}_{npqr}(\lambda) = O\left(\frac{1}{\lambda}\right) \quad \text{as} \quad |\lambda| \to \infty.$$

Taking the inverse Laplace transform of $a^1_{npqr}(\lambda)$, we obtain

$$a^1_{npqr}(t) = a^{10}_{npqr}\delta(t) + \sum_{i=1}^{\infty} B_i^{npqr} e^{\lambda_i t}, \tag{7.31}$$

where, according to (7.30),

$$B_i^{npqr} = \lambda_i(A_i^{npqr} + A_{ii}^{npqr}) + C_i^{npqr} + C_{ii}^{npqr} + 2\sum_{j>i} \frac{\lambda_i^2 A_{ij}^{npqr} + \lambda_i C_{ij}^{npqr}}{\lambda_i - \lambda_j} \tag{7.32}$$

and Theorem 7.4 is proved.

We now use Theorems 7.3 and 7.4 to obtain an explicit representation for the relaxation tensor.

PROPOSITION 7.7. *The following representation for the effective tensor* (3.14) *holds:*

(7.33)

$$\hat{a}_{npqr}(t) = a^0_{npqr}\delta(t) + a^1_{npqr}\chi(t) + \sum_{i=1}^{\infty} B_i^{npqr} e^{\lambda_i t} \quad (\lambda_i < 0,\ \lambda_i \to -0, \quad \text{as } i \to \infty),$$

where

$$a^0_{npqr} = \mu \mathbf{I}_{npqr} + 2\mu \int_K \sum_{kl} e_{kl}[\underline{u}_0^{np}] e_{kl}[\underline{u}_0^{qr}] dx + a^{10}_{npqr},$$

$$a^1_{npqr} = \sum_j{}' \langle C_1^{ij} \underline{\psi}^{np}(\underline{x}^j) \underline{\psi}^{qr}(\underline{x}^j) \rangle + \langle \underline{\underline{C}} \underline{b}_0^{np}, \underline{b}_0^{qr} \rangle.$$

Constant tensors a^{10}_{npqr} *and* B_i^{npqr} *are defined in* (7.30), (7.32), *and* (7.29).

The representation (7.33) follows from (7.16) and (7.31).

An important question for the effective rheology of a composite medium is whether there are infinitely many relaxation times present. We now show that the presence of finite relaxations times $-\lambda_i^{-1}$ is caused by the asymmetry of the particles Q^i_ε. Roughly speaking, our calculations show that the finite relaxation times arise due to

the rotation of particles in a symmetric flow, and the effect of this rotation is negligible (to the leading term in the homogenization limit) if particles are symmetric.

PROPOSITION 7.8. *Suppose that the periodicity cell K is invariant with respect to reflections about all three coordinate planes (e.g., a spherical particle Q in a cubic periodicity cell K). Then the effective tensor $a_{npqr}(t)$ defined in* (4.8) *can be computed as follows:*

$$a_{npqr}(t) = a^0_{npqr}\delta(t) + a^1_{npqr}\chi(t), \tag{7.34}$$

where $\delta(t)$ is the Dirac delta-function and $\chi(t)$ is the Heaviside function.

$$a^0_{npqr} = \mu \mathbf{I}_{npqr} + 2\mu \int_K \sum e_{kl}[\underline{u}^{np}_0] e_{kl}[\underline{u}^{qr}_0] dx,$$

$$a^1_{npqr} = \sum_j \langle C^{ij}_1 \underline{\psi}^{np}(\underline{x}^j) \underline{\psi}^{qr}(\underline{x}^j) \rangle,$$

and $\underline{u}^{np}_0(\underline{x})$ is the solution of the problem (7.18) *for $\underline{\underline{C}} \equiv 0$ and $\underline{b}^{np}_0 = 0$.*

We now outline the proof of this proposition. (The details are presented in [1].) Assume that particles Q^i_ε are spherical balls. Then in the problem (7.18), Q is a ball, and taking into account the invariance of this problem under $x_i \to -x_i$, $i = 1, 2, 3$, transformations and the uniqueness of its solution, we conclude that $\underline{b}_0 = 0$ and $\underline{\underline{\sigma}}_0 = 0$. Then it follows from the uniqueness of the solution of the problem (7.19) that $\underline{v}(\underline{x}, \lambda) \equiv 0$ and $\underline{d}(\lambda) = 0$. Next (7.17) implies that $a^1_{npqr}(\lambda) = 0$, and according to (7.31), $a^{10}_{npqr} = 0$ and $B^{npqr}_i = 0$. Thus, in the case of the balls, the effective tensor $a_{npqr}(t)$ has the form (7.34). Note that the equality (7.34) means that in the case of symmetric particles the effective medium is viscoelastic with no memory.

Acknowledgments. Part of this work was done when E. K. was visiting Penn State University. His visit was supported by NSF grant DMS-9971999. We are pleased to thank A. Leonov who brought to our attention the problem of polymeric compounds with interacting filling particles and suggested that we develop a mathematical model for such compounds. We are also grateful to R. Lakes for discussions on the viscoelastic properties of composites and to A. Leonov for very helpful discussions.

REFERENCES

[1] L. V. BERLYAND AND E. YA. KHRUSLOV, *Non-Newtonian Homogenized Model of a Newtonian Fluid with Interacting Particles*, available online at http://www.math.psu.edu/beryland.
[2] R. BURRIDGE AND J. B. KELLER, *Poroelasticity equations derived from microstructure*, J. Acoust. Soc. Amer., 70 (1981), pp. 1140–1146.
[3] R. P. GILBERT AND A. MIKELIC, *Homogenization of acoustic properties of a seabed*, Nonlinear Anal., 40 (2000), pp. 185–212.
[4] A. I. LEONOV, *On rheology of filled polymers*, J. Rheol., 34 (1990), pp. 1039–1068.
[5] R. G. LARSON, *The Structure and Rheology of Complex Fluids*, Oxford University Press, New York, Oxford, 1999.
[6] H. GOLDSTEIN, *Classical Mechanics*, 3rd ed., Addison-Wesley Press, Cambridge, MA, 2002.
[7] G. MASE, *Theory and Problems of Continuum Mechanics*, McGraw-Hill, New York, 1970.
[8] R. CHRISTENSEN, *Mechanics of Composite Materials*, John Wiley & Sons, New York, 1982.
[9] N. D. KOPACHEVSKIY, S. G. KREIN, AND N. Z. KA, *Operator Methods in Linear Hydrodymamics*, Nauka, Moscow, 1989.
[10] O. A. LADYZHENSKAIA, *The Mathematical Theory of Viscous Incompressible Flow*, Translated from Russian by Richard A. Silverman, Gordon and Breach, New York, 1963.
[11] A. I. MARKUSHEVICH, *Theory of Analytic Functions*, Nauka, Moscow, 1968 (in Russian).

[12] V. V. Jikov, S. M. Kozlov, and O. A. Oleynik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.
[13] O. A. Oleynik, A. S. Shamaev, and G. A. Yosifian, *Mathematical Problems in Elasticity and Homogenization*, North–Holland, Amsterdam, 1992.
[14] J. N. Pernin and E. Jacquet, *Elasticity and viscoelasticity in highly heterogeneous composite medium: Threshold phenomenon and homogenization*, Internat. J. Engrg. Sci., 39 (2001), pp. 1655–1689.
[15] J. Sanchez-Hubert, *Asymptotic study of the macroscopic behavior of solid-liquid mixture*, Math. Methods Appl. Sci., 2 (1980), pp. 1–11.
[16] E. Sanchez-Palencia, *Non-homogeneous Media and Vibration Theory*, Springer-Verlag, New York, 1980.
[17] W. B. Russel, D. A. Saville, and W. R. Schowalter, *Colloidal Dispersions*, Cambridge University Press, Cambridge, UK, 1989.
[18] R. Temam, *Navier–Stokes Equations*, North–Holland, Amsterdam, New York, Oxford, 1979.
[19] M. J. Wang, *Effect of polymer-filler and filler-filler interactions on dynamic properties of filled vulcanizates*, Rubb. Chem. Technol., 71 (1998), pp. 520–589.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 1035–1056

SEMI-INFINITE ARRAYS OF ISOTROPIC POINT SCATTERERS. A UNIFIED APPROACH*

C. M. LINTON† AND P. A. MARTIN‡

Abstract. We solve the two-dimensional problem of acoustic scattering by a semi-infinite periodic array of identical isotropic point scatterers, i.e., objects whose size is negligible compared to the incident wavelength and which are assumed to scatter incident waves uniformly in all directions. This model is appropriate for scatterers on which Dirichlet boundary conditions are applied in the limit as the ratio of wavelength to body size tends to infinity. The problem is also relevant to the scattering of an E-polarized electromagnetic wave by an array of highly conducting wires. The actual geometry of each scatterer is characterized by a single parameter in the equations, related to the single-body scattering problem and determined from a harmonic boundary-value problem. Using a mixture of analytical and numerical techniques, we confirm that a number of phenomena reported for specific geometries are in fact present in the general case (such as the presence of shadow boundaries in the far field and the vanishing of the circular wave scattered by the end of the array in certain specific directions). We show that the semi-infinite array problem is equivalent to that of inverting an infinite Toeplitz matrix, which in turn can be formulated as a discrete Wiener–Hopf problem. Numerical results are presented which compare the amplitude of the wave diffracted by the end of the array for scatterers having different shapes.

Key words. scattering, semi-infinite array, Foldy's method, discrete Wiener–Hopf

AMS subject classifications. 74J20, 78A45

DOI. 10.1137/S0036139903427891

1. Introduction.

Many methods exist for studying wave interactions with finite arrays of scatterers. For some simple geometries, methods based on separation of variables can be used. For example, the scattering of a plane wave by an arbitrary finite array of circular cylinders can be reduced to the solution of a rapidly convergent infinite system of linear equations.

For more complicated geometries a different method is needed. One possibility is to express the solution to the multiple scattering problem in terms of the individual scattering characteristics of the elements that make up the array. This leads to the so-called T-matrix approach, which has been used extensively in acoustics and in other fields. Another technique is to formulate the problem as an integral equation by, for example, representing the solution as a distribution of dipoles over all the scatterers. This leads to an integral equation of the second kind for the unknown dipole strength. Discretization of the integral equations typically leads to large, full systems of algebraic equations.

As the size of an array increases, solutions to scattering problems rapidly become computationally expensive. In contrast, the case of an infinite periodic array excited by a plane wave is usually a much simpler proposition. This is because the periodicity allows us to formulate the problem on a single "cell" of the array, with periodic boundary conditions. In terms of integral equations this necessitates the use of a

*Received by the editors May 12, 2003; accepted for publication October 1, 2003; published electronically April 14, 2004.
http://www.siam.org/journals/siap/64-3/42789.html
†Department of Mathematical Sciences, Loughborough University, Leicestershire LE11 3TU, UK (c.m.linton@lboro.ac.uk). The research of this author was supported in part by EPSRC Overseas Travel Grant GR/S31204/01.
‡Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (pamartin@mines.edu).

more complicated Green's function, the efficient computation of which may be an issue; see [16].

Methods by which solutions to infinite array scattering problems can be applied to shed light on associated large finite array problems have been applied previously in the design of large phased array antennas, early examples using a Fourier windowing approach [13, 24], in which some fairly crude assumptions are made about the field near each scatterer. Recently, methods based on integral equations have been devised in which the basic idea is to formulate an integral equation for the difference between the infinite array and finite array solutions; see [26, 21], for example. For large finite one-dimensional arrays this leads to problems formulated on semi-infinite arrays. Associated with these so-called fringe integral equation methods is the analysis of Green's functions for semi-infinite arrays; see [5, 6, 17].

Scattering by a semi-infinite array is thus a problem of considerable interest from both a practical and a theoretical point of view, and very few results are available. Perhaps the only attempt at a general theory is that of Millar [18, 19] based on the analysis of a nonlinear integral equation. Some results have been derived previously for the case of small, widely spaced circular cylinders [12, 11] and for the strip grating at low frequencies [22, 23]. In this paper we consider the general case of a semi-infinite array of identical scatterers which are each small with respect to the incident wavelength and under the assumption of Dirichlet boundary conditions on the scatterers. This problem was considered in [20], where a number of asymptotic results were derived. Our approach is based on Foldy's method [8], which since 1945 has found wide application in multiple scattering problems; for a recent application, see [2], for example. We show that the problems considered in [12, 11] and [22, 23] are special cases, and we construct a general system of equations in which the geometry of the scatterer is characterized by a single parameter. The system of equations can be inverted numerically or, since it is of Toeplitz type, it can be solved explicitly via the discrete Wiener–Hopf technique.

The semi-infinite array problem is closely related to the fully infinite array (i.e., diffraction grating) problem and, following a description of Foldy's method in section 2, we next solve this for the same class of scatterers in section 3. The semi-infinite grating problem is formulated in section 4, including a detailed description of the form of the far field and of the behavior at resonance frequencies. Finally, in section 5 we show how the integral equation approach used in [22, 23] reduces to exactly the same equations as those found in section 4. Many of the technical details are relegated to the appendices.

2. Foldy's method. The classic work on acoustic scattering by semi-infinite gratings is that of Hills and Karp [12, 11], who consider small sound-soft circular cylinders. Their formulation is based on a technique due to Foldy [8]. Foldy considers isotropic point scatterers, meaning that "in the neighborhood of the jth scatterer," the scattered field "will behave like" $A_j G(\mathbf{r}-\mathbf{r}_j)$, where the jth scatterer is centered at $\mathbf{r}_j$, A_j is an unknown amplitude, and G is the free-space Green's function; in two dimensions $G(\mathbf{r}) = H_0(kr)$, where $r = |\mathbf{r}|$ and $H_0 \equiv H_0^{(1)}$ is a Hankel function. Foldy represents the total field as

$$u(\mathbf{r}) = u_{\mathrm{inc}}(\mathbf{r}) + \sum_j A_j G(\mathbf{r}-\mathbf{r}_j), \tag{2.1}$$

where the sum is over all the scatterers. The so-called external field is

$$u_n(\mathbf{r}) \equiv u(\mathbf{r}) - A_n G(\mathbf{r}-\mathbf{r}_n) = u_{\rm inc}(\mathbf{r}) + \sum_{\substack{j \\ j\neq n}} A_j G(\mathbf{r}-\mathbf{r}_j), \tag{2.2}$$

which can be regarded as the "incident field" for the nth scatterer.

Now, characterize the scattering properties of the scatterers by

$$A_n = f_n u_n(\mathbf{r}_n), \tag{2.3}$$

where f_n is "the scattering coefficient for the nth scatterer." Thus, the scattered field is determined by the value of the external field at the center of the scatterer, $\mathbf{r}_n$, together with the quantity f_n (which we will come back to later). Then, (2.2) gives

$$u_n(\mathbf{r}) = u_{\rm inc}(\mathbf{r}) + \sum_{\substack{j \\ j\neq n}} f_j u_j(\mathbf{r}_j) G(\mathbf{r}-\mathbf{r}_j).$$

Evaluating this equation at $\mathbf{r}_n$ gives, after using (2.3),

$$f_n^{-1} A_n = u_{\rm inc}(\mathbf{r}_n) + \sum_{\substack{j \\ j\neq n}} A_j G(\mathbf{r}_n-\mathbf{r}_j), \tag{2.4}$$

which is a linear system of algebraic equations for the amplitudes A_j. The total field is then given by (2.1). When the scatterers are identical $f_n = f_0$, say. This quantity depends on the geometry of the scatterers and is discussed next.

The parameter f_0. The problem of scattering by a small sound-soft cylinder is a problem of low-frequency asymptotics. Using the general theory of Kleinman and Vainberg [14], we find that

$$u_{\rm sc}(\mathbf{r}) \approx f_0 u_{\rm inc}(\mathbf{0}) H_0(kr), \tag{2.5}$$

where the origin is inside the cylinder's cross section S and

$$-\frac{1}{f_0} = \frac{2{\rm i}}{\pi}(\ln k\ell - \delta). \tag{2.6}$$

The complex constant δ occurs in the asymptotic approximation

$$H_0(w) = \frac{2{\rm i}}{\pi}(\ln w - \delta) + O(w^2 \log w) \quad \text{as } w \to 0;$$

thus, $\delta = \ln 2 - C + {\rm i}\pi/2$, where $C \approx 0.5772$ is Euler's constant. The length ℓ in (2.6) depends on the geometry of S. It is determined by solving the following two-dimensional exterior Dirichlet problem for Laplace's equation: $\nabla^2 v = 0$ outside S, $v = 0$ on S, and

$$v = \ln(r/\ell) + o(1) \quad \text{as } r \to \infty.$$

Thus, for an ellipse with semimajor axis a and semiminor axis b, we obtain

$$\ell = \frac{1}{2}(a+b).$$

In particular, for circles of radius a, we have $\ell = a$, and then (2.6) is consistent with $-f_0^{-1} = H_0(ka)$, which is the approximation used by Hills and Karp [12]. Note that our formula for f_0, (2.6), does not depend on the orientation of S.

Let us make a few remarks. First, the approximation (2.5) is a rigorous, asymptotic approximation, valid for small sound-soft cylinders of any cross section. It is not merely a far-field result, but is valid in the near field too. It states that soft cylinders scatter isotropically—the scattered field does not depend on the direction of observation. None of these results is true for sound-hard cylinders (Neumann problem) or for penetrable cylinders (transmission problem), and thus Foldy's original method should be modified for nonsoft cylinders.

3. Infinite grating. We begin with the grating problem and consider the scattering of a plane wave

$$u_{\rm inc} = \mathrm{e}^{\mathrm{i}(\beta x + \alpha y)}, \tag{3.1}$$

where $\alpha = k\sin\psi$ and $\beta = k\cos\psi$, by an infinite row of identical scatterers, located at $(x, y) = (ms, 0)$, $m = 0, \pm1, \pm2, \ldots$, where s is the spacing. We will use polar coordinates (r_m, θ_m) centered on the mth scatterer and defined by

$$x - ms = r_m\cos\theta_m, \qquad y = r_m\sin\theta_m,$$

and we will write (r, θ) for (r_0, θ_0). In terms of (r_m, θ_m) the incident wave is

$$u_{\rm inc} = I_m \mathrm{e}^{\mathrm{i}kr_m\cos(\theta_m - \psi)}, \tag{3.2}$$

where

$$I_m = \mathrm{e}^{\mathrm{i}\beta ms}.$$

For future convenience we define the scattering angles ψ_m, $m = 0, \pm1, \pm2, \ldots$, by

$$\psi_m = \arccos\left(\frac{\beta_m}{k}\right), \qquad \beta_m = \beta + \frac{2m\pi}{s}.$$

If $|\beta_m| < k$, i.e., if

$$-1 < \cos\psi + \frac{2m\pi}{ks} < 1,$$

we say that $m \in \mathcal{M}$ and then $0 < \psi_m < \pi$. These correspond to the angles at which plane waves are scattered from an infinite grating; see (3.9) below. If $|\beta_m| > k$, then ψ_m is no longer real and the appropriate branch of the arccos function is given by

$$\arccos t = \begin{cases} \mathrm{i}\,\mathrm{arccosh}\, t, & t > 1, \\ \pi - \mathrm{i}\,\mathrm{arccosh}(-t), & t < -1, \end{cases} \tag{3.3}$$

with $\mathrm{arccosh}\, t = \ln(t + \sqrt{t^2 - 1})$ for $t > 1$.

Let us apply Foldy's method to the problem of the scattering of a plane wave by an infinite row of identical (small) sound-soft scatterers. The system (2.4) becomes (with B_n as the unknowns)

$$f_0^{-1}B_n = u_{\rm inc}(ns, 0) + \sum_{\substack{m=-\infty \\ m\neq n}}^{\infty} B_m H_0(ks|m-n|), \quad n = 0, \pm1, \pm2, \ldots . \tag{3.4}$$

We have $u_{\rm inc}(ms,0) = \mathrm{e}^{\mathrm{i}\beta sm} = I_m$, and quasi-periodicity (see (3.2)) gives $B_m = I_m B_0$, and then (3.4) gives

$$-f_0^{-1}B_0 + B_0 \sum_{\substack{m=-\infty \\ m\neq n}}^{\infty} H_0(ks|m-n|)I_{m-n} = -1, \quad n = 0, \pm 1, \pm 2, \ldots . \tag{3.5}$$

Hence, $B_0 = (f_0^{-1} - \sigma(\psi))^{-1}$, where

$$\sigma(\psi) = \sum_{\substack{m=-\infty \\ m\neq n}}^{\infty} H_0(ks|m-n|)I_{m-n} = \sum_{j=1}^{\infty}(I_j + I_{-j})H_0(kjs). \tag{3.6}$$

It will be convenient to define a quantity $\mathcal{K}$ by

$$\mathcal{K} = -1/B_0 = \sigma(\psi) - f_0^{-1}, \tag{3.7}$$

so that $B_n = -I_n/\mathcal{K}$.

The far field. From (2.1)

$$u = u_{inc} - \frac{1}{\mathcal{K}} \sum_{m=-\infty}^{\infty} I_m H_0(kr_m).$$

If we insert the integral representation (A.1), we get

$$u = u_{inc} - \frac{1}{\mathcal{K}} \sum_{m=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\beta ms} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-k\gamma(t)|y|}}{\gamma(t)} \mathrm{e}^{\mathrm{i}k(x-ms)t}\, \mathrm{d}t.$$

Now use the Poisson summation formula (B.1) to get

$$u = u_{inc} - 2 \sum_{m=-\infty}^{\infty} \frac{\mathrm{e}^{\mathrm{i}kr\cos(\theta - \mathrm{sgn}(y)\psi_m)}}{ks\mathcal{K}\sin\psi_m}, \tag{3.8}$$

where we have used $\gamma(\beta_m/k) = -\mathrm{i}\sin\psi_m$. Note that $-\mathrm{i}\psi_m$ is real and positive if $|\beta_m/k| > 1$, and so the terms in the sum for these values of m decay as $|y| \to \infty$. The far field involves only those m for which $m \in \mathcal{M}$ and thus, as $y \to \pm\infty$,

$$u \sim u_{inc} - 2 \sum_{m\in\mathcal{M}} \frac{\mathrm{e}^{\mathrm{i}kr\cos(\theta \mp \psi_m)}}{ks\mathcal{K}\sin\psi_m}. \tag{3.9}$$

The scattered field, which is symmetric about the x-axis, thus consists of a number of plane waves, that number increasing as ks increases. In $y > 0$, these waves make angles ψ_m with the positive x-axis. For sufficiently small ks there is just one plane wave corresponding to $m = 0$.

Resonance. For large j we have

$$(I_j + I_{-j})H_0(kjs) \sim (\mathrm{e}^{\mathrm{i}j\beta s} + \mathrm{e}^{-\mathrm{i}j\beta s})\sqrt{\frac{2}{\pi jks}}\mathrm{e}^{\mathrm{i}(kjs - \frac{1}{4}\pi)},$$

and so the sum in (3.6) fails to converge if $(k \pm \beta)s = 2n\pi$ for some integer n. This condition corresponds to $\beta_n = \pm k$ for some integer n, which implies that $\psi_n = 0$ or

$\psi_n = \pi$. An alternative expression for $\sigma(\psi)$ is given in (C.1), which shows that as ψ_n approaches 0 or π,

$$\mathcal{K} \sim \sigma(\psi) \sim 2(ks \sin \psi_n)^{-1}.$$

Such a situation is termed resonance, and all the coefficients B_m tend to zero in this limit. The field is not zero, though. Indeed, from (3.8) we have

$$u = u_{\rm inc} - \mathrm{e}^{\mathrm{i}kx \cos \psi_n}$$

in this limit, so that the scattered field reduces to a wave propagating along the grating, either towards $x = \infty$ ($\psi_n = 0$) or towards $x = -\infty$ ($\psi_n = \pi$). For simplicity, we will exclude the possibility that ks is an integer multiple of π so that we cannot satisfy $\cos \psi_n = -1$ and $\cos \psi_m = 1$ simultaneously.

4. Semi-infinite grating. Suppose now that we have a semi-infinite grating of scatterers located along the positive x-axis at $(x, y) = (ms, 0)$, $m = 0, 1, 2, \ldots$. Again, as the scatterers are identical, we have $f_m \equiv f_0$. Then, the scattered field is given from (2.1) by

$$u_{\rm sc} = \sum_{n=0}^{\infty} A_n H_0(kr_n), \tag{4.1}$$

where the coefficients A_n are found to satisfy

$$A_n - f_0 \sum_{\substack{m=0 \\ m \neq n}}^{\infty} A_m H_0(ks|m-n|) = f_0 I_n, \quad n = 0, 1, 2, \ldots , \tag{4.2}$$

which is equivalent to [12, (3.1-1)] (apart from a missing A_n) and [20, (41)]. If we write $A_n = I_n B_0 + C_n$, where $B_0 = -\mathcal{K}^{-1}$ is the solution to the corresponding infinite grating problem, then we find that, for $n = 0, 1, 2, \ldots$,

$$C_n - f_0 \sum_{\substack{m=0 \\ m \neq n}}^{\infty} C_m H_0(ks|m-n|) = \frac{f_0}{\mathcal{K}} \sum_{j=n+1}^{\infty} I_{n-j} H_0(kjs). \tag{4.3}$$

Both (4.2) and (4.3) can be written in terms of Toeplitz matrices, which can be inverted either directly using numerical truncation or via the discrete Wiener–Hopf technique as described in Appendix E. Note that in order to compute the slowly convergent sum on the right-hand side of (4.3) we use (C.2).

For large n, the sum over j on the right-hand side of (4.3) satisfies

$$\sum_{j=n+1}^{\infty} I_{n-j} H_0(kjs) \sim -\sqrt{\frac{2}{\pi kns}}\, \frac{\mathrm{e}^{-\frac{1}{4}\mathrm{i}\pi} \mathrm{e}^{\mathrm{i}ksn}}{[1 - (-1)^q \mathrm{e}^{-\mathrm{i}(k-\beta)s/2}]},$$

where we have used the asymptotic form for the Hankel function with large argument and (D.1), and q is such that $2q\pi < (k - \beta)s < 2(q + 1)\pi$ (i.e., $\beta_q < k < \beta_{q+1}$). One might expect, therefore, that $C_n = O(n^{-1/2})$ as $n \to \infty$. In fact, calculations show that $C_n = O(n^{-3/2})$ as $n \to \infty$. The same decay rate is observed in [22] and is consistent with behavior in the equivalent half-plane problem. The reason for this faster-than-expected decay is the presence of the phase $\exp(\mathrm{i}ksn)$ in the large n behavior of the right-hand side; see (E.8), (E.10), and (E.12).

The far field. The approximation $C_m = 0$ leads to what is known as the Kirchhoff solution. In this case, (4.1) becomes, after the substitution of the integral representation for the Hankel function given by (A.1) and use of (B.2),

$$u_{\rm sc}^{\rm K} = \frac{\mathrm{i}}{\pi\mathcal{K}} \int\!\!\!\!\!\!-_{-\infty}^{\infty} \frac{\mathrm{e}^{-k\gamma(t)|y|}}{\gamma(t)} \frac{\mathrm{e}^{\mathrm{i}kxt}}{1-\mathrm{e}^{\mathrm{i}s(\beta-kt)}}\,\mathrm{d}t.$$

The total field is given by

$$u_{\rm sc} = u_{\rm sc}^{\rm K} - \frac{\mathrm{i}}{\pi}\int_{-\infty}^{\infty} \frac{\mathrm{e}^{-k\gamma(t)|y|}}{\gamma(t)} \mathrm{e}^{\mathrm{i}kxt} \sum_{n=0}^{\infty} C_n \mathrm{e}^{-\mathrm{i}knst}\,\mathrm{d}t.$$

From (F.1), for $-\pi < \theta < \pi$,

(4.4)
$$u_{\rm sc}^{\rm K}(r,\theta) \sim \frac{\mathrm{e}^{\mathrm{i}(kr-\frac{1}{4}\pi)}}{\mathcal{K}(\mathrm{e}^{\mathrm{i}ks(\cos\psi-\cos\theta)}-1)}\sqrt{\frac{2}{\pi kr}} - \sum_{\substack{m\in\mathcal{M}\\ \psi_m>|\theta|}} \frac{2\mathrm{e}^{\mathrm{i}kr\cos(|\theta|-\psi_m)}}{ks\mathcal{K}\sin\psi_m} \quad \text{as } kr\to\infty,$$

with the addition of the correction term $\tilde{I}$ given by (F.2) when $|\theta|$ is close to ψ_p, and

$$u_{\rm sc} \sim u_{\rm sc}^{\rm K} + \sqrt{\frac{2}{\pi kr}}\mathrm{e}^{\mathrm{i}(kr-\frac{1}{4}\pi)} \sum_{n=0}^{\infty} C_n \mathrm{e}^{-\mathrm{i}kns\cos\theta}. \tag{4.5}$$

Just as in the infinite grating problem, the scattered field is symmetric about the x-axis. To simplify the discussion of the far field we will assume that $y > 0$ (i.e., $0 < \theta < \pi$). If we define

$$\tilde{H}(kr) = \sqrt{\frac{2}{\pi kr}}\mathrm{e}^{\mathrm{i}(kr-\frac{1}{4}\pi)}$$

and

$$g(\theta,\psi) = \frac{1}{\mathcal{K}(\mathrm{e}^{\mathrm{i}ks(\cos\psi-\cos\theta)}-1)} + \sum_{n=0}^{\infty} C_n \mathrm{e}^{-\mathrm{i}kns\cos\theta}, \tag{4.6}$$

then we have

$$u_{\rm sc} \sim g(\theta,\psi)\tilde{H}(kr) - \sum_{\substack{m\in\mathcal{M}\\ \psi_m>\theta}} \frac{2\mathrm{e}^{\mathrm{i}kr\cos(\theta-\psi_m)}}{ks\mathcal{K}\sin\psi_m}. \tag{4.7}$$

The far field thus consists of a circular wave of "amplitude" $g(\theta,\psi)$ and a set of plane waves. These plane waves propagate in the same directions as for the infinite grating but do not exist everywhere. The plane wave making an angle ψ_m with the positive x-axis is found only in the sector $\theta < \psi_m$. It is apparent that the coefficients C_n affect only the circular wave, but that the plane wave field is determined solely by the Kirchhoff solution. In the numerical results presented below we will thus focus only on the circular wave.

The lines $\theta = \psi_m$ are known as shadow boundaries, and the circular wave becomes infinite as the shadow boundaries are approached. In fact near these lines we should add a term to $g(\theta,\psi)$ given from (F.2) by

$$\tilde{g}(r,\theta,\psi) = \frac{\mathrm{i}(1+2\mathrm{i}\zeta_p\mathrm{e}^{-\mathrm{i}\zeta_p^2}F(\zeta_p))}{2ks\mathcal{K}\sin\frac{1}{2}(\theta-\psi_p)\sin\psi_p},$$

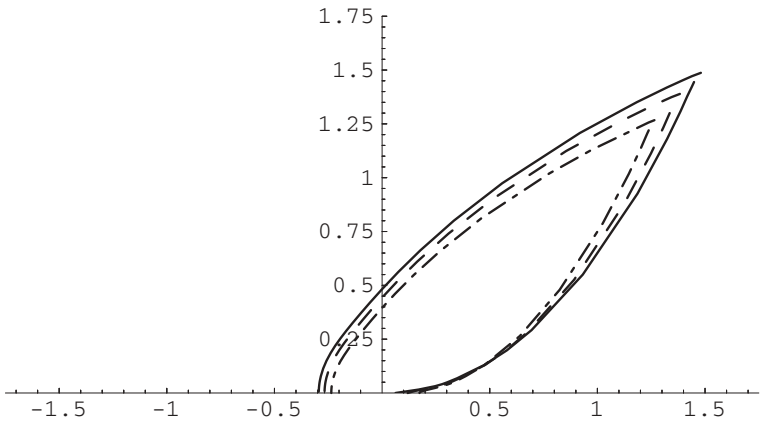


FIG. 4.1. *The amplitude of the circular wave, $|g(\theta,\psi)+\tilde{g}(r,\theta,\psi)|$, for three different scatterers when $ka=0.05$, $ks=2$, $kr=20$, and $\psi=\pi/4$. The solid line corresponds to an array of circles of radius a, the dashed line corresponds to ellipses with $a/b=2$, and the dash-dot line corresponds to plates of length $2a$.*

where $\zeta_p = \sqrt{2kr}\sin\frac{1}{2}|\theta-\psi_p|$ and F is the Fresnel integral defined in (F.3). The combination $g+\tilde{g}$ is bounded as $\theta\to\psi_p$ for any r, but the limit is different from each side. Since $F(0)=\frac{1}{2}\sqrt{\pi}\exp(\mathrm{i}\pi/4)$, the discontinuity in $g+\tilde{g}$ as θ passes through ψ_p exactly cancels the extra residue contribution that appears in the sum in (4.7) as the shadow boundary is crossed.

Hills and Karp [12] introduced the characteristic angles $\psi_m(0)$. These are the real scattering angles when the incident wave angle is zero, i.e.,

$$\psi_m(0)=\arccos\left(1+\frac{2m\pi}{ks}\right),\quad m=-[ks/\pi],\ldots,-1,0.$$

It follows from (4.6) and (E.13) that in the full solution the circular wave vanishes in these directions, i.e., $g(\psi_m(0),\psi)=0$. Hills and Karp state this result only for large ks, but it is true for any value of ks, provided that the Wiener–Hopf factorization described in Appendix E exists. Calculations in [22] suggest that this result is true for moderate values of ks (about 3.5) but perhaps not for small ks (about 0.7). Our calculations based on (4.3) indicate that $g(0,\psi)=0$ for all ks. (For $ks<\pi$ there is only one characteristic angle, namely, $\theta=0$.)

In Figures 4.1–4.3 we show as polar plots the amplitude of the circular wave, $|g(\theta,\psi)+\tilde{g}(r,\theta,\psi)|$, for three different scatterers when $ka=0.05$, $\psi=\pi/4$, and $kr=20$. In Figure 4.1, $ks=2$ and the scatterers are fairly close together, whereas in Figure 4.3, $ks=10$ and the scatterers are well separated. Figure 4.2 represents an intermediate case with $ks=5$. In each case the three different (discontinuous) curves correspond to an array of circles of radius a (solid lines), ellipses with semimajor axis a and semiminor axis $\frac{1}{2}a$ (dashed lines), and plates of length $2a$ (dash-dot lines). The quantity f_0 is calculated from (2.6) in each case with $\ell=a$, $\frac{3}{4}a$, and $\frac{1}{2}a$, respectively.

It can be seen that the general form of each of the curves is the same but that circles produce the diffracted wave with the largest amplitude, while the plates produce the smallest effect. The scales on the three figures are not the same, and it is clear that the amplitude of the diffracted wave generally diminishes as ks increases (though not necessarily for a given observation angle). The numerical results were obtained by direct truncation of (4.3) and checked against (E.10). In order to represent the zeros

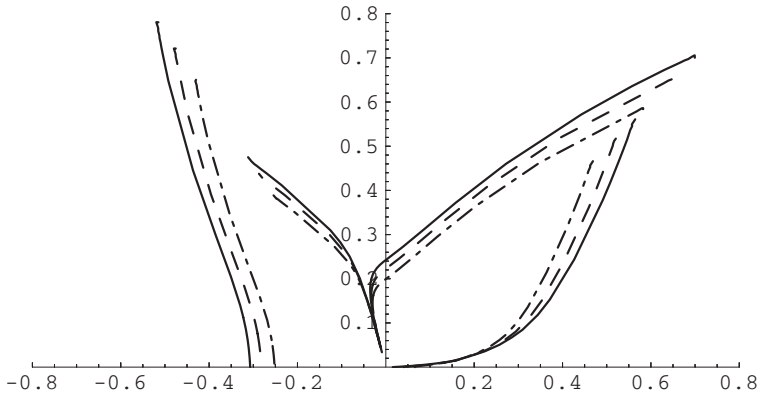


FIG. 4.2. *The amplitude of the circular wave, $|g(\theta,\psi)+\tilde{g}(r,\theta,\psi)|$, for three different scatterers when $ka = 0.05$, $ks = 5$, $kr = 20$, and $\psi = \pi/4$. The solid line corresponds to an array of circles of radius a, the dashed line corresponds to ellipses with $a/b = 2$, and the dash-dot line corresponds to plates of length $2a$.*

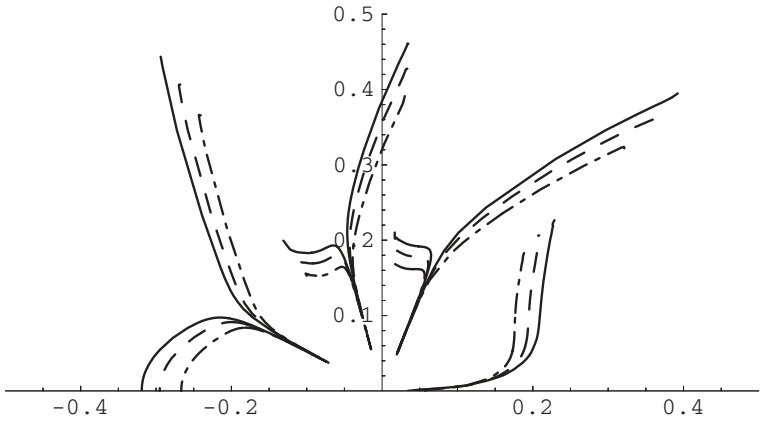


FIG. 4.3. *The amplitude of the circular wave, $|g(\theta,\psi)+\tilde{g}(r,\theta,\psi)|$, for three different scatterers when $ka = 0.05$, $ks = 10$, $kr = 20$, and $\psi = \pi/4$. The solid line corresponds to an array of circles of radius a, the dashed line corresponds to ellipses with $a/b = 2$, and the dash-dot line corresponds to plates of length $2a$.*

in the directions $\psi_m(0)$ with reasonable accuracy, a 200×200 system of equations was used, though the main features of the solution are accurately represented if a much smaller truncation is used.

In Figure 4.1 there is just one predominant scattering direction corresponding to the direction of the incident wave, and the amplitude of the wave is zero for $\theta = \psi_0(0) = 0$. In Figure 4.2 there are two predominant scattering directions corresponding to $\psi_0 = \pi/4$ and $\psi_{-1} = \arccos(1/\sqrt{2} - 2\pi/5) \approx 0.685\pi$. The amplitude of the wave is zero in the directions $\psi_0(0) = 0$ and $\psi_{-1}(0) = \arccos(1 - 2\pi/5) \approx 0.583\pi$. In Figure 4.3 there are three predominant scattering directions corresponding to $\psi_0 = \pi/4$, $\psi_{-1} = \arccos(1/\sqrt{2} - \pi/5) \approx 0.475\pi$, and $\psi_{-2} = \arccos(1/\sqrt{2} - 2\pi/5) \approx 0.685\pi$. The amplitude of the wave is zero in the directions $\psi_0(0) = 0$, $\psi_{-1}(0) = \arccos(1 - \pi/5) \approx 0.379\pi$, $\psi_{-2}(0) = \arccos(1 - 2\pi/5) \approx 0.583\pi$, and $\psi_{-3}(0) = \arccos(1 - 3\pi/5) \approx 0.846\pi$.

Inward and outward resonance. For fixed n the sum on the right-hand side of (4.3) converges unless $(k-\beta)s = 2m\pi$ for some integer m (in other words, unless $\psi_m = 0$ for some integer m). Following Hills and Karp [12], we call this *inward resonance*. We also use the term *outward resonance* for the case when $\psi_m = \pi$ for some integer m. In either situation $1/\mathcal{K} = 0$. Provided we do not have inward resonance, $C_n \to 0$ as $n \to \infty$.

For outward resonance, we get simply $C_n = 0$ and

$$u_{\rm sc} = u_{\rm sc}^{\rm K} = -\frac{1}{\mathcal{K}} \sum_{m=0}^{\infty} I_m H_0(kr_m) = \frac{\mathrm{i}}{\pi\mathcal{K}} \int\!\!\!\!\!\!\diagup_{-\infty}^{\infty} \frac{\mathrm{e}^{-k\gamma(t)|y|}}{\gamma(t)} \frac{\mathrm{e}^{\mathrm{i}kxt}}{1-\mathrm{e}^{\mathrm{i}s(\beta-kt)}}\,\mathrm{d}t.$$

As $\psi_n \to \pi$ we have $\mathcal{K} \sim 2/(ks\sin\psi_n)$, and in order to find the singular behavior of the integral in this limit (which corresponds to one of the poles of the integrand coinciding with the branch point of γ at $t=-1$; the branch cut extending to $-\mathrm{i}\infty$) we deform the contour so that it passes above the pole at β_n/k and thus pick up a contribution $2\pi\mathrm{i}\mathrm{e}^{-\mathrm{i}kx}/(ks\sin\psi_n)$, the remaining integral being finite. Thus, for outward resonance, we obtain

$$u_{\rm sc} = -\mathrm{e}^{-\mathrm{i}kx}$$

in agreement with [12, (3.6-1)].

In the inward resonance case $\psi_n \to 0$, the same arguments show that the Kirchhoff solution $u_{\rm sc}^{\rm K}$ approaches $-\mathrm{e}^{\mathrm{i}kx}$, but the region of existence shrinks to the line $\theta = 0$, and this no longer represents the total scattered field; see [11]. We have not considered the case where inward and outward resonance occur together (which can only happen if ks is an integer multiple of π). Some results for this case can be found in [20].

5. Semi-infinite strip grating. Here we will demonstrate that the analysis given in [22, 23] for a semi-infinite strip grating can be reduced to the general form given in section 4. Consider first an infinite set of strips $S_n = (ns-a, ns+a)$, $n = 0, \pm1, \pm2, \ldots$ $(s > 2a)$, on which we have $u = 0$ and write $S = \bigcup_{n=-\infty}^{\infty} S_n$. We wish to solve the Helmholtz equation in $y > 0$ with

$$\begin{aligned} \frac{\partial u_{\rm sc}}{\partial y} &= 0 \qquad \text{on } y = 0,\ x \notin S, \\ u_{\rm sc} &= -\mathrm{e}^{\mathrm{i}\beta x} \quad \text{on } y = 0,\ x \in S. \end{aligned}$$

Define $v_{\rm sc}(x) \equiv \partial u_{\rm sc}/\partial y|_{y=0}$. Then we have the integral equation

$$\int_S v_{\rm sc}(\xi) G(x-\xi, 0)\,\mathrm{d}\xi = -\mathrm{e}^{\mathrm{i}\beta x}, \quad x \in S,$$

where $G(X,Y) = -\frac{1}{2}\mathrm{i}H_0(k\sqrt{X^2+Y^2})$ is the free-space Green's function. Equivalently, since $v_{\rm sc}(\xi + ms) = I_m v_{\rm sc}(\xi)$,

$$\sum_{m=-\infty}^{\infty} I_m \int_{-a}^{a} v_{\rm sc}(\xi) G(x-\xi-ms, 0)\,\mathrm{d}\xi = -\mathrm{e}^{\mathrm{i}\beta x}, \quad |x| < a,$$

which can be written

$$\int_{-a}^{a} v_{\rm sc}(\xi) G_\beta(x-\xi, 0)\,\mathrm{d}\xi = -\mathrm{e}^{\mathrm{i}\beta x}, \quad |x| < a, \tag{5.1}$$

where we have defined

$$G_\beta(X,Y)=\sum_{m=-\infty}^{\infty} I_m G(X-ms,Y)=-\frac{\mathrm{i}}{ks}\sum_{m=-\infty}^{\infty}\frac{\mathrm{e}^{\mathrm{i}k|Y|\sin\psi_m}\mathrm{e}^{\mathrm{i}\beta_m X}}{\sin\psi_m}$$

using (B.1). The scattered field is then represented by

$$u_{\mathrm{sc}}(x,y)=\int_S v_{\mathrm{sc}}(\xi)G(x-\xi,y)\,\mathrm{d}\xi=\int_{-a}^{a} v_{\mathrm{sc}}(\xi)G_\beta(x-\xi,y)\,\mathrm{d}\xi.$$

Now consider a semi-infinite set of strips $S_n=(ns-a,ns+a)$, $n=0,1,2,\ldots$ $(s>2a)$, and write $S^+=\bigcup_{n=-\infty}^{\infty}S_n$. We have

$$\int_{S^+} v_{\mathrm{sc}}^+(\xi)G(x-\xi,0)\,\mathrm{d}\xi=-\mathrm{e}^{\mathrm{i}\beta x},\quad x\in S^+,$$

and

$$u_{\mathrm{sc}}^+(x,y)=\int_{S^+} v_{\mathrm{sc}}^+(\xi)G(x-\xi,y)\,\mathrm{d}\xi. \tag{5.2}$$

Write $u_{\mathrm{sc}}^+=\phi+u_{\mathrm{sc}}$ and $\nu(x)\equiv\partial\phi/\partial y|_{y=0}$. Then the integral equation becomes

$$\sum_{m=0}^{\infty}\int_{-a}^{a}\nu(\xi+ms)G(x-\xi-ms,0)\,\mathrm{d}\xi+\int_{-a}^{a}v_{\mathrm{sc}}(\xi)G_\beta^+(x-\xi,0)\,\mathrm{d}\xi=-\mathrm{e}^{\mathrm{i}\beta x},\quad x\in S^+, \tag{5.3}$$

where we have defined

$$G_\beta^+(X,Y)=\sum_{m=0}^{\infty}I_m G(X-ms,Y)=-\frac{1}{2\pi}\int\!\!\!\!\!\!\diagup_{-\infty}^{\infty}\frac{\mathrm{e}^{-k\gamma(t)|Y|}}{\gamma(t)}\frac{\mathrm{e}^{\mathrm{i}kXt}}{1-\mathrm{e}^{\mathrm{i}s(\beta-kt)}}\,\mathrm{d}t, \tag{5.4}$$

using (B.2). Equation (5.3) is the starting point for the numerical calculations given in [22, 23]. Alternatively, using (5.1),

$$\sum_{m=0}^{\infty}\int_{-a}^{a}\nu(\xi+ms)G(x-\xi-ms,0)\,\mathrm{d}\xi=\int_{-a}^{a}v_{\mathrm{sc}}(\xi)G_\beta^-(x-\xi,0)\,\mathrm{d}\xi,\quad x\in S^+, \tag{5.5}$$

where $G_\beta^-=G_\beta-G_\beta^+$.

Under the assumption that $ka\ll 1$, we make the approximations (as in [22, 23])

$$v_{\mathrm{sc}}(x)=\frac{2\mathrm{i}B}{\pi\sqrt{a^2-x^2}},\qquad \nu(x+ms)=\frac{2\mathrm{i}C_m}{\pi\sqrt{a^2-x^2}},\quad m=0,1,2,\ldots,$$

with $|x|<a$ in all cases. To determine B we average (5.1) so that

$$\frac{2\mathrm{i}B}{\pi}\int_{-a}^{a}\int_{-a}^{a}\frac{G_\beta(x-\xi,0)}{\sqrt{a^2-\xi^2}}\,\mathrm{d}\xi\,\mathrm{d}x=-\int_{-a}^{a}\mathrm{e}^{\mathrm{i}\beta x}\,\mathrm{d}x=-\frac{2}{\beta}\sin\beta a\approx -2a$$

or

$$B=-\left(\sum_{m=-\infty}^{\infty}I_m\mathcal{G}_m\right)^{-1}, \tag{5.6}$$

where we have defined

$$\mathcal{G}_n = \frac{\mathrm{i}}{\pi a}\int_{-a}^{a}\int_{-a}^{a}\frac{G(x-\xi-ns,0)}{\sqrt{a^2-\xi^2}}\,\mathrm{d}\xi\,\mathrm{d}x, \tag{5.7}$$

for which approximate values, valid for $ka \ll 1$, can be derived; see (G.1) and (G.2).

To determine C_m we average (5.5) so that for each $n = 0, 1, 2, \ldots$

$$\sum_{m=0}^{\infty} C_m \int_{ns-a}^{ns+a}\int_{-a}^{a}\frac{G(x-\xi-ms,0)}{\sqrt{a^2-\xi^2}}\,\mathrm{d}\xi\,\mathrm{d}x = B\int_{ns-a}^{ns+a}\int_{-a}^{a}\frac{G_\beta^-(x-\xi,0)}{\sqrt{a^2-\xi^2}}\,\mathrm{d}\xi\,\mathrm{d}x$$

or

$$\sum_{m=0}^{\infty} C_m\mathcal{G}_{m-n} = B\sum_{m=-\infty}^{-1} I_m\mathcal{G}_{m-n},$$

which, once B is determined from (5.6), is an infinite system of algebraic equations for the unknowns.

If we substitute the approximate values for $\mathcal{G}_n$ ($n > 0$) from (G.2), we get $B = -1/(\mathcal{G}_0 + \sigma(\psi))$, and then

$$C_n + \frac{1}{\mathcal{G}_0}\sum_{\substack{m=0\\ \neq n}}^{\infty} C_m H_0(k|n-m|s) = -\frac{1/\mathcal{G}_0}{\mathcal{G}_0+\sigma(\psi)}\sum_{j=n+1}^{\infty} I_{n-j}H_0(kjs), \tag{5.8}$$

which is of exactly the same form as (4.3), since $\mathcal{G}_0 = -1/f_0$ (compare (G.1) and (2.6) with $\ell = a/2$).

The field is then given, from (5.2), by

$$\begin{aligned} u_{\mathrm{sc}}^+(x,y) &= \int_{S^+}(\nu(\xi)+v_{\mathrm{sc}}(\xi))G(x-\xi,y)\,\mathrm{d}\xi \\ &= \frac{2\mathrm{i}}{\pi}\sum_{m=0}^{\infty} C_m\int_{-a}^{a}\frac{G(x-\xi-ms,y)}{\sqrt{a^2-\xi^2}}\,\mathrm{d}\xi + \frac{2\mathrm{i}B}{\pi}\int_{-a}^{a}\frac{G_\beta^+(x-\xi,y)}{\sqrt{a^2-\xi^2}}\,\mathrm{d}\xi. \end{aligned}$$

The last integral is, using (5.4),

$$-\frac{1}{2\pi}\fint_{-\infty}^{\infty}\int_{-a}^{a}\frac{\mathrm{e}^{-\mathrm{i}k\xi t}\,\mathrm{d}\xi}{\sqrt{a^2-\xi^2}}\frac{\mathrm{e}^{\mathrm{i}kxt-k\gamma(t)|y|}}{\gamma(t)(1-\mathrm{e}^{\mathrm{i}s(\beta-kt)})}\,\mathrm{d}t = -\frac{1}{2}\fint_{-\infty}^{\infty}\frac{J_0(kat)\mathrm{e}^{\mathrm{i}kxt-k\gamma(t)|y|}}{\gamma(t)(1-\mathrm{e}^{\mathrm{i}s(\beta-kt)})}\,\mathrm{d}t$$

and similarly for the first integral so that

$$\begin{aligned} u_{\mathrm{sc}}^+(x,y) = &-\frac{\mathrm{i}}{\pi}\sum_{m=0}^{\infty} C_m\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-k\gamma(t)|y|}}{\gamma(t)}J_0(kat)\mathrm{e}^{\mathrm{i}k(x-ms)t}\,\mathrm{d}t \\ &-\frac{\mathrm{i}B}{\pi}\fint_{-\infty}^{\infty}\frac{\mathrm{e}^{-k\gamma(t)|y|}}{\gamma(t)}\frac{J_0(kat)\mathrm{e}^{\mathrm{i}kxt}}{1-\mathrm{e}^{\mathrm{i}s(\beta-kt)}}\,\mathrm{d}t, \end{aligned}$$

which, via the results in Appendix F and utilizing the fact that $ka \ll 1$, leads to precisely the same far-field asymptotics as that given in (4.5).

Note that exactly the same far-field is generated for a semi-infinite array of angled plates, since f_0 is independent of the plate orientation. Indeed, the plates in the array may all be oriented in different directions.

6. Conclusion. Using a mixture of analysis (discrete Wiener–Hopf) and computation, we have studied the problem of acoustic scattering by a semi-infinite periodic array of identical isotropic point scatterers, i.e., scatterers which are small compared to the incident wavelength and on which Dirichlet boundary conditions are applied. The actual geometry is characterized by a single parameter in the equations, related to the single-body scattering problem. Numerical results have been presented which show the effect of the shape of the scatterer on the form of the circular wave diffracted by the end of the array.

Computations for semi-infinite arrays under less restrictive assumptions than that of isotropic point scatterers do not appear to be available in the literature, and we are currently extending the techniques developed in this paper to study problems in which the size of the individual scatterers is not necessarily small, and to include boundary conditions other than those of Dirichlet type.

Appendix A. Integral representations for Hankel functions. We start from the integral representation, valid for $0 < \theta < \pi$ (i.e., $y > 0$),

$$H_0(kr) = -\frac{\mathrm{i}}{\pi}\int_{-\infty}^{\infty+\mathrm{i}\pi} \mathrm{e}^{\mathrm{i}kx\cosh\alpha}\mathrm{e}^{ky\sinh\alpha}\,\mathrm{d}\alpha.$$

This integral can be converted into a single integral along the real axis. We first split the integral into three parts, namely $(-\infty, 0)$, $(0, \mathrm{i}\pi)$, and $(\mathrm{i}\pi, \infty + \mathrm{i}\pi)$, and make the substitutions $\alpha = -\operatorname{arccosh} t$, $\alpha = \mathrm{i}\arccos t$, and $\alpha = \mathrm{i}\pi + \operatorname{arccosh}(-t)$, respectively. This leads, noting that $H_0(kr)$ is symmetric in y, to

$$H_0(kr) = -\frac{\mathrm{i}}{\pi}\int_{-\infty}^{\infty} \frac{\mathrm{e}^{-k\gamma(t)|y|}}{\gamma(t)}\mathrm{e}^{\mathrm{i}kxt}\,\mathrm{d}t, \tag{A.1}$$

valid for all y, where

$$\gamma(t) = \begin{cases} -\mathrm{i}\sqrt{1-t^2}, & |t| \le 1, \\ \sqrt{t^2-1}, & |t| > 1. \end{cases} \tag{A.2}$$

Appendix B. Summation formulas. We can define two generalized functions by

$$\sum_{m=1}^{\infty}\cos mu = -\frac{1}{2} + \pi\sum_{m=-\infty}^{\infty}\delta(u-2m\pi), \qquad \sum_{m=1}^{\infty}\sin mu = \frac{1}{2}\cot\frac{1}{2}u$$

(see [9, section 2.4] for more details), from which we can construct the generalized functions

$$\sum_{m=-\infty}^{\infty}\mathrm{e}^{\pm\mathrm{i}mu} = 2\pi\sum_{m=-\infty}^{\infty}\delta(u-2m\pi)$$

and

$$\sum_{m=0}^{\infty}\mathrm{e}^{\pm\mathrm{i}mu} = \frac{1}{1-\mathrm{e}^{\pm\mathrm{i}u}} + \pi\sum_{m=-\infty}^{\infty}\delta(u-2m\pi).$$

Hence

$$\sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty} f(u)\mathrm{e}^{-\mathrm{i}mu}\,\mathrm{d}u = 2\pi\sum_{m=-\infty}^{\infty} f(2m\pi), \tag{B.1}$$

which is the Poisson summation formula, and

$$\sum_{m=0}^{\infty}\int_{-\infty}^{\infty} f(u)\mathrm{e}^{-\mathrm{i}mu}\,\mathrm{d}u = \int\!\!\!\!\!\!-_{-\infty}^{\infty}\frac{f(u)}{1-\mathrm{e}^{-\mathrm{i}u}}\,\mathrm{d}u + \pi\sum_{m=-\infty}^{\infty} f(2m\pi) = \int\!\!\!\!\!\!-_{-\infty}^{\infty}\frac{f(u)}{1-\mathrm{e}^{-\mathrm{i}u}}\,\mathrm{d}u, \tag{B.2}$$

where the notation means that the contour passes below the poles of the integrand.

Appendix C. Schlömilch series. The quantity $\sigma(\psi)$ is defined by (3.6). An alternative representation is

$$\sigma(\psi) = -1 - \frac{2\mathrm{i}}{\pi}\left(C + \ln\frac{ks}{4\pi}\right) + \frac{2}{ks\sin\psi_0} + \sum_{\substack{m=-\infty\\ m\neq 0}}^{\infty}\left(\frac{2}{ks\sin\psi_m} + \frac{\mathrm{i}}{\pi|m|}\right), \tag{C.1}$$

where $C \approx 0{\cdot}5772$ is Euler's constant. The efficient computation of this series is discussed in [16].

Another important series is

$$S = \sum_{m=1}^{\infty}\mathrm{e}^{-\mathrm{i}m\beta s}H_0(kms).$$

To derive an alternative representation more convenient for computation we write

$$\begin{aligned} 2S - \sigma &= \sum_{m=1}^{\infty}(\mathrm{e}^{-\mathrm{i}m\beta s} - \mathrm{e}^{\mathrm{i}m\beta s})H_0(kms)\\ &= \sum_{m=-\infty}^{\infty}(\mathrm{e}^{-\mathrm{i}|m|\beta s} - \mathrm{e}^{-\mathrm{i}m\beta s})H_0(k|m|s) = \sum_{m=-\infty}^{\infty} f(2m\pi), \end{aligned}$$

where

$$f(u) = (\mathrm{e}^{-\mathrm{i}|u|\beta s/2\pi} - \mathrm{e}^{-\mathrm{i}u\beta s/2\pi})H_0(k|u|s/2\pi).$$

The Poisson summation formula (B.1) then gives

$$\begin{aligned} 2S - \sigma &= \frac{1}{2\pi}\sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty}(\mathrm{e}^{-\mathrm{i}|u|\beta s/2\pi} - \mathrm{e}^{-\mathrm{i}u\beta s/2\pi})H_0(k|u|s/2\pi)\mathrm{e}^{-\mathrm{i}mu}\,\mathrm{d}u\\ &= \frac{1}{s}\sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty}(\mathrm{e}^{-\mathrm{i}|v|\beta} - \mathrm{e}^{-\mathrm{i}v\beta})H_0(k|v|)\mathrm{e}^{-\mathrm{i}m2\pi v/s}\,\mathrm{d}v\\ &= \frac{1}{s}\sum_{m=-\infty}^{\infty}\int_{0}^{\infty}(\mathrm{e}^{-\mathrm{i}v\beta_{-m}} - \mathrm{e}^{\mathrm{i}v\beta_m})H_0(kv)\,\mathrm{d}v\\ &= \frac{4}{\pi ks}\left(\frac{\frac{1}{2}\pi - \psi}{\sin\psi} + \sum_{m=1}^{\infty}\left[\frac{\frac{1}{2}\pi - \psi_m}{\sin\psi_m} + \frac{\frac{1}{2}\pi - \psi_{-m}}{\sin\psi_{-m}}\right]\right), \end{aligned} \tag{C.2}$$

where we have used [10, equations 6.671(7), (11)]. Note that

$$\frac{\frac{1}{2}\pi-\psi_m}{\sin\psi_m}+\frac{\frac{1}{2}\pi-\psi_{-m}}{\sin\psi_{-m}}\sim\frac{\beta ks^2}{4m^2\pi^2}\left(\pi\mathrm{i}-2-2\ln\left(\frac{4m\pi}{ks}\right)\right)+O(m^{-3})\quad\text{as } m\to\infty.$$

As $\psi_p\to 0$ for some integer p,

$$S\sim 2(ks\sin\psi_p)^{-1}, \tag{C.3}$$

but as $\psi_p\to\pi$, S is bounded since the singularity in the sum in (C.2) exactly cancels that in σ.

Appendix D. Asymptotics of a sum. Consider the sum

$$S_p=\sum_{j=p+1}^{\infty}\frac{\mathrm{e}^{\mathrm{i}j\theta}}{j^{1/2}}.$$

Denote by Γ_p the contour which runs from $\mathrm{i}(p+\frac{1}{2})-\infty$ to $\mathrm{i}(p+\frac{1}{2})+\infty$ and is closed in the upper half-plane. Then, provided $0<\theta<2\pi$,

$$\begin{aligned}S_p&=\frac{1}{2\pi\mathrm{i}}\left\{-2\pi\mathrm{i}^{\frac{1}{2}}\int_{\Gamma_p}\frac{\mathrm{e}^{\theta t}\,\mathrm{d}t}{t^{\frac{1}{2}}(1-\mathrm{e}^{2\pi t})}\right\}=-\frac{1}{\mathrm{i}^{\frac{1}{2}}}\int_{\mathrm{i}(p+\frac{1}{2})-\infty}^{\mathrm{i}(p+\frac{1}{2})+\infty}\frac{\mathrm{e}^{\theta t}\,\mathrm{d}t}{t^{\frac{1}{2}}(1-\mathrm{e}^{2\pi t})}\\&=-\frac{\mathrm{e}^{\mathrm{i}\theta(p+\frac{1}{2})}}{\mathrm{i}^{\frac{1}{2}}}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\theta u}\,\mathrm{d}u}{[u+\mathrm{i}(p+\frac{1}{2})]^{\frac{1}{2}}(1+\mathrm{e}^{2\pi u})}\\&\sim\frac{\mathrm{i}\mathrm{e}^{\mathrm{i}\theta(p+\frac{1}{2})}}{p^{\frac{1}{2}}}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\theta u}\,\mathrm{d}u}{1+\mathrm{e}^{2\pi u}}=\frac{\mathrm{i}\mathrm{e}^{\mathrm{i}\theta(p+\frac{1}{2})}}{2p^{\frac{1}{2}}\sin\frac{1}{2}\theta}=\frac{-p^{-\frac{1}{2}}\mathrm{e}^{\mathrm{i}\theta p}}{1-\mathrm{e}^{-\frac{\mathrm{i}\theta}{2}}}\quad\text{as } p\to\infty.\end{aligned}$$

If $2m\pi<\theta<2(m+1)\pi$,

$$\sum_{j=p+1}^{\infty}\frac{\mathrm{e}^{\mathrm{i}j\theta}}{j^{\frac{1}{2}}}=\sum_{j=p+1}^{\infty}\frac{\mathrm{e}^{\mathrm{i}j(\theta-2m\pi)}}{j^{\frac{1}{2}}}\sim\frac{-p^{-\frac{1}{2}}\mathrm{e}^{\mathrm{i}\theta p}}{1-(-1)^m\mathrm{e}^{-\frac{\mathrm{i}\theta}{2}}}\quad\text{as } p\to\infty. \tag{D.1}$$

Appendix E. Inversion of symmetric Toeplitz matrices. Each of (4.2), (4.3), and (5.8) is of the form

$$\sum_{m=0}^{\infty}T_{nm}X_m=R_n,\quad n=0,1,2,\dots, \tag{E.1}$$

i.e., $\mathbf{T}X=R$, where $\mathbf{T}$ is a Toeplitz matrix whose elements are given by $T_{nm}=T_{mn}=t_{n-m}$ with

$$t_m=\begin{cases}1, & m=0,\\ -f_0H_0(k|m|s), & \text{otherwise.}\end{cases}$$

The matrix $\mathbf{T}$ can be inverted using the discrete Wiener–Hopf technique; the symmetry of $\mathbf{T}$ is not required for this approach but it simplifies the analysis. Here we follow the method as described in [27]. No fully rigorous theory appears to exist which includes the particular matrix $\mathbf{T}$ that occurs in our problem. If the elements of the matrix satisfied $\sum|t_m|<\infty$, or the function $\sum t_m\exp(\mathrm{i}m\theta)<\infty$ for all $\theta\in(-\pi,\pi)$,

then we could appeal to the general theory given in [4], [15, section 13]. It is assumed throughout this appendix that we are dealing with a nonresonant case.

First we rewrite (E.1) as

$$\sum_{m=-\infty}^{\infty} t_{n-m} X_m^+ = R_n^+ + R_n^-, \quad n = 0, \pm 1, \pm 2, \dots ,$$

where X_n^+ and R_n^+ are equal to X_n and R_n, respectively, when $n \geq 0$, and are zero if $n < 0$ and $R_n^- = 0$ for $n \geq 0$, but are otherwise unknown. If we multiply the nth equation by z^n and then sum over n, we get, after writing $n = m + \nu$ on the left-hand side,

$$\sum_{\nu=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} t_\nu X_m^+ z^{m+\nu} = \sum_{n=-\infty}^{\infty} R_n^+ z^n + \sum_{n=-\infty}^{\infty} R_n^- z^n,$$

which can be written

$$K(z) X_+(z) = R_+(z) + R_-(z), \tag{E.2}$$

where $X_+(z) = \sum_{n=-\infty}^{\infty} X_n^+ z^n = \sum_{n=0}^{\infty} X_n z^n$, $R_+(z) = \sum_{n=0}^{\infty} R_n z^n$, $R_-(z) = \sum_{n=-\infty}^{-1} R_n^- z^n$, and

$$K(z) = \sum_{\nu=-\infty}^{\infty} t_\nu z^\nu = 1 - f_0 \sum_{\substack{\nu=-\infty \\ \nu \neq 0}}^{\infty} H_0(ks|\nu|) z^\nu.$$

Note that

$$K(\mathrm{e}^{\pm \mathrm{i} ks \cos\theta}) = 1 - f_0 \sigma(\theta), \tag{E.3}$$

where $\sigma(\theta)$ is given by either (3.6) or (C.1), and, in particular, $K(\mathrm{e}^{\pm \mathrm{i}\beta s}) = -f_0 \mathcal{K}$, where $\mathcal{K}$ is defined by (3.7). Here $X_+(z)$ is analytic in some disk centered on the origin, and it is reasonable to assume that the radius of convergence is greater than or equal to one. Similar remarks pertain to $R_+(z)$. On the other hand, we can assume that $R_-(z)$ is analytic in the region exterior to the unit disk. One device that can be used in scattering problems is to let the wavenumber k have a small positive imaginary part, which is equivalent to allowing for a small amount of dissipation in the acoustic medium. This will ensure that the plus functions are analytic for $|z| < \rho_2$ and the minus functions are analytic for $|z| > \rho_1$, with $\rho_1 < 1 < \rho_2$. The solution is then obtained by letting the imaginary part of k tend to zero at the end of the calculation. This also takes care of the fact that in our case $K(z)$ actually has singularities on the unit circle, namely inverse square-root branch points at $z = \exp(\pm \mathrm{i} ks)$.

The solution method is based on a factorization $K(z) = K_+(z) K_-(z)$, where $K_+(z)$ (resp., $K_-(z)$) is analytic and nonzero inside (resp., outside) and on $|z| = 1$. Given such a factorization we have $\ln K(z) = \ln K_+(z) + \ln K_-(z) = q_+(z) + q_-(z)$, say, where $q_+(z)$ (resp., $q_-(z)$) is analytic inside (resp., outside) and on $|z| = 1$. From Cauchy's integral formula, writing $q = q_+ + q_-$,

$$q_-(z) = q_-(\infty) - \frac{1}{2\pi \mathrm{i}} \oint_{|\zeta|=1} \frac{q(\zeta)}{\zeta - z} \, \mathrm{d}\zeta, \quad |z| > 1.$$

Here we must assume that it is possible to choose a single-valued branch of $\ln K(z)$ in some neighborhood of the unit circle. Note that $K(z) = K(1/z)$, so we can normalize the factorization by requiring that $K_+(z) = K_-(1/z)$, in which case $q_+(z) = q_-(1/z)$. It follows that

$$q_+(z) = q_+(0) - \frac{1}{2\pi \mathrm{i}} \oint_{|\zeta|=1} \frac{zq(\zeta)}{z\zeta - 1}\,\mathrm{d}\zeta, \quad |z| < 1.$$

The required factorization of $K(z)$ (which is unique) is thus given by

$$\begin{aligned} K_+(z) &= \frac{1}{\lambda_0} \exp\left(\frac{1}{2\pi \mathrm{i}} \oint_{|\zeta|=1} \frac{z \ln K(\zeta)}{1 - z\zeta}\,\mathrm{d}\zeta\right), \quad |z| < 1, \\ K_-(z) &= \frac{1}{\lambda_0} \exp\left(\frac{1}{2\pi \mathrm{i}} \oint_{|\zeta|=1} \frac{\ln K(\zeta)}{z - \zeta}\,\mathrm{d}\zeta\right), \quad |z| > 1, \end{aligned} \tag{E.4}$$

where $\lambda_0 = \mathrm{e}^{-q_+(0)}$. However, from Cauchy's integral formula

$$q_+(0) = \frac{1}{2\pi \mathrm{i}} \oint_{|z|=1} \frac{q_+(z)}{z}\,\mathrm{d}z = \frac{1}{4\pi \mathrm{i}} \oint_{|z|=1} \frac{q_+(z) + q_-(z)}{z}\,\mathrm{d}z,$$

and thus we have

$$\lambda_0 = \exp\left(-\frac{1}{4\pi \mathrm{i}} \oint_{|z|=1} \frac{\ln K(z)}{z}\,\mathrm{d}z\right).$$

With this factorization we can rearrange (E.2) as follows:

$$K_+ X_+ - \left(\frac{R_+}{K_-}\right)_+ = \left(\frac{R_+}{K_-}\right)_- + \frac{R_-}{K_-}, \tag{E.5}$$

in which we have further separated the function $R_+(z)/K_-(z)$ into the sum of a function analytic inside $|z| = 1$ and one analytic outside this circle. Liouville's theorem then implies that both sides must equal a constant. The sum-split of R_+/K_- is performed so that $(R_+/K_-)_-$ tends to zero as $z \to \infty$. We also have $R_-/K_- \to 0$ since $K_-(z)$ tends to a nonzero constant in this limit, and so both sides of (E.5) must in fact be zero.

We have thus established that $X_+ = (R_+/K_-)_+/K_+$ and hence that

$$X_m^+ = \frac{1}{2\pi \mathrm{i}} \oint_{|z|=1} \frac{(R_+/K_-)_+(z)}{K_+(z)} z^{-m-1}\,\mathrm{d}z.$$

Now $(R_+/K_-)_+(z) = \sum_{n=0}^{\infty} R_n (z^n/K_-)_+(z)$ and

$$\left(\frac{z^n}{K_-}\right)_+(z) = \sum_{j=0}^{\infty} a_j^{(n)} z^j \quad \text{with } a_j^{(n)} = \frac{1}{2\pi \mathrm{i}} \oint_{|z|=1} \frac{z^{n-j-1}}{K_-(z)}\,\mathrm{d}z,$$

the final integral being zero if $j > n$, since then the integrand is regular and nonzero for $|z| > 1$ and decays at infinity faster than $1/z$. Thus we have

$$X_m^+ = \frac{1}{2\pi \mathrm{i}} \sum_{n=0}^{\infty} R_n \oint_{|z|=1} \sum_{j=0}^{n} a_j^{(n)} \frac{z^{j-m-1}}{K_+(z)}\,\mathrm{d}z. \tag{E.6}$$

Now if we define $\lambda_\mu = a_{n-\mu}^{(n)}$ (with $\lambda_\mu = 0$ if $\mu < 0$), then

$$\lambda_\mu = \frac{1}{2\pi \mathrm{i}} \oint_{|z|=1} \frac{z^{\mu-1}}{K_-(z)}\,\mathrm{d}z,$$

and so λ_μ, $\mu = 0, 1, 2, \ldots$, are the coefficients in the Laurent series for $1/K_-(z)$, i.e., $[K_-(z)]^{-1} = \sum_{\mu=0}^{\infty} \lambda_\mu z^{-\mu}$, from which

$$\frac{1}{K_+(z)} = \frac{1}{K_-(1/z)} = \sum_{\mu=0}^{\infty} \lambda_\mu z^\mu. \tag{E.7}$$

Note that $\lambda_0 = 1/K_+(0)$, in agreement with (E.4). The coefficients λ_μ can be calculated without knowledge of the functions $K_\pm$ since

$$\lambda_\mu = \frac{1}{\mu!}\frac{\mathrm{d}^\mu}{\mathrm{d}z^\mu}\left[\frac{1}{K_+(z)}\right]_{z=0},$$

and the right-hand side can be evaluated from (E.4) in terms of the weakly singular integrals

$$\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}m\theta} \ln[K(\mathrm{e}^{\mathrm{i}\theta})]\,\mathrm{d}\theta, \quad m = 0, 1, \ldots .$$

In order to compute K on the unit circle we use (E.3) and (C.1).

The presence of square-root singularities in $K(z)$ at $z = \exp(\pm \mathrm{i}ks)$ implies, after letting the imaginary part of k tend to zero, a singularity in $K_+(z)$ at $z = \exp(-\mathrm{i}ks)$, i.e.,

$$[K_+(\mathrm{e}^{-\mathrm{i}ks})]^{-1} = \sum_{\mu=0}^{\infty} \lambda_\mu \mathrm{e}^{-\mathrm{i}\mu ks} = 0. \tag{E.8}$$

The function $[K_+(z)]^{-1}$ is smooth everywhere on the unit circle except at the point $z = \exp(-\mathrm{i}ks)$, where its derivative has a square-root singularity. We thus expect (see [25, p. 441]) that

$$\lambda_\mu = O(\mu^{-3/2}) \quad \text{as } \mu \to \infty.$$

If (E.7) is substituted into (E.6), we get

$$X_m^+ = \frac{1}{2\pi \mathrm{i}} \sum_{n=0}^{\infty} R_n \oint_{|z|=1} \sum_{j=0}^{n} \lambda_{n-j} z^{j-m-1} \sum_{\mu=0}^{\infty} \lambda_\mu z^\mu \,\mathrm{d}z = \sum_{n=0}^{\infty} R_n \sum_{j=0}^{\min(n,m)} \lambda_{n-j}\lambda_{m-j}. \tag{E.9}$$

We have thus shown that the elements of the (symmetric) inverse matrix $\mathbf{T}^{-1}$ (written T_{mn}^{-1}, $m \geq 0$, $n \geq 0$) are given by

$$T_{mn}^{-1} = \sum_{j=0}^{\min(m,n)} \lambda_{n-j}\lambda_{m-j}.$$

The final expression in (E.9) can be rearranged to give

$$X_m = \sum_{p=0}^{\infty} \sum_{q=0}^{m} \lambda_p \lambda_q R_{m+p-q}. \tag{E.10}$$

The method of Hills and Karp [12]. For the particular case of (4.2), we have $R_n = f_0 I_n$ and so

$$R_+(z) = \sum_{n=0}^{\infty} R_n z^n = \frac{f_0}{1 - z\mathrm{e}^{\mathrm{i}\beta s}}.$$

This is the analytic continuation of the series into the entire complex plane except the point $z = \exp(-\mathrm{i}\beta s)$. Note that, assuming k to have a positive imaginary part, the singularity at $z = \exp(-\mathrm{i}\beta s)$ is exterior to the unit circle. With $A_+(z) = \sum_{n=0}^{\infty} A_n z^n$ we then have

$$K(z)A_+(z) = \frac{f_0}{1 - z\mathrm{e}^{\mathrm{i}\beta s}} + R_-(z). \tag{E.11}$$

Equation (E.11) is [12, (3.1-3)]. It is a single equation for two unknown functions, namely $A_+(z)$ and $R_-(z)$.

The split into plus and minus functions can now be carried out more simply than in the general case. We have

$$K_+(z)A_+(z) - \frac{f_0}{K_-(\mathrm{e}^{-\mathrm{i}\beta s})(1 - z\mathrm{e}^{\mathrm{i}\beta s})} = \frac{f_0}{(1 - z\mathrm{e}^{\mathrm{i}\beta s})}\left(\frac{1}{K_-(z)} - \frac{1}{K_-(\mathrm{e}^{-\mathrm{i}\beta s})}\right) + \frac{R_-(z)}{K_-(z)},$$

and Liouville's theorem shows that both sides are zero so that

$$A_+(z) = \frac{f_0}{K_+(z)K_-(\mathrm{e}^{-\mathrm{i}\beta s})(1 - z\mathrm{e}^{\mathrm{i}\beta s})}.$$

It follows that

$$A_m = \frac{f_0 I_m}{K_-(\mathrm{e}^{-\mathrm{i}\beta s})} \sum_{q=0}^{m} \lambda_q I_{-q} = f_0 I_m \sum_{p=0}^{\infty} \lambda_p I_p \sum_{q=0}^{m} \lambda_q I_{-q}$$

in agreement with (E.10).

Now, since $A_n = I_n B_0 + C_n$ and $B_0 = -1/\mathcal{K} = f_0/K(\exp(\pm\mathrm{i}\beta s))$,

$$\sum_{n=0}^{\infty} C_n z^n \equiv C_+(z) = \frac{f_0}{K_-(\mathrm{e}^{-\mathrm{i}\beta s})(1 - z\mathrm{e}^{\mathrm{i}\beta s})}\left(\frac{1}{K_+(z)} - \frac{1}{K_+(\mathrm{e}^{-\mathrm{i}\beta s})}\right).$$

The coefficients C_n decay at the same rate as λ_n, i.e.,

$$C_n = O(n^{-\frac{3}{2}}) \quad \text{as } n \to \infty, \tag{E.12}$$

and for exactly the same reasons. This decay rate for the edge effects was noted in [20, equation 76]. From (E.8), we have

$$C_+(\mathrm{e}^{-\mathrm{i}ks}) = -\frac{f_0}{K(\mathrm{e}^{-\mathrm{i}\beta s})(1 - \mathrm{e}^{\mathrm{i}(\beta - k)s})},$$

or equivalently,

$$\sum_{n=0}^{\infty} C_n \mathrm{e}^{-\mathrm{i}nks} = \frac{1}{\mathcal{K}(1 - \mathrm{e}^{\mathrm{i}ks(\cos\psi - 1)})}. \tag{E.13}$$

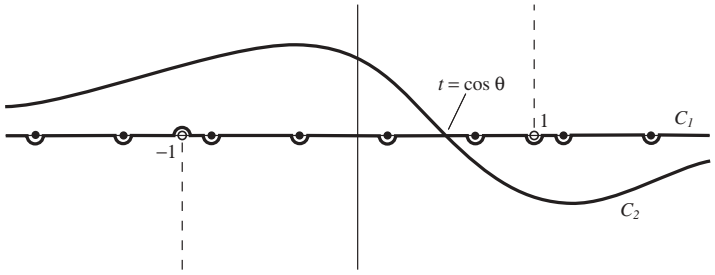


FIG. F.1. *Contours C_1 and C_2. The dashed lines are branch cuts for the function $\gamma(t)$, and the solid circles are zeros of $1 - \mathrm{e}^{\mathrm{i}s(\beta - kt)}$ (i.e., $t = \beta_m/k$).*

Appendix F. Asymptotics of an integral. Consider the integral

$$I(kr) = \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-k\gamma(t)y}\mathrm{e}^{\mathrm{i}kxt} f(t)}{\gamma(t)(1 - \mathrm{e}^{\mathrm{i}s(\beta - kt)})}\,\mathrm{d}t = \int_{C_1} \frac{f(t)\mathrm{e}^{krg(t)}}{\gamma(t)(1 - \mathrm{e}^{\mathrm{i}s(\beta - kt)})}\,\mathrm{d}t,$$

where $y = r\sin\theta > 0$,

$$g(t) = -\gamma(t)\sin\theta + \mathrm{i}t\cos\theta,$$

and C_1 is the contour shown in Figure F.1. Here the branch of $\gamma(t)$ (defined for real t by (A.2)) is indicated by the branch cuts shown in the figure, and f is assumed regular throughout the complex t-plane. We will assume that $|\beta_m/k| \neq 1$ holds for all m so that none of the poles of the integrand coincide with the branch points of γ. The function g has one simple saddle point in the complex t-plane at $t = \cos\theta$ and

$$g(\cos\theta) = \mathrm{i}, \qquad g''(\cos\theta) = -\mathrm{i}/\sin^2\theta.$$

In order to derive the asymptotics of I for large kr we need to deform the contour C_1 into the path of steepest descent. This is the curve on which $\operatorname{Imag} g = 1$, which passes through the saddle point, making an angle $-\pi/4$ with the positive real t-axis. This curve crosses the real axis again at $t = 1/\cos\theta$. In deforming the contour, we pick up contributions from the poles on the real axis over which we pass. Only those poles between -1 and 1 give any contribution in the limit as $kr \to \infty$, the others leading to exponentially small terms. Hence we can deform the contour back down to the real axis to produce C_2 without affecting the asymptotics.

Hence, as $kr \to \infty$,

$$\begin{aligned} I(kr) &\sim \int_{C_2} \frac{f(t)\mathrm{e}^{krg(t)}}{\gamma(t)(1 - \mathrm{e}^{\mathrm{i}s(\beta - kt)})}\,\mathrm{d}t + 2\pi\mathrm{i} \sum_{\substack{m\in\mathcal{M} \\ \beta_m < k\cos\theta}} \frac{f(\beta_m/k)\mathrm{e}^{\mathrm{i}kr\cos(\theta - \psi_m)}}{ks\sin\psi_m} \\ &\sim \frac{f(\cos\theta)\mathrm{e}^{krg(\cos\theta) - \frac{1}{4}\pi\mathrm{i}}}{-\mathrm{i}\sin\theta(1 - \mathrm{e}^{\mathrm{i}s(\beta - k\cos\theta)})}\sqrt{\frac{2\pi}{kr|g''(\cos\theta)|}} \\ &\quad + 2\pi\mathrm{i} \sum_{\substack{m\in\mathcal{M} \\ \cos\psi_m < \cos\theta}} \frac{f(\beta_m/k)\mathrm{e}^{\mathrm{i}kr\cos(\theta - \psi_m)}}{ks\sin\psi_m} \\ &= \frac{\mathrm{i}f(\cos\theta)\mathrm{e}^{\mathrm{i}(kr - \frac{1}{4}\pi)}}{1 - \mathrm{e}^{\mathrm{i}ks(\cos\psi - \cos\theta)}}\sqrt{\frac{2\pi}{kr}} + 2\pi\mathrm{i} \sum_{\substack{m\in\mathcal{M} \\ \psi_m > \theta}} \frac{f(\beta_m/k)\mathrm{e}^{\mathrm{i}kr\cos(\theta - \psi_m)}}{ks\sin\psi_m}, \end{aligned} \tag{F.1}$$

where we have used $\gamma(\cos\theta) = -\mathrm{i}\sin\theta$ and the asymptotics of the integral along the steepest descent contour are given, for example, by [7, equation 4.2.1b]. To obtain the asymptotics valid for $y < 0$ we simply replace θ by $-\theta$ in (F.1). It is implicit in the above that the saddle point of g does not coincide with any of the zeros of $1 - \mathrm{e}^{\mathrm{i}s(\beta-kt)}$. In other words, we have assumed that $|\theta| \neq \psi_m$ for any integer m.

Uniform asymptotics valid as $\psi_p \to |\theta|$ can be obtained; see [7, section 4.4], for example. A lengthy calculation shows that we must add a term

$$\tilde{I} = \frac{\sqrt{\pi}\mathrm{e}^{\mathrm{i}(kr-\frac{1}{4}\pi)}}{ks\zeta_p \sin\psi_p} \operatorname{sgn}(|\theta| - \psi_p) f(\beta_p/k) \left(1 + 2\mathrm{i}\zeta_p \mathrm{e}^{-\mathrm{i}\zeta_p^2} F(\zeta_p)\right) \tag{F.2}$$

to the right-hand side of (F.1). Here $\zeta_p = \sqrt{2kr}\sin(\frac{1}{2}||\theta| - \psi_p|)$, and

$$F(v) = \int_v^\infty \mathrm{e}^{\mathrm{i}u^2}\,\mathrm{d}u \quad \left(0 < \arg u < \frac{1}{2}\pi \text{ as } u \to \infty\right) \tag{F.3}$$

is a Fresnel integral. Since (see [3, p. 67])

$$F(v) \sim \frac{\mathrm{i}}{2v}\mathrm{e}^{\mathrm{i}v^2}\left(1 + \sum_{n=1}^{\infty} \frac{(2n-1)!!}{(2\mathrm{i}v^2)^n}\right) \quad \text{as } v \to \infty, \quad -\frac{1}{2}\pi < \arg v < \pi,$$

we have

$$\tilde{I} \sim -\frac{\sqrt{\pi}\mathrm{e}^{\mathrm{i}(kr-\frac{1}{4}\pi)}}{2\mathrm{i}ks\zeta_p^3 \sin\psi_p} \operatorname{sgn}(|\theta| - \psi_p) f(\beta_p/k) \quad \text{as } \zeta_p \to \infty.$$

Appendix G. The quantities $\mathcal{G}_n$. From (5.7), we have

$$\mathcal{G}_n = \frac{1}{2\pi a}\int_{-a}^{a}\int_{-a}^{a} \frac{H_0(k|x - \xi - ns|)}{\sqrt{a^2 - \xi^2}}\,\mathrm{d}\xi\,\mathrm{d}x.$$

For $n = 0$, with $ka \ll 1$, we can approximate this by

$$\begin{aligned}\mathcal{G}_0 &\approx \frac{1}{2\pi^2 a}\int_{-a}^{a}\int_{-a}^{a} \frac{\pi + 2\mathrm{i}(C + \ln\frac{1}{2}k|x-\xi|)}{\sqrt{a^2-\xi^2}}\,\mathrm{d}\xi\,\mathrm{d}x \\ &= 1 + \frac{2\mathrm{i}C}{\pi} + \frac{\mathrm{i}}{\pi^2 a}\int_{-a}^{a}\int_{-a}^{a} \frac{\ln\frac{1}{2}k|x-\xi|}{\sqrt{a^2-\xi^2}}\,\mathrm{d}\xi\,\mathrm{d}x = 1 + \frac{2\mathrm{i}}{\pi}\left(C + \ln\frac{1}{4}ka\right).\end{aligned} \tag{G.1}$$

In [22] the $O((ka)^2)$ terms in $\mathcal{G}_0$ are also evaluated, but this seems to be inconsistent with the level of approximation being used. For $n \neq 0$ we use Neumann's addition theorem [1, equation 9.1.75], which shows that

$$H_0(k|n|s \pm k(x-\xi)) = \sum_{m=-\infty}^{\infty} H_{\pm m}(k|n|s) J_m(k(x-\xi)) \approx H_0(k|n|s)$$

since $ka \ll 1$. Hence

$$\mathcal{G}_n \approx H_0(k|n|s). \tag{G.2}$$

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, New York, 1965.

[2] C. Aristégui and Y. C. Angel, *New results for isotropic point scatterers: Foldy revisited*, Wave Motion, 36 (2002), pp. 383–399.

[3] J. J. Bowman, T. B. A. Senior, and P. L. E. Uslenghi, eds., *Electromagnetic and Acoustic Scattering by Simple Shapes*, revised ed., Hemisphere, New York, 1987.

[4] A. Calderón, F. Spitzer, and H. Widom, *Inversion of Toeplitz matrices*, Illinois J. Math., 3 (1959), pp. 490–498.

[5] F. Capolino, M. Albani, S. Maci, and L. B. Felsen, *Frequency-domain Green's function for a planar periodic semi-infinite phased array—Part* I*: Truncated Floquet wave formulation*, IEEE Trans. Antennas and Propagation, 48 (2000), pp. 67–74.

[6] F. Capolino, M. Albani, S. Maci, and L. B. Felsen, *Frequency-domain Green's function for a planar periodic semi-infinite phased array—Part* II*: Diffracted wave phenomenology*, IEEE Trans. Antennas and Propagation, 48 (2000), pp. 75–85.

[7] L. B. Felsen and N. Marcuvitz, *Radiation and Scattering of Waves*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[8] L. L. Foldy, *The multiple scattering of waves* I*. General theory of isotropic scattering by randomly distributed scatterers*, Phys. Rev., 67 (1945), pp. 107–119.

[9] I. M. Gel'fand and G. E. Shilov, *Generalized Functions. Volume* 1*. Properties and Operations*, Academic Press, New York, 1964.

[10] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products*, 4th ed., Academic Press, New York, 1980.

[11] N. L. Hills, *Semi-infinite diffraction gratings.* II*. Inward resonance*, Comm. Pure Appl. Math., 18 (1965), pp. 389–395.

[12] N. L. Hills and S. N. Karp, *Semi-infinite diffraction gratings*—I, Comm. Pure Appl. Math., 18 (1965), pp. 203–233.

[13] A. Ishimaru, R. J. Coe, G. E. Miller, and W. P. Geren, *Finite periodic structure approach to large scanning array problems*, IEEE Trans. Antennas and Propagation, 33 (1985), pp. 1213–1220.

[14] R. Kleinman and B. Vainberg, *Full low-frequency asymptotic expansion for second-order elliptic equations in two dimensions*, Math. Methods Appl. Sci., 17 (1994), pp. 989–1004.

[15] M. G. Krein, *Integral equations on a half-line with kernel depending upon the difference of the arguments*, Amer. Math. Soc. Transl. Ser. 2, 22 (1962), pp. 163–288.

[16] C. M. Linton, *The Green's function for the two-dimensional Helmholtz equation in periodic domains*, J. Engrg. Math., 33 (1998), pp. 377–402.

[17] S. Maci, F. Capolino, and L. B. Felsen, *Three-dimensional Green's function for planar rectangular phased dipole arrays*, Wave Motion, 34 (2001), pp. 263–279.

[18] R. F. Millar, *Plane wave spectra in grating theory.* III*. Scattering by a semiinfinite grating of identical cylinders*, Canad. J. Phys., 42 (1964), pp. 1149–1184.

[19] R. F. Millar, *On a non-linear integral equation occurring in diffraction theory*, Proc. Cambridge Philos. Soc., 62 (1966), pp. 249–261.

[20] R. F. Millar, *Plane wave spectra in grating theory.* V*. Scattering by a semi-infinite grating of isotropic scatterers*, Canad. J. Phys., 44 (1966), pp. 2839–2874.

[21] A. Neto, S. Maci, G. Vecchi, and M. Sabbadini, *A truncated Floquet wave diffraction method for the full wave analysis of large phased arrays—Part* I*: Basic principles and* 2-D *cases*, IEEE Trans. Antennas and Propagation, 48 (2000), pp. 594–600.

[22] M. Nishimoto and H. Ikuno, *Analysis of electromagnetic wave diffraction by a semi-infinite strip grating and evaluation of end-effects*, Progress in Electromagnetics Research, 23 (1999), pp. 39–58.

[23] M. Nishimoto and H. Ikuno, *Space-wavenumber analysis of field scattered from a semi-infinite strip grating*, Electrical Engineering in Japan, 132 (2000), pp. 1–8.

[24] A. J. Roscoe and R. A. Perrott, *Large finite array analysis using infinite array data*, IEEE Trans. Antennas and Propagation, 42 (1994), pp. 983–992.

[25] E. C. Titchmarsh, *The Theory of Functions*, 2nd ed., Oxford University Press, London, 1939.

[26] B. Tomasic and A. Hessel, *Analysis of finite arrays—A new approach*, IEEE Trans. Antennas and Propagation, 47 (1999), pp. 555–565.

[27] W. Wasylkiwskyj, *Mutual coupling effects in semi-infinite arrays*, IEEE Trans. Antennas and Propagation, 21 (1973), pp. 277–285.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 1057–1076

SELECTIVE ACOUSTIC FOCUSING USING TIME-HARMONIC REVERSAL MIRRORS*

CHRISTOPHE HAZARD† AND KARIM RAMDANI‡

Abstract. A mathematical study of the focusing properties of acoustic fields obtained by a time-reversal process is presented. The case of time-harmonic waves propagating in a nondissipative medium containing sound-soft obstacles is considered. In this context, the so-called D.O.R.T. method (decomposition of the time-reversal operator in French) was recently proposed to achieve selective focusing by computing the eigenelements of the *time-reversal operator*. The present paper describes a justification of this technique in the framework of the far field model, i.e., for an ideal time-reversal mirror able to reverse the far field of a scattered wave. Both cases of closed and open mirrors, that is, surrounding completely or partially the scatterers, are dealt with. Selective focusing properties are established by an asymptotic analysis for small and distant obstacles.

Key words. acoustic scattering, time-reversal, far field operator, small obstacles

AMS subject classifications. 35B40, 35P25, 45A05, 74J20

DOI. 10.1137/S0036139903428732

1. Introduction.

Acoustic time-reversal has known in the last few years a significant growth of interest, covering a large number of applications (medical imaging, nondestructive testing, etc.). The main idea of this phenomenon is to take advantage of the reversibility of the wave equation in a nondissipative unknown medium to backpropagate signals to the sources that emitted them. Today, the physical literature (cf. [9] for more details) on this topic is quite rich. Meanwhile, some mathematical works started to deal with different aspects of time-reversal phenomena: see, for instance, [2, 4] for time-reversal in the time domain, [14] for time-reversal in the frequency domain, and [15] for time-reversal in random media.

In this work, we present a mathematical analysis of the so-called D.O.R.T. method (decomposition of the time-reversal operator in French), detailed in [16] to achieve selective focusing on diffracting obstacles using time-reversal mirrors (TRM) which are able to emit and receive acoustic waves. In the frequency domain, this method can be described as follows: the TRM first emits an acoustic wave in a homogeneous and nondissipative medium containing some unknown obstacles, and then measures the diffracted field. The measured field is then conjugated (reversing time amounts to a conjugation when the time dependence is of the form $e^{i\omega t}$), and re-emitted. The time-reversal operator T is the operator obtained by iterating this procedure twice. The experimental results obtained in [16] show that the number of nonzero (or significant) eigenvalues of T is exactly the number of obstacles contained in the propagation medium. Furthermore, the corresponding eigenvectors generate incident waves that focus selectively on the obstacles. Our aim here is to present a mathematical justification of these results related to selective focusing using TRMs; we will show that these results are not true in general, but do hold for small and distant obstacles with distinct reflectivities.

*Received by the editors May 26, 2003; accepted for publication (in revised form) September 10, 2003; published electronically April 14, 2004.
http://www.siam.org/journals/siap/64-3/42873.html
†Laboratoire SMP, CNRS / ENSTA, 32 Boulevard Victor, 75739 Paris Cedex 15, France (hazard@ensta.fr).
‡INRIA (Projet CORIDA) and Institut Elie Cartan, University of Nancy I, POB 239, Vandœuvre-les-Nancy, 54506 France (ramdani@loria.fr).

Let us point out that the inverse problem which consists of recovering the location of the obstacles from the scattering data is not dealt with in this paper. We are mainly concerned with qualitative properties of the eigenvectors of the time-reversal operator. The eigenvectors corresponding to significant eigenvalues span some kind of *relevant subspace* in the sense that they contain nearly all the information about the obstacles which can be extracted from the time-reversal operator. The others span the so-called *noise subspace*, which represents some kind of quasi-null space. This point of view meets the basics of the so-called MUSIC algorithm used in signal processing and imaging, and its generalization, the linear sampling method (LSM), used in inverse scattering (for a short presentation of these methods, see, for instance, [3], and for more details, cf. [12]). These methods answer the inverse problem by using a convenient characterization of the relevant subspace. In the context of scattering, the link between the scattering data and the unknown locations of the obstacles is made by means of point sources: if the radiated field produced by a given point source has a nonzero component in the relevant subspace, the point belongs to one scatterer, otherwise it is outside. But as mentioned in [3], both MUSIC and LSM use the noise subspace: the question of recovering the geometric information directly from the relevant subspace remains open.

The paper is organized as follows. We first deal with a TRM which entirely surrounds the obstacles. In section 2, we describe the mathematical model used to analyze time-reversal phenomena in the framework of time harmonic waves in the far field model, i.e., for an ideal TRM able to reverse the asymptotic behavior at large distance of a scattered wave. This will in particular lead us to express the time-reversal operator by means of the far field operator, well known in scattering theory. Section 3 recalls some results obtained in [14], concerning the global focusing properties of the eigenvectors of the time-reversal operator. The main result of the paper, which concerns selective focusing, is given in section 4. It provides a mathematical justification of the D.O.R.T. method for the problem of scattering by several small and distant obstacles. In section 5, we generalize the results obtained in the previous sections to the case of open mirrors (i.e., mirrors which do not completely surround the scatterers). The main ingredient for the proof of our main result is formula (4.2), which provides the asymptotic behavior of the scattering amplitude for the diffraction by many small obstacles. This formula, which is of independent interest, is proved in the appendix.

2. Mathematical setting of the problem and definition of the time-reversal operator. Consider a TRM completely surrounding a collection of sound-soft obstacles, located in a homogeneous medium of celerity c. During the emission step, the TRM illuminates the obstacles with an incident wave u_I which is supposed to be a Herglotz wave. Such waves are superpositions of planes waves $u_I^\alpha(x) = \exp(ik\alpha\cdot x)$ of direction $\alpha \in S^2$ (S^2 denotes the unit sphere in $\mathbb{R}^3$, $k = \omega/c$ is the wavenumber, and ω is the frequency). More precisely, given a directional distribution $f \in L^2(S^2)$, we suppose that the incident field emitted by the TRM has the form

$$u_I(x) = \int_{S^2} f(\alpha)\, u_I^\alpha(x)\, d\alpha = \int_{S^2} f(\alpha)\, e^{ik\alpha\cdot x}\, d\alpha. \tag{2.1}$$

We assume that the TRM is located far enough from the obstacles, so that its influence on the diffracted field can be neglected. Moreover, the TRM is supposed to measure the far field corresponding to the diffracted field.

Let Ω denote the propagation domain located outside the obstacles and let ν be the outgoing normal to Ω on its boundary $\Gamma = \partial\Omega$. When illuminated by the incident plane wave $u_I^\alpha(x) = e^{ik\alpha\cdot x}$ of direction $\alpha \in S^2$, the obstacles generate the diffracted field u_D^α that solves the classical Dirichlet exterior problem:

$$\left\{\begin{array}{rcll} \Delta u_D^\alpha + k^2 u_D^\alpha &=& 0, & (\Omega)\\ u_D^\alpha &=& -u_I^\alpha, & (\Gamma)\\ \lim\limits_{R\longrightarrow+\infty}\displaystyle\int_{S_R}\left|\frac{\partial u_D^\alpha}{\partial\nu} - iku_D^\alpha\right|^2 dx &=& 0, & \end{array}\right.$$

where S_R is the sphere $\{x \in \mathbb{R}^3;\ \|x\| = R\}$ and where $\partial u_D^\alpha/\partial\nu$ denotes the radial derivative of u_D^α on S_R.

It is well known (cf. [7]) that the far field asymptotics of the diffracted field in a given direction $\beta \in S^2$ is given by the formula

$$u_D^\alpha(\beta\|x\|) = \frac{e^{ik\|x\|}}{\|x\|}A(\alpha,\beta) + O(\|x\|^{-2}),$$

where the bound $O(\|x\|^{-2})$ is uniform for all $\beta \in S^2$, and where $A(\alpha,\beta)$ is known as the *scattering amplitude*. This function satisfies some remarkable properties (cf. [7]), which are summarized in the following.

PROPOSITION 2.1. *The scattering amplitude $A(\cdot,\cdot)$ is given by the formula*

$$A(\alpha,\beta) = \frac{1}{4\pi}\int_\Gamma \frac{\partial u_T^\alpha}{\partial\nu}(y)\,\overline{u_I^\beta(y)}\,d\Gamma_y, \tag{2.2}$$

where $u_T^\alpha = u_I^\alpha + u_D^\alpha$ denotes the total field associated with the incident field u_I^α. Furthermore, $A(\cdot,\cdot)$ defines an analytic function on $S^2\times S^2$ and satisfies the reciprocity relation

$$A(\alpha,\beta) = A(-\beta,-\alpha). \tag{2.3}$$

Remark 1. This reciprocity relation simply states that the behavior of the diffracted field observed in the direction β when the scatterers are illuminated by a plane wave of direction α, is identical to its behavior in the direction $-\alpha$ under an incident plane wave with direction $-\beta$. This property is a direct consequence of the symmetry of the Green function of the diffraction problem (which follows itself from the self-adjointness of the Dirichlet Laplacian).

Note that in (2.2), the integral actually represents the duality product between $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$ since $\partial u_T^\alpha/\partial\nu$ belongs to the latter in general. We keep this simplified notation in what follows.

By linearity, it follows from the results above that when illuminated by the Herglotz wave (2.1) associated with a given directional distribution $f \in L^2(S^2)$, the scattering obstacles generate the diffracted field u_D

$$u_D(x) = \int_{S^2} f(\alpha)\,u_D^\alpha(x)\,d\alpha.$$

Furthermore, the asymptotic behavior of u_D is given by the formula

$$u_D(\beta\|x\|) = \frac{e^{ik\|x\|}}{\|x\|}Ff(\beta) + O(\|x\|^{-2}),$$

where the far field $Ff(\beta)$ in the direction $\beta \in S^2$ is simply given by the relation

$$Ff(\beta) = \int_{S^2} A(\alpha, \beta)\, f(\alpha)\, d\alpha. \tag{2.4}$$

The integral operator $F : L^2(S^2) \longrightarrow L^2(S^2)$ with kernel $A(\cdot, \cdot)$ is known in the literature as the *far field operator*. Its properties are given in the following.

PROPOSITION 2.2. *The far field operator $F : L^2(S^2) \longrightarrow L^2(S^2)$ defined by equation* (2.4) *is a compact and normal operator. Its adjoint is the operator $F^* : L^2(S^2) \longrightarrow L^2(S^2)$ defined by*

$$F^* f = \overline{RF\overline{Rf}} \quad \forall f \in L^2(S^2), \tag{2.5}$$

where R is the symmetry operator defined by $Rf(\alpha) = f(-\alpha)\ \forall \alpha \in S^2$.

Proof. The compactness of the integral operator F follows immediately from the analyticity of its kernel $A(\cdot, \cdot)$. The fact that F is a normal operator is a well-known result, which is proved, for instance, in [5] (see Corollary 2.5). The adjoint F^* of F is the integral operator with kernel

$$A^*(\alpha, \beta) = \overline{A(\beta, \alpha)} = \overline{A(-\alpha, -\beta)},$$

where we have used the reciprocity relation (2.3). Formula (2.5) follows. □

Remark 2. In fact, in [5], it is proved more precisely that

$$FF^* = F^*F = \frac{2\pi}{ik}(F - F^*). \tag{2.6}$$

Since the far field operator F is related to the scattering matrix by the relation $S = I + (ik/2\pi)F$, formula (2.6) can be seen as an equivalent formulation of the fact that the scattering operator S is unitary, which is a classical result in scattering theory (cf. [13]).

We are now able to give a rigorous definition of the time-reversal operator. During the time-reversal process, when a Herglotz wave associated with a density $f \in L^2(S^2)$ is emitted by the TRM, the far field corresponding to the diffracted field is measured, conjugated, and then re-emitted by the TRM. The new emission is characterized by the Herglotz wave associated with the density $g \in L^2(S^2)$ defined by

$$g = \overline{RFf}.$$

In this relation, the presence of the symmetry operator R is due to the fact that during the time-reversal process, the far field measured in a given direction $\beta \in S^2$ is used to define the new incident plane wave in the direction $-\beta$. The time-reversal operator T is then obtained by iterating this scheme once again, and thus, we have

$$Tf = \overline{RFg} = \overline{RF\overline{RFf}}.$$

Thanks to (2.5) and using the fact that F is a normal operator, we finally get the following.

PROPOSITION 2.3. *The time-reversal operator $T : L^2(S^2) \longrightarrow L^2(S^2)$ is given by*

$$T = F^*F = FF^*. \tag{2.7}$$

It is the integral operator with kernel

$$t(\alpha,\beta) = \frac{1}{4\pi}\int_{\Gamma\times\Gamma} j_0(k\|y-z\|)\,\frac{\partial u_T^\alpha}{\partial\nu}(y)\,\overline{\frac{\partial u_T^\beta}{\partial\nu}(z)}\;d\Gamma_y\,d\Gamma_z, \tag{2.8}$$

where $u_T^\alpha = u_I^\alpha + u_D^\alpha$ *denotes the total field associated with the incident field* u_I^α*, and* $j_0(\xi) = \sin(\xi)/\xi$ *is the spherical Bessel function of order* 0.

Proof. Since (2.7) has been already proved, we only have to show the second part of the proposition. From (2.7), it follows that T is the integral operator with kernel

$$t(\alpha,\beta) = \int_{S^2} A(\alpha,\gamma)\,\overline{A(\beta,\gamma)}\,d\gamma. \tag{2.9}$$

Substituting expression (2.2) of the scattering amplitude in the above relations and inverting the integrals over S^2 with the integrals over Γ, we find that

$$t(\alpha,\beta) = \frac{1}{(4\pi)^2}\int_{\Gamma\times\Gamma}\left(\int_{S^2}\overline{u_I^\gamma(y)}\,u_I^\gamma(z)\,d\gamma\right)\frac{\partial u_T^\alpha}{\partial\nu}(y)\,\overline{\frac{\partial u_T^\beta}{\partial\nu}(z)}\;d\Gamma_y\,d\Gamma_z.$$

Equation (2.8) follows then from the identity (cf. [1, p. 155])

$$\int_{S^2}\overline{u_I^\gamma(y)}\,u_I^\gamma(z)\,d\gamma = \int_{S^2} e^{ik\gamma\cdot(z-y)}\,d\gamma = 4\pi j_0(k\|y-z\|). \tag{2.10}$$

□

3. Global focusing. The time-reversal operator $T = F^*F : L^2(S^2) \longrightarrow L^2(S^2)$ is clearly a positive and self-adjoint operator. Moreover, by Proposition 2.2, it is also a compact operator. Besides the value 0, its spectrum is thus constituted of a finite or countable sequence of positive eigenvalues admitting 0 as the only possible accumulation point. In this section, we see how these eigenvectors can be used to generate incident waves that focus acoustic on the diffracting obstacles. These global focusing results (namely Propositions 3.2 and 3.3) actually are a reformulation of results obtained in [14]. First, we recall a classical result from linear operators theory (see, for instance, [20, p. 442]).

PROPOSITION 3.1. *Let* N *be a compact and normal on a Hilbert space* H*. If* $\lambda_1, \lambda_2, \dots$ *is the sequence of all nonzero eigenvalues of* N*, arranged such that* $|\lambda_1| \geq |\lambda_2| \geq \cdots$*, and if* $\varphi_1, \varphi_2, \dots$ *is a corresponding orthonormal sequence of eigenvectors, then* $|\lambda_1|^2 \geq |\lambda_2|^2 \geq \cdots$ *is the sequence of all nonzero eigenvalues of* $N^*N = NN^*$*, and* $\varphi_1, \varphi_2, \dots$ *is a corresponding orthonormal sequence of eigenvectors.*

This proposition shows that the nonzero eigenvalues of the time reversal operator $T = F^*F = FF^*$ are exactly the positive numbers $|\lambda_1|^2 \geq |\lambda_2|^2 \geq \cdots$, where the complex numbers $(\lambda_p)_{p\geq 1}$ denote the nonzero eigenvalues of the normal compact far field operator F. Furthermore, the corresponding eigenvectors $(f_p)_{p\geq 1}$ of F are exactly the eigenvectors of $T = F^*F$. Consequently, it suffices to analyze the focusing properties of the eigenvectors of the far field F to obtain the same results for the time reversal operator T.

Let us first deal with the largest eigenvalue of the far field operator. Then, we have the following.

PROPOSITION 3.2. *Let* λ_1 *be the largest eigenvalue (in modulus) of* F*, and let* $f_1 \in L^2(S^2)$ *be an eigenvector of* F *associated with* λ_1*. Then,*

$$\sup_{f\in L^2(S^2)\,,\,f\neq 0}\frac{\|Ff\|^2_{L^2(S^2)}}{\|f\|^2_{L^2(S^2)}} = \frac{\|Ff_1\|^2_{L^2(S^2)}}{\|f_1\|^2_{L^2(S^2)}} = |\lambda_1|^2.$$

In other words, the incident Herglotz wave $u_I^1(x) = \int_{S^2} f_1(\alpha)\, e^{ik\alpha\cdot x}\, d\alpha$ *is, among all the possible Herglotz waves, the one that maximizes the energy scattered by the obstacles.*

Proof. The proposition is a straightforward consequence of the Min-Max principle. Indeed, applying this principle to the positive self-adjoint and bounded operator $T = F^*F$, we can write that the largest eigenvalue $|\lambda_1|^2$ of T satisfies

$$|\lambda_1|^2 = \sup_{f\in L^2(S^2)\,,\, f\neq 0} \frac{(Tf,f)_{L^2(S^2)}}{\|f\|^2_{L^2(S^2)}} = \sup_{f\in L^2(S^2)\,,\, f\neq 0} \frac{\|Ff\|^2_{L^2(S^2)}}{\|f\|^2_{L^2(S^2)}}. \qquad \square$$

Roughly speaking, this result says that the "best" way to illuminate a family of obstacles with Herglotz waves is to use a Herglotz wave u_I^1 corresponding to an eigenvector f_1 of F (or T) associated with its largest eigenvalue λ_1. The physical reason explaining this property is that the incident field generated by an eigenvector f_p associated with any eigenvalue $\lambda_p \neq 0$ of F, focuses on the obstacles. More precisely, the following result holds true (see [14]).

PROPOSITION 3.3. *Let* $\lambda_p \neq 0$ *be an eigenvalue of* F *and* $f_p \in L^2(S^2)$, $f_p \neq 0$, *an eigenvector of* F *associated with* λ_p. *Then, the Herglotz wave* $u_{I,p}$ *associated with* f_p *and defined by* $u_{I,p}(x) = \int_{S^2} f_p(\alpha)\, u_I^\alpha(x)\, d\alpha = \int_{S^2} f_p(\alpha)\, e^{ik\alpha\cdot x}\, d\alpha$, *has the following form:*

$$u_{I,p}(x) = \frac{1}{\lambda_p} \int_\Gamma j_0(k\|x-y\|)\, \frac{\partial u_{T,p}}{\partial \nu}(y)\, d\Gamma_y, \tag{3.1}$$

where $u_{T,p} = u_{I,p} + u_{D,p}$ *denotes the total field associated with the incident field* $u_{I,p}$.

Proof. Since $f_p(\beta) = \lambda_p^{-1} F f_p(\beta) = \lambda_p^{-1} \int_{S^2} A(\alpha,\beta)\, f_p(\alpha)\, d\alpha$, we obtain by using expression (2.2) of $A(\alpha,\beta)$ that

$$\begin{aligned} f_p(\beta) &= (4\pi\lambda_p)^{-1} \int_{S^2} \int_\Gamma \frac{\partial u_T^\alpha}{\partial \nu}\, \overline{u_I^\beta}\, d\Gamma\; f_p(\alpha)\, d\alpha \\ &= (4\pi\lambda_p)^{-1} \int_\Gamma \int_{S^2} \frac{\partial u_T^\alpha}{\partial \nu}\, f_p(\alpha)\, d\alpha\; \overline{u_I^\beta}\, d\Gamma. \end{aligned}$$

But by superposition, the integral $\int_{S^2} \partial u_T^\alpha/\partial\nu\, f_p(\alpha)\, d\alpha$ is nothing but the normal derivative of the total field $u_{T,p}$ associated with the incident field $u_{I,p}$, and thus

$$f_p(\beta) = (4\pi\lambda_p)^{-1} \int_\Gamma \frac{\partial u_{T,p}}{\partial \nu}\, \overline{u_I^\beta}\, d\Gamma. \tag{3.2}$$

We can now obtain the expression of the incident field generated by the eigenvector f_p. From (3.2), we have

$$\begin{aligned} u_{I,p}(x) &= \int_{S^2} f_p(\beta)\, u_I^\beta(x)\, d\beta \\ &= (4\pi\lambda_p)^{-1} \int_{S^2} \int_\Gamma \frac{\partial u_{T,p}}{\partial \nu}\; \overline{u_I^\beta}\, d\Gamma\; u_I^\beta(x)\, d\beta \\ &= (4\pi\lambda_p)^{-1} \int_\Gamma \left(\int_{S^2} u_I^\beta(x)\, \overline{u_I^\beta(y)}\, d\beta \right) \frac{\partial u_{T,p}}{\partial \nu}(y)\, d\Gamma_y. \end{aligned}$$

Formula (3.1) follows then from identity (2.10). $\square$

Since $j_0(\xi) = \sin(\xi)/\xi$, formula (3.1) shows that, as expected, the incident field $u_{I,p}(x)$ generated by an eigenvector f_p of F (or T) decreases like $r(x)^{-1}$ if $r(x)$ denotes the distance of x to the obstacles. In this sense, one can say that $u_{I,p}$ focuses on the obstacles located in the propagation medium. Furthermore, the quality of this focusing (given by the amplitude of the far field) is exactly given by the magnitude of the eigenvalue λ_p, since

$$|\lambda_p| = \frac{\|Ff_p\|_{L^2(S^2)}}{\|f_p\|_{L^2(S^2)}}.$$

4. Selective focusing. The aim of this section is to propose a mathematical justification of the so-called D.O.R.T. method presented in [16] and briefly described in the introduction of this paper. Roughly speaking, we answer the two following questions.

(i) Is the number of obstacles contained in a homogeneous medium equal to the number of "significant" eigenvalues of the far field operator F (or, equivalently, to those of the time-reversal operator $T = F^*F = FF^*$)?

(ii) If so, do the associated eigenvectors selectively focus on the obstacles?

As can be seen from the numerical experiments presented in [6], the answer to the first question is, in general, negative (there can be several "significant" eigenvalues even when there is just one scatterer). We will confirm this result by studying in subsection 4.1 the special case of a single spherical obstacle. Nevertheless, we will show that the answer becomes positive provided the obstacles considered are small enough. Under this assumption, we show in subsection 4.2 that selective focusing can be achieved using the eigenvectors of the far field operator.

4.1. Diffraction by a single spherical obstacle. In this subsection, we deal with the case where the scatterer is a sphere of radius $a > 0$. For this particular geometry, an explicit formula can be obtained for the eigenvalues of the far field mapping and thus for those of the time-reversal operator. The results of this subsection are classical and can be found, for instance, in [7]. In particular, formula (3.30) in [7] shows that for any given density

$$f = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} a_n^m Y_n^m \in L^2(S^2),$$

we have

$$Ff(\beta) = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} \frac{4i\pi}{k} \frac{j_n(ka)}{h_n^1(ka)} a_n^m Y_n^m(\beta).$$

Here Y_n^m denotes the usual spherical harmonics, j_n and h_n^1 are, respectively, the spherical Bessel and Hankel functions of order n.

Since the spherical harmonics constitute an orthonormal basis of $L^2(S^2)$, this formula shows that the following result holds.

PROPOSITION 4.1. *The eigenvalues of the far field operator F in the case of a single sound-soft spherical scatterer of radius a are given by*

$$\lambda_n = \frac{4i\pi}{k} \frac{j_n(ka)}{h_n^1(ka)} \qquad \forall n \geq 1. \tag{4.1}$$

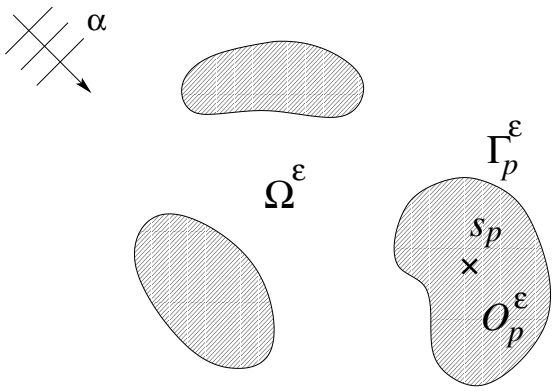


FIG. 4.1.

The eigenspace associated with the eigenvalue λ_n is the vector space of dimension $2n+1$ with basis Y_n^m, for $|m| \leq n$.

Remark 3. Equation (4.1) shows that the eigenvalues λ_n of the far field operator F satisfy $|\lambda_n| \leq 4\pi/k$ (recall that $h_n^1 = j_n + iy_n$). This property can also be obtained from the fact that the scattering operator $S = I + (ik/2\pi)F$ is unitary. Indeed, this property implies that the eigenvalues λ_n lie on the circle of radius $4\pi/k$ centered at $(0, 2\pi/k)$.

Proposition 4.1 shows in particular that *the number of nonzero eigenvalues is not necessarily equal to the number of obstacles*. However, in the case of a point scatterer or in the case of the low-frequency scattering (both cases which correspond to the asymptotic limit $ka \longrightarrow 0$), this result becomes true. Indeed, using the asymptotic behavior of Bessel and Hankel functions, we easily see that the eigenvalues λ_n given by (4.1) satisfy

$$\lambda_n \sim -\frac{4\pi^2}{k} \frac{(ka/2)^{2n+1}}{\Gamma(n+1/2)\Gamma(n+3/2)}$$

when ka goes to zero (and n is fixed). Thus, λ_{n+1}/λ_n decreases like $(ka)^2$, and hence, one can consider that the only significant eigenvalue in the limit case $ka \longrightarrow 0$ is the largest one λ_1. This observation suggests that the number of nonzero eigenvalues can be related to the number of obstacles when the obstacles are small. The next subsection provides a justification of this statement.

4.2. Diffraction by several small obstacles. Consider a family of obstacles $\{\mathcal{O}_p^\varepsilon;\ p = 1, N\}$ depending on a small parameter ε, where each $\mathcal{O}_p^\varepsilon$ is the image of a reference open set $\mathcal{O}_p$ (which is assumed to contain the origin) by a dilation of ratio ε centered at a given point $s_p \in \mathbb{R}^3$ (see Figure 4.1):

$$\mathcal{O}_p^\varepsilon = \left\{ x \in \mathbb{R}^3;\ \xi = \frac{x - s_p}{\varepsilon} \in \mathcal{O}_p \right\}.$$

Of course the "centers" s_p are chosen different so that for small enough ε, the obstacles do not intersect.

The main ingredient to show that selective focusing can be achieved using the eigenvectors of the far field operator when ε is small enough is given by the following

result, which provides the asymptotic behavior of the scattering amplitude $A^\varepsilon(\alpha,\beta)$ associated with the family of obstacles $\{\mathcal{O}_p^\varepsilon\}$.

PROPOSITION 4.2. *There exist N positive constants $C_1,\dots,C_N$ depending only on the geometry of the reference obstacles $\mathcal{O}_1,\dots,\mathcal{O}_N$ (called the "capacities" of these obstacles) such that*

$$\frac{A^\varepsilon(\alpha,\beta)}{\varepsilon} = A^{(1)}(\alpha,\beta) + O(\varepsilon) \quad \textit{with} \quad A^{(1)}(\alpha,\beta) = \frac{-1}{4\pi}\sum_{p=1,N} C_p\, u_I^\alpha(s_p)\,\overline{u_I^\beta(s_p)}, \tag{4.2}$$

where the bound $O(\varepsilon)$ is uniform for all $\alpha,\beta\in S^2$.

For the sake of clarity, the—rather technical—proof of this proposition is given in the appendix.

Remark 4. The capacity of a spherical soft obstacle of radius a is $C = 4\pi a$ (since the solution to (A.7) is simply given in this case by $V(x) = a/\|x\|$).

Thanks to Proposition 4.2, we know that the far field operator F^ε of the family of obstacles $\{\mathcal{O}_p^\varepsilon;\ p=1,N\}$ satisfies

$$\left\|\varepsilon^{-1}F^\varepsilon - F^{(1)}\right\|_{\mathcal{L}(L^2(S^2))} = \sup_{f\in L^2(S^2)\setminus\{0\}} \frac{\|(\varepsilon^{-1}F^\varepsilon - F^{(1)})f\|_{L^2(S^2)}}{\|f\|_{L^2(S^2)}} = O(\varepsilon),$$

where $F^{(1)}$ is the integral operator on $L^2(S^2)$ with kernel $A^{(1)}$:

$$F^{(1)}f\,(\beta) = \int_{S^2} A^{(1)}(\alpha,\beta)\, f(\alpha)\, d\alpha.$$

Since F^ε is compact and normal, perturbation theory [11] ascertains the continuity of any finite system of eigenvalues as well as of the associated total eigenprojection. More precisely, assume that $\lambda^{(1)}$ is an isolated eigenvalue of $F^{(1)}$ with finite multiplicity m, which implies that $\lambda^{(1)}\neq 0$.

(i) Then for small enough ε, the spectrum of $\varepsilon^{-1}F^\varepsilon$ can be separated into two parts. On one hand, the so-called $\lambda^{(1)}$-*group* consists of $m'\le m$ eigenvalues λ_j^ε, with $j=1$ to m', having a constant multiplicity m_j for $\varepsilon\neq 0$, and which are continuous near $\varepsilon = 0$, namely

$$|\lambda_j^\varepsilon - \lambda^{(1)}| = O(\varepsilon).$$

Moreover, the total multiplicity $\sum_{j=1,m'} m_j$ of the $\lambda^{(1)}$-group coincide with the multiplicity m of $\lambda^{(1)}$. On the other hand, the complementary of the $\lambda^{(1)}$-group in the spectrum of $\varepsilon^{-1}F^\varepsilon$ lies outside a vicinity of $\lambda^{(1)}$.

(ii) The total projection P^ε for the $\lambda^{(1)}$-group, i.e., the sum of the orthogonal projections on the eigenspaces associated with the λ_j^ε, is continuous at $\varepsilon = 0$, and

$$\left\|P^\varepsilon - P^{(1)}\right\|_{\mathcal{L}(L^2(S^2))} = O(\varepsilon),$$

where $P^{(1)}$ is the eigenprojection associated with $\lambda^{(1)}$.

Notice that in general, one cannot assert the existence of a continuous family of eigenvectors associated respectively with the λ_j^ε. However, for our particular choice of geometric perturbation (ε-dilation), such a result holds, since the perturbation actually is analytic with respect to ε (which is easily deduced from the appendix). But this result is of poor practical interest.

An eigenvalue of $\varepsilon^{-1}F^\varepsilon$ either belongs to some $\lambda^{(1)}$-group for a nonzero eigenvalue $\lambda^{(1)}$ of $F^{(1)}$, or vanishes as ε tends to 0. In the latter case, the above result does not apply; perturbation theory only provides the continuity of nonstationary eigenelements. So it remains to study the spectral properties of $F^{(1)}$, whose *degenerate kernel* will be rewritten in the form

$$A^{(1)}(\alpha,\beta) = -\sum_{p=1,N} C_p\,\overline{e_p(\alpha)}\,e_p(\beta), \quad \text{where} \quad e_p(\alpha) = \frac{e^{-ik\alpha\cdot s_p}}{2\sqrt{\pi}} \ (p=1,N). \tag{4.3}$$

Remark 5. Each e_p appears as a normalized function of $L^2(S^2)$ corresponding to an incident Herglotz wave $u_{I,p}$ which focuses on the pth obstacle, for

$$u_{I,p}(x) = \int_{S^2} e_p(\alpha)\,u_I^\alpha(x)\,d\alpha = 2\sqrt{\pi}\,j_0(k\|x-s_p\|),$$

by virtue of (2.10).

The above expression of $A^{(1)}$ then yields

$$F^{(1)}f = -\sum_{p=1,N} C_p\ (f,e_p)_{L^2(S^2)}\ e_p. \tag{4.4}$$

Proposition 4.3. *The limit far field operator* (4.4) *is a negative self-adjoint operator with finite rank N (the number of obstacles) and whose spectral radius cannot be smaller than the greatest capacity C_p of the obstacles.*

In the case where the wavelength $\ell = 2\pi/k$ is small compared with the minimum distance $d = \min_{1\le p\ne q\le N}\|s_p - s_q\|$ between the obstacles, the family $\{e_p;\ p=1,N\}$ defined in (4.3) *provides an approximate basis of eigenvectors associated with the approximate eigenvalues $\{-C_p;\ p=1,N\}$:*

$$F^{(1)}e_p = -C_p\,e_p + O\left(\frac{\ell}{d}\right). \tag{4.5}$$

Proof. The bilinear form associated with $F^{(1)}$,

$$(F^{(1)}f,f')_{L^2(S^2)} = -\sum_{p=1,N} C_p\ (f,e_p)_{L^2(S^2)}\ \overline{(f',e_p)_{L^2(S^2)}},$$

is clearly negative and self-adjoint, and so is $F^{(1)}$. The range of $F^{(1)}$ is spanned by $\{e_p;\ p=1,N\}$. To see that this family is linearly independent, suppose that

$$\sum_{p=1,N} z_p\,e_p = 0 \text{ with } z_p\in\mathbb{C}.$$

It is clear that the function $e_p\in L^2(S^2)$ is nothing but the far field corresponding to a point source located at the point s_p. Consequently, the above relation simply states that we have chosen a superposition of point sources located at the points $(s_p)_{p=1,N}$ whose far field vanishes. Thus, by Rellich's lemma, the field is identically zero. Hence, all the coefficients $(z_p)_{p=1,N}$ of the linear combination must also vanish. The linear independence of the family $\{e_p;\ p=1,N\}$ is thus established.

The lower bound for the spectral radius follows from the fact that

$$\left|(F^{(1)}e_q,e_q)_{L^2(S^2)}\right| = \sum_{p=1,N} C_p\,\left|\ (e_p,e_q)_{L^2(S^2)}\right|^2 \ge C_q \quad \text{for } q=1,N,$$

since the e_p are normalized in $L^2(S^2)$. On the other hand, nothing can be said in general about the gap between 0 and the other eigenvalues, which may be arbitrarily close to the former. This actually depends on the constructive or destructive interactions between the different obstacles, which are measured by the following scalar products (see (2.10)):

$$(e_p, e_q)_{L^2(S^2)} = j_0(k\|s_p - s_q\|) = \frac{\sin(k\|s_p - s_q\|)}{k\|s_p - s_q\|}.$$

These relations show in particular that

$$(e_p, e_q)_{L^2(S^2)} = \begin{cases} 1 & \text{for } q = p, \\ O\left(\dfrac{\ell}{d}\right) & \text{for } q \neq p, \end{cases}$$

which means that $\{e_p;\ p = 1, N\}$ is close to an orthogonal basis of the range of $F^{(1)}$ when $\ell \ll d$. The estimate (4.5) follows: each e_p is an approximate eigenvector. □

What are the practical consequences of the above results as regards selective focusing? Mainly that the eigenvectors of the time-reversal operator (or the far field operator) will produce selective focusing acoustic waves if

(i) the obstacles are small enough, compared to the wavelength,

(ii) the smallest distance between them is large, compared again to the wavelength,

(iii) their capacities are all distinct.

Indeed in this case all the nonzero eigenvalues of $F^{(1)}$ will be simple: the diagonalization of the time-reversal operator will then yield approximations of the focusing densities e_p.

But if one of these assumptions is missing, the nice focusing properties will disappear, at least for some groups of eigenvectors.

On one hand, if the interactions between the obstacles become significant, i.e., when $d/\ell = O(1)$, these properties may reduce to the purely global focusing presented in section 3. In particular, for very low frequencies, the situation $\varepsilon \ll d \ll \ell$ may occur. In this case we have

$$e_p = \tilde{e} + O\left(\frac{d}{\ell}\right) \quad \text{with } \tilde{e}(\alpha) = \frac{e^{-ik\,\alpha\cdot\tilde{s}}}{2\sqrt{\pi}},$$

where $\tilde{s}$ may be chosen as a convex combination of the s_p. As a consequence

$$F^{(1)} f = -\left(\sum_{p=1,N} C_p\right)\ (f, \tilde{e})_{L^2(S^2)}\ \tilde{e} + O\left(\frac{d}{\ell}\right),$$

which shows that the cluster of obstacles behaves like a unique obstacle which accumulates their respective capacities; only one significant eigenvalue of the time-reversal operator may be observed. Of course, for several distant clusters, we shall recover selective focusing on each cluster.

On the other hand, if some of the obstacles have neighboring capacities, the time-reversal operator may admit nonsimple eigenvalues. In this situation, the diagonalization of the latter cannot choose the selective focusing densities among all their linear combinations which compose the corresponding eigenspace.

5. Open time-reversal mirrors. In this section, we consider the case of a TRM that does not entirely surround the obstacle. Given a subset $\widehat{S}$ of S^2, we assume that the TRM can emit plane waves of directions $\alpha \in \widehat{S}$, and measures the far field in the opposite directions $\beta \in (-\widehat{S})$. One emission-diffraction-reception cycle is described by the directional far field operator

$$\widehat{F} = \widehat{P}_- \, F \, \widehat{P}_+^* : L^2(+\widehat{S}) \longrightarrow L^2(-\widehat{S}),$$

where $\widehat{P}_\pm$ are the restriction operators from $L^2(S^2)$ to $L^2(\pm\widehat{S})$, and thus their respective adjoints $\widehat{P}_\pm^* : L^2(\pm\widehat{S}) \longrightarrow L^2(S^2)$ are the operators of continuation by 0 outside $\pm\widehat{S}$. Note here that $\widehat{F}$ appears as the integral operator

$$\widehat{F}f(\beta) = \int_{+\widehat{S}} A(\alpha,\beta)\, f(\alpha)\, d\alpha \quad \text{for } \beta \in -\widehat{S}.$$

The time-reversal operator $\widehat{T}$ in the case of an open TRM is then defined by

$$\widehat{T}f = \overline{\widehat{R}\,\widehat{F}\overline{\widehat{R}\,\widehat{F}f}},$$

where $\widehat{R} : L^2(-\widehat{S}) \longrightarrow L^2(+\widehat{S})$ is the restriction of the symmetry operator defined in section 2 (*i.e.*, $\widehat{R}f(\alpha) = f(-\alpha)$ for $\alpha \in \widehat{S}$).

But one can easily check that $\widehat{R}\,\widehat{P}_- = \widehat{P}_+ R$ and $\widehat{P}_+^*\,\widehat{R} = R\,\widehat{P}_-^*$, and since these operators commute with the conjugation, we have by virtue of (2.5)

$$\widehat{F}^* f = \widehat{P}_+ F^* \widehat{P}_-^* f = \overline{\widehat{P}_+ R F \overline{R \widehat{P}_-^* f}} = \overline{\widehat{R}\,\widehat{P}_- F \widehat{P}_+^* \overline{\widehat{R} f}} = \overline{\widehat{R}\,\widehat{F}\,\overline{\widehat{R} f}}.$$

Hence, we can state the following result.

PROPOSITION 5.1. *The time-reversal operator $\widehat{T}$ for an open TRM is given by*

$$\widehat{T} = \widehat{F}^* \widehat{F} : L^2(\widehat{S}) \longrightarrow L^2(\widehat{S}).$$

Thus, it is the integral operator with kernel

$$\widehat{t}(\alpha,\beta) = \int_{-\widehat{S}} A(\alpha,\gamma)\, \overline{A(\beta,\gamma)}\, d\gamma \quad \text{for } \alpha,\beta \in \widehat{S}. \tag{5.1}$$

Moreover, $\widehat{T}$ defines a compact positive and self-adjoint operator.

Besides the value 0, the spectrum $\widehat{T}$ is thus constituted of a finite or countable sequence of positive eigenvalues $(\widehat{\mu}_p)_{p\geq 1}$ admitting 0 for only possible accumulation point. The largest eigenvalue $\widehat{\mu}_1$ of $\widehat{T}$ is thus given by

$$\widehat{\mu}_1 = \sup_{f \in L^2(\widehat{S})\,,\, f\neq 0} \frac{\left(\widehat{T}f, f\right)_{L^2(\widehat{S})}}{\|f\|^2_{L^2(S^2)}} = \sup_{f \in L^2(\widehat{S})\,,\, f\neq 0} \frac{\|\widehat{F}f\|^2_{L^2(-\widehat{S})}}{\|f\|^2_{L^2(+\widehat{S})}}.$$

This expression shows in particular that the incident field corresponding to an eigenvector associated with this eigenvalue maximizes the diffracted field in the direction of the TRM. Our goal now is to see if the global and selective properties proved respectively in sections 3 and 4 for closed mirrors still hold in the case of an open TRM. The

main difference between both situations is that in the latter one, the directional far field operator $\widehat{F}$ is not anymore normal (the range of $\widehat{F}^*\widehat{F}$ is contained in $L^2(\widehat{S})$, when that of $\widehat{F}\widehat{F}^*$ is contained in $L^2(-\widehat{S})$). Consequently, the eigenelements of $\widehat{T} = \widehat{F}^*\widehat{F}$ cannot be directly related to those of $\widehat{F}$. Contrary to the case of a closed TRM, the spectral analysis need thus to be carried on the time reversal operator and not on the far field one. Nevertheless, as we are going to see now, all the focusing results obtained previously still hold.

5.1. Global focusing. In this subsection, we prove a global focusing property similar to the one given in Proposition 3.3. More precisely, we have the following result.

PROPOSITION 5.2. *Let $\widehat{\mu}_p \neq 0$ be an eigenvalue of $\widehat{T}$ and $\widehat{f}_p \in L^2(\widehat{S})$ be a corresponding eigenvector. Then, the Herglotz wave $\widehat{u}_{I,p}$ associated with $\widehat{f}_p$ and defined by $\widehat{u}_{I,p}(x) = \int_{\widehat{S}} \widehat{f}_p(\alpha)\, u_I^\alpha(x)\, d\alpha$ can be written in the form*

$$\widehat{u}_{I,p}(x) = \int_\Gamma \widehat{\jmath}(k(x-y))\, h_p(y)\, d\Gamma \tag{5.2}$$

for some density h_p, where

$$\widehat{\jmath}(k(x-y)) = \int_{\widehat{S}} u_I^\beta(x)\, \overline{u_I^\beta(y)}\, d\beta = \int_{\widehat{S}} e^{ik\beta\cdot(x-y)}\, d\beta. \tag{5.3}$$

Proof. Like in the proof of Proposition 3.3, formula (5.2) will be proved if we can write $\widehat{f}_p$ in the form

$$\widehat{f}_p(\beta) = \int_\Gamma h_p\, \overline{u_I^\beta}\, d\Gamma \tag{5.4}$$

for a given density h_p. Indeed, if such a relation holds, then

$$\widehat{u}_{I,p}(x) = \int_{\widehat{S}} \widehat{f}_p(\beta)\, u_I^\beta(x)\, d\beta = \int_{\widehat{S}} \int_\Gamma h_p\, \overline{u_I^\beta}\, d\Gamma\ u_I^\beta(x)\, d\beta.$$

Equation (5.2) follows then by inverting the integrals over $\widehat{S}$ and Γ.

Thus, it only remains to prove (5.4). We first write that for all $\beta \in \widehat{S}$,

$$\widehat{f}_p(\beta) = \frac{1}{\widehat{\mu}_p}\widehat{T}\widehat{f}_p(\beta) = \frac{1}{\widehat{\mu}_p}\int_{\widehat{S}} \widehat{t}(\alpha,\beta)\widehat{f}_p(\alpha)\, d\alpha. \tag{5.5}$$

Thanks to the reciprocity relation (2.3), formula (5.1) can be written

$$\widehat{t}(\alpha,\beta) = \int_{-\widehat{S}} A(\alpha,\gamma)\, \overline{A(-\gamma,-\beta)}\, d\gamma.$$

Using the integral representation (2.2) in the above relation, we get after some simple computations that

$$\widehat{t}(\alpha,\beta) = \int_\Gamma h_p^\alpha\, \overline{u_I^\beta}\, d\Gamma, \tag{5.6}$$

where the density h_p^α is given by

$$h_p^\alpha(x) = \frac{1}{(4\pi)^2} \int_{-\widehat{S}} \int_\Gamma \frac{\partial u_T^\alpha}{\partial \nu} \, \overline{u_I^\gamma} \, d\Gamma \, \overline{\frac{\partial u_T^{-\gamma}}{\partial \nu}(x)} \, d\gamma.$$

Combining (5.5) and (5.6), one obtains the claimed relation (5.4), with the density $h_p(x) = \widehat{\mu}_p^{-1} \int_{\widehat{S}} h_p^\alpha(x) \, \widehat{f}_p(\alpha) \, d\alpha$. ∎

It is well known in oscillatory integrals theory that the function $\widehat{\jmath}(x)$ defined by (5.3) satisfies (one can use the stationary phase theorem; see, for instance, Theorem 1 in [19, p. 322])

$$\widehat{\jmath}(x) = \mathcal{O}\left(\|x\|^{-1}\right). \tag{5.7}$$

In the directions which are not covered by the TRM (i.e., when $x/\|x\| \notin \pm S$), one can in fact obtain a faster decay for $\widehat{\jmath}(x)$, since we have then $\widehat{\jmath}(x) = \mathcal{O}\left(\|x\|^{-3/2}\right)$.

Thanks to (5.7), formula (5.2) shows thus that the incident field generated by an eigenvector of $\widehat{T}$ focuses on the obstacles located in the propagation medium.

5.2. Diffraction by several small obstacles. Now we turn to the analysis of the selective focusing in the case of a TRM partially surrounding several small obstacles. The assumptions made on the geometry of the small scatterers are identical to those made in section 4. Let us recall that the main difference with section 4 is that since $\widehat{F}$ is not normal, the spectral analysis can no longer be achieved on $\widehat{F}$ but has to be carried out directly on the time-reversal operator $\widehat{T}$. In this subsection, we are going to see that the selective focusing results obtained in section 4 can be extended to the case of an open mirror.

Using the asymptotic formula (4.2) of the scattering amplitude, one easily gets that the kernel $\widehat{t}^{\,\varepsilon}(\alpha,\beta)$ of the time-reversal operator $\widehat{T}^\varepsilon = (\widehat{F}^\varepsilon)^* \widehat{F}^\varepsilon$ satisfies

$$\frac{\widehat{t}^{\,\varepsilon}(\alpha,\beta)}{\varepsilon^2} = \widehat{t}^{\,(1)}(\alpha,\beta) + O(\varepsilon),$$

where

$$\widehat{t}^{\,(1)}(\alpha,\beta) = \int_{-\widehat{S}} A^{(1)}(\alpha,\gamma) \, \overline{A^{(1)}(\beta,\gamma)} \, d\gamma \qquad \forall \alpha,\beta \in \widehat{S}$$

and where $A^{(1)}(\cdot,\cdot)$ is the degenerate kernel defined in (4.2). Since $\widehat{T}^\varepsilon$ is compact and self-adjoint, classical results of perturbation theory show again that for small ε, the spectral elements of $\varepsilon^{-2}\widehat{T}^\varepsilon$ can be approximated by those of the integral operator $\widehat{T}^{(1)}$ with kernel $\widehat{t}^{\,(1)}(\cdot,\cdot)$, which also reads $\widehat{T}^{(1)} = (\widehat{F}^{(1)})^* \widehat{F}^{(1)}$, where the operator $\widehat{F}^{(1)} : L^2(\widehat{S}) \longrightarrow L^2(-\widehat{S})$ is defined by

$$\widehat{F}^{(1)} f(\beta) = -\sum_{p=1,N} C_p \, (f, e_p)_{L^2(\widehat{S})} \, e_p(\beta) \qquad \text{for } \beta \in -\widehat{S}.$$

If we define the normalized functions $\{\widehat{e}_p;\ p = 1, N\}$ in $L^2(\widehat{S})$ and $L^2(-\widehat{S})$ by

$$\widehat{e}_p(\alpha) = (4\pi\widehat{r}\,)^{-1/2} \, e^{-ik\alpha\cdot s_p}, \tag{5.8}$$

where $\widehat{r} = \mathrm{mes}(\widehat{S})/(4\pi)$ is the *opening ratio* of the TRM, then

$$\widehat{F}^{(1)} f = -\widehat{r} \sum_{p=1,N} C_p \, (f, \widehat{e}_p)_{L^2(\widehat{S})} \, \widehat{e}_p \ \text{ in } L^2(-\widehat{S}). \tag{5.9}$$

Hence, for all $f \in L^2(\widehat{S})$, we have

$$\widehat{T}^{(1)} f = \widehat{r}^{\,2} \sum_{q=1,N} C_q \left(\sum_{p=1,N} C_p \, (f, \widehat{e}_p)_{L^2(\widehat{S})} (\widehat{e}_p, \widehat{e}_q)_{L^2(-\widehat{S})} \right) \widehat{e}_q. \tag{5.10}$$

We can now state the main result of this subsection.

PROPOSITION 5.3. *The limit time reversal operator* $\widehat{T}^{(1)} : L^2(\widehat{S}) \longrightarrow L^2(\widehat{S})$ *defined by* (5.10) *is a self-adjoint operator with finite rank* N *(the number of obstacles).*

Furthermore, if the wavelength $\ell = 2\pi/k$ *is small compared with the minimum distance between the obstacles, the family* $\{\widehat{e}_p;\ p = 1, N\}$ *defined in* (5.8) *provides an approximate basis of eigenvectors of* $\widehat{T}^{(1)}$ *associated with the approximate eigenvalues* $(\widehat{r}\, C_p)^2$*:*

$$\widehat{T}^{(1)} \widehat{e}_p = (\widehat{r}\, C_p)^2 \, \widehat{e}_p + O\left(\frac{\ell}{d}\right). \tag{5.11}$$

Proof. The fact that $\widehat{T}^{(1)}$ is of rank N follows from the fact that the family $\{\widehat{e}_p; p = 1, N\}$ is linearly independent in $L^2(\widehat{S})$(see the proof of Proposition 5.3). Equation (5.11) follows from (5.10) combined with the fact that

$$(\widehat{e}_p, \widehat{e}_q)_{L^2(\widehat{S})} = \overline{(\widehat{e}_p, \widehat{e}_q)_{L^2(-\widehat{S})}} = \begin{cases} 1 & \text{for } p = q, \\ O\left(\dfrac{\ell}{d}\right) & \text{for } p \neq q. \end{cases}$$

The last estimate follows from the relation

$$(\widehat{e}_p, \widehat{e}_q)_{L^2(\widehat{S})} = (4\pi\hat{r})^{-1} \int_{\widehat{S}} e^{ik\beta\cdot(s_p - s_q)} \, d\beta = (4\pi\hat{r})^{-1} \widehat{\jmath}(k(s_p - s_q))$$

and from the decay property (5.7) of $\widehat{\jmath}$ for $p \neq q$. ∎

Remark 6. Contrary to the case of a closed mirror (compare Propositions 4.3 and 5.3), we have not been able to compare the spectral radius of $\hat{T}^{(1)}$ with the greatest value taken by the quantities $(\hat{r}C_p)^2$.

As in the case of a closed mirror, Proposition 5.3 shows that the eigenvectors of the time-reversal operator for an open mirror will produce selective focusing acoustic waves if

(i) the obstacles are small enough, compared to the wavelength,
(ii) the smallest distance between them is large, compared to the wavelength,
(iii) their capacities are all distinct.

Indeed in this case all the nonzero eigenvalues of $\widehat{T}^{(1)}$ will be simple: the diagonalization of the time-reversal operator will then yield approximations of the focusing densities $\widehat{e}_p$ since each $\widehat{e}_p$ generates an incident Herglotz wave $\widehat{u}_{I,p}$ which focuses on the pth obstacle for

$$\widehat{u}_{I,p}(x) = \int_{\widehat{S}} \widehat{e}_p(\alpha) \, u_I^{\alpha}(x) \, d\alpha = \frac{1}{\sqrt{4\pi\widehat{r}}} \, \widehat{\jmath}(k(x - s_p)) = O\left(\frac{\ell}{\|x - s_p\|}\right).$$

Appendix A. Asymptotics for small obstacles. We detail here a constructive proof of the asymptotic behavior (4.2) of the scattering amplitude for small obstacles, claimed in Proposition 4.2. This result is formally derived in other papers (see, e.g.,

[17, 18]). A more abstract proof based on potential theory was recently proposed in [8].

The idea of our proof is to rewrite the scattering problem as a *regular* perturbation of a Fredholm equation in a *fixed* Hilbert space, in the sense that it does not depend on the size, say ε, of the obstacles:

$$(I + \mathbb{K}^\varepsilon)\varphi^\varepsilon = g^\varepsilon. \tag{A.1}$$

We obtain such a formulation by means of a variant of the integral method introduced by Jami and Lenoir [10], which has the advantage to involve nonsingular kernels, contrary to usual integral equations (for which perturbation theory requires more complicated arguments).

Consider the family of obstacles $\{\mathcal{O}_p^\varepsilon;\ p = 1, N\}$ introduced in subsection 4.2. We denote by Γ_p^ε (respectively, Γ_p) the boundary of $\mathcal{O}_p^\varepsilon$ (respectively, of $\mathcal{O}_p$), $\Gamma^\varepsilon = \bigcup_{p=1,N} \Gamma_p^\varepsilon$ and $\mathcal{O}^\varepsilon = \bigcup_{p=1,N} \mathcal{O}_p^\varepsilon$. Our exterior Dirichlet problem for the diffracted field u^ε reads

$$\begin{cases} \Delta u^\varepsilon + k^2 u^\varepsilon = 0 \ \text{ in } \mathbb{R}^3 \setminus \overline{\mathcal{O}^\varepsilon}, \\ u^\varepsilon = f \ \text{ on } \Gamma^\varepsilon, \\ R.C., \end{cases} \tag{A.2}$$

where $R.C.$ stands for the outgoing radiation condition, and $f = -u_I^\alpha$ is the Dirichlet datum associated with an incident plane wave $u_I^\alpha(x) = \exp(ik\alpha\cdot x)$ of direction $\alpha \in S^2$.

Reduction to a bounded domain. Around each reference obstacle $\mathcal{O}_p$, we delimit a bounded part D_p of its exterior by a fictitious boundary Σ_p which does not intersect Γ_p. We denote by D_p^ε and Σ_p^ε the images of D_p and Σ_p by the same dilation as for $\mathcal{O}_p^\varepsilon$, as well as $D^\varepsilon = \bigcup_{p=1,N} D_p^\varepsilon$ and $\Sigma^\varepsilon = \bigcup_{p=1,N} \Sigma_p^\varepsilon$.

The Jami–Lenoir method consists of introducing a transparent boundary condition on Σ^ε which is derived from the usual integral representation of u^ε. Here, in order to get rid of the normal derivative of u^ε on Γ^ε, the single-layer potential is re-expressed as a volume potential by Green's formula. Indeed it is easy to see that near Σ^ε we have

$$u^\varepsilon = f \overset{\Gamma^\varepsilon}{*} \frac{\partial G_k}{\partial \nu} + k^2 u^\varepsilon \overset{D^\varepsilon}{*} (\chi^\varepsilon G_k) - \nabla u^\varepsilon \overset{D^\varepsilon}{*} \nabla(\chi^\varepsilon G_k),$$

where the different "convolutions" represent, respectively, the surface double-layer potential

$$\left\{ f \overset{\Gamma^\varepsilon}{*} \frac{\partial G_k}{\partial \nu} \right\}(x) = \int_{\Gamma^\varepsilon} f(y)\, \frac{\partial G_k}{\partial \nu_y}(x-y)\, d\gamma_y$$

and the volume potentials

$$\begin{aligned} \{u^\varepsilon \overset{D^\varepsilon}{*} (\chi^\varepsilon G_k)\}(x) &= \int_{D^\varepsilon} u^\varepsilon(y)\, \chi^\varepsilon(y) G_k(x-y)\, dy, \\ \{\nabla u^\varepsilon \overset{D^\varepsilon}{*} \nabla(\chi^\varepsilon G_k)\}(x) &= \int_{D^\varepsilon} \nabla u^\varepsilon(y) \cdot \nabla_y(\chi^\varepsilon(y) G_k(x-y))\, dy. \end{aligned}$$

In the above expressions, G_k stands for the outgoing Green function of $\Delta + k^2$, *i.e.*, $G_k(x) = -\exp(ik|x|)/(4\pi|x|)$, and χ^ε denotes a family of regular cutoff functions $(\chi_p^\varepsilon)_{p=1,\dots,N}$ defined by the ε-dilation: $\chi_p^\varepsilon(x) = \chi_p((x - s_p)/\varepsilon)$ if $x \in D_p^\varepsilon$, where each

χ_p is equal to 1 in a vicinity of Γ_p and 0 in a vicinity of Σ_p. Note that these integrals involve regular kernels when x is near Σ^ε.

As a consequence if u^ε solves (A.2), its restriction v^ε to D^ε satisfies

$$\text{(A.3)}\quad \left\{ \begin{array}{l} \Delta v^\varepsilon + k^2 v^\varepsilon = 0 \ \text{ in } D^\varepsilon, \\ v^\varepsilon = f \ \text{ on } \Gamma^\varepsilon, \\ Z^\varepsilon v^\varepsilon = Z^\varepsilon \left\{ f \overset{\Gamma^\varepsilon}{*} \dfrac{\partial G_k}{\partial \nu} + \ k^2 v^\varepsilon \overset{D^\varepsilon}{*} (\chi^\varepsilon G_k) - \nabla v^\varepsilon \overset{D^\varepsilon}{*} \nabla(\chi^\varepsilon G_k) \right\} \ \text{ on } \Sigma^\varepsilon, \end{array} \right.$$

where Z^ε stands for the boundary operator $(\partial/\partial\nu + i/\varepsilon)$ on Σ^ε.

Conversely, the solution to this problem extends outside Σ^ε (by the integral representation) to the solution to (A.2) (thanks to the term involving i/ε which prevents the so-called *irregular frequencies* from being real; see [10]).

The limiting process. In order to work in a functional framework independent of ε, we perform in each subdomain D_p^ε the change of variable $\xi = (x - s_p)/\varepsilon$. By denoting $\varphi_p^\varepsilon(\xi) = v^\varepsilon(x)$ and $f_p^\varepsilon(\xi) = f(x)$, for $x \in D_p^\varepsilon$, as well as

$$G_{pq}^\varepsilon(\xi,\eta) = G_k(s_p - s_q + \varepsilon(\xi - \eta)) \text{ for } \xi \in \overline{D_p} \text{ and } \eta \in \overline{D_q},$$

problem (A.3) amounts to a family of N problems set on the domains D_p coupled by the transparent boundary conditions written on Σ_p:

$$\text{(A.4)}\quad \left\{ \begin{array}{l} \Delta \varphi_p^\varepsilon + (\varepsilon k)^2 \varphi_p^\varepsilon = 0 \ \text{ in } D_p, \\ \varphi_p^\varepsilon = f_p^\varepsilon \ \text{ on } \Gamma_p, \\ Z\varphi_p^\varepsilon = Z \displaystyle\sum_{q=1,N} \left\{ \varepsilon f_q^\varepsilon \overset{\Gamma_q}{*} \frac{\partial G_{pq}^\varepsilon}{\partial \nu} \right. \\ \qquad\qquad \left. + \ \varepsilon^3 k^2 \varphi_q^\varepsilon \overset{D_q}{*} (\chi_q G_{pq}^\varepsilon) - \varepsilon \nabla \varphi_q^\varepsilon \overset{D_q}{*} \nabla(\chi_q G_{pq}^\varepsilon) \right\} \ \text{ on } \Sigma_p, \end{array} \right.$$

where $Z = (\partial/\partial\nu + i)$ on Σ_p.

We are now able to define the formal limit of the latter problem. Let G_0 be the limit of $G_{\varepsilon k}$ when ε tends to 0, i.e., $G_0(x) = -1/(4\pi|x|)$. Notice that

$$\text{(A.5)}\quad \begin{array}{ll} G_{pq}^\varepsilon(\xi,\eta) = G_k(s_p - s_q) + O(\varepsilon) & \text{if } p \neq q, \\ G_{pp}^\varepsilon(\xi,\eta) = \varepsilon^{-1} G_0(\xi - \eta) + O(1) & \text{if } p = q, \end{array}$$

where these formulas hold uniformly in any compact subset of $\overline{D_p} \times \overline{D_q}$ which does not contain points of the diagonal when $p = q$, and can be derived with respect to ξ or η. Hence the formal limit of problem (A.4) reads as

$$\text{(A.6)}\quad \left\{ \begin{array}{l} \Delta \varphi_p^0 = 0 \ \text{ in } D_p, \\ \varphi_p^0 = f_p^0 = -e^{ik\alpha\cdot s_p} \ \text{ on } \Gamma_p, \\ Z\varphi_p^0 = Z \left\{ f_p^0 \overset{\Gamma_p}{*} \dfrac{\partial G_0}{\partial \nu} - \nabla \varphi_p^0 \overset{D_p}{*} \nabla(\chi_p G_0) \right\} \ \text{ on } \Sigma_p, \end{array} \right.$$

which correspond to a family of uncoupled problems. Each of them amounts to solving an exterior Laplace equation. More precisely, we can write that $\varphi_p^0 = -u_I^\alpha(s_p)\, V_p$, where V_p is the *static* potential solution to

$$\text{(A.7)}\quad \left\{ \begin{array}{l} \Delta V_p = 0 \ \text{ in } \mathbb{R}^3 \setminus \mathcal{O}_p, \\ V_p = 1 \ \text{ on } \Gamma_p, \\ V_p = O(1/x) \ \text{ as } |x| \to \infty. \end{array} \right.$$

Convergence. Consider the closed subspace of the usual Sobolev space $H^1(D_p)$ given by $\mathcal{H}_p = \{\psi_p \in H^1(D_p);\ \psi_p = 0 \text{ on } \Gamma_p\}$. The variational formulation of (A.4) appears as a coupled system of variational equations:

$$\begin{aligned}
&\text{Find } \varphi_p^\varepsilon \in f_p^\varepsilon + \mathcal{H}_p,\ p = 1, N, \text{ such that}\\
&\int_{D_p} \nabla\varphi_p^\varepsilon \cdot \overline{\nabla\psi_p} - (\varepsilon k)^2 \int_{D_p} \varphi_p^\varepsilon\, \overline{\psi_p} + i \int_{\Sigma_p} \varphi_p^\varepsilon\, \overline{\psi_p}\, d\sigma\\
&+ \int_{\Sigma_p} Z \left\{ \sum_{q=1,N} \varepsilon^3 k^2 \varphi_q^\varepsilon \overset{D_q}{*} (\chi_q G_{pq}^\varepsilon) - \varepsilon \nabla\varphi_q^\varepsilon \overset{D_q}{*} \nabla(\chi_q G_{pq}^\varepsilon) \right\} \overline{\psi_p}\, d\sigma\\
&= \int_{\Sigma_p} Z \left\{ \sum_{q=1,N} \varepsilon f_q^\varepsilon \overset{\Gamma_q}{*} \frac{\partial G_{pq}^\varepsilon}{\partial\nu} \right\} \overline{\psi_p}\, d\sigma \quad \forall \psi_p \in \mathcal{H}_p,\ p = 1, N.
\end{aligned}$$

Adding these equations yields the announced Fredholm equation (A.1) in the Hilbert space $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_N$ which can be equipped with the scalar product

$$(\varphi, \psi) = \sum_{p=1,N} \int_{D_p} \nabla\varphi_p \cdot \overline{\nabla\psi_p}.$$

Indeed, define $\varphi^\varepsilon = (\varphi_1^\varepsilon, \ldots, \varphi_N^\varepsilon)$, $f^\varepsilon = (f_1^\varepsilon, \ldots, f_N^\varepsilon)$ and, respectively, the operator $\mathbb{K}^\varepsilon$ defined in $\mathcal{H}$ and $g^\varepsilon \in \mathcal{H}$ by

$$\begin{aligned}
(\mathbb{K}^\varepsilon\varphi, \psi) &= \sum_{p=1,N} -(\varepsilon k)^2 \int_{D_p} \varphi_p\, \overline{\psi_p} + i \int_{\Sigma_p} \varphi_p\, \overline{\psi_p}\, d\sigma\\
&+ \int_{\Sigma_p} Z \left\{ \sum_{q=1,N} \varepsilon^3 k^2 \varphi_q \overset{D_q}{*} (\chi_q G_{pq}^\varepsilon) - \varepsilon \nabla\varphi_q \overset{D_q}{*} \nabla(\chi_q G_{pq}^\varepsilon) \right\} \overline{\psi_p}\, d\sigma\\
(g^\varepsilon, \psi) &= \sum_{p=1,N} \int_{\Sigma_p} Z \left\{ \sum_{q=1,N} \varepsilon f_q^\varepsilon \overset{\Gamma_q}{*} \frac{\partial G_{pq}^\varepsilon}{\partial\nu} \right\} \overline{\psi_p}\, d\sigma
\end{aligned}$$

for all $\varphi = (\varphi_1, \ldots, \varphi_N)$ and $\psi = (\psi_1, \ldots, \psi_N)$ in $\mathcal{H}$. Then our coupled system reads as follows:

$$\text{Find } \varphi^\varepsilon \in f^\varepsilon + \mathcal{H} \text{ such that } (I + \mathbb{K}^\varepsilon)\varphi^\varepsilon = g^\varepsilon. \tag{A.8}$$

And of course we have a similar expression of the limit problem (A.6) with

$$\begin{aligned}
(\mathbb{K}^0\varphi, \psi) &= \sum_{p=1,N} i \int_{\Sigma_p} \varphi_p\, \overline{\psi_p}\, d\sigma - \int_{\Sigma_p} Z \left\{ \nabla\varphi_p \overset{D_p}{*} \nabla(\chi_q G_0) \right\} \overline{\psi_p}\, d\sigma,\\
(g^0, \psi) &= \sum_{p=1,N} \int_{\Sigma_p} Z \left\{ f_p^0 \overset{\Gamma_p}{*} \frac{\partial G_0}{\partial\nu} \right\} \overline{\psi_p}\, d\sigma.
\end{aligned}$$

Note that the uniqueness of the solution to (A.2) (respectively, (A.7)) implies that $I + \mathbb{K}^\varepsilon$ (respectively, $I + \mathbb{K}^0$) is injective, and thus bijective thanks to the following.

LEMMA A.1. *$\mathbb{K}^\varepsilon$ defines a family of compact operators in $\mathcal{H}$ which satisfies*

$$\|\mathbb{K}^\varepsilon - \mathbb{K}^0\| = \sup_{\varphi,\psi\in\mathcal{H}\setminus\{0\}} \frac{(\mathbb{K}^\varepsilon - \mathbb{K}^0)\varphi, \psi)}{\|\varphi\|\|\psi\|} = O(\varepsilon). \tag{A.9}$$

Proof. Consider, for instance, the part of $\mathbb{K}^\varepsilon$ corresponding to the operator $\mathbb{T}^\varepsilon_{pq}$ given by

$$\begin{aligned}(\mathbb{T}^\varepsilon_{pq}\varphi,\psi) &= \int_{\Sigma_p} Z\left\{\varepsilon\nabla\varphi_q \overset{D_q}{*} \nabla(\chi_q G^\varepsilon_{pq})\right\}\overline{\psi_p}\,d\sigma \\ &= \int_{\Sigma_p}\int_{D_q} \nabla\varphi_q(\eta)\cdot\nabla_\eta(\varepsilon\chi_q(\eta)Z_\xi G^\varepsilon_{pq}(\xi-\eta))\,d\eta\,\overline{\psi_p}\,d\sigma_\xi.\end{aligned}$$

We detail the proof only for the latter; similar arguments can be used for the other terms involved in the definition of $\mathbb{K}^\varepsilon$.

The compactness of $\mathbb{T}^\varepsilon_{pq}$ can be easily deduced from that of its adjoint. Indeed, using Schwarz inequality yields

$$\begin{aligned}\|(\mathbb{T}^\varepsilon_{pq})^*\psi\| &= \sup_{\varphi\in\mathcal{H}\setminus\{0\}}\frac{(\mathbb{T}^\varepsilon_{pq}\varphi,\psi)}{\|\varphi\|} \le C^\varepsilon_{pq}\,\|\psi\|_{L^2(\Sigma_p)}, \quad \text{where} \\ C^\varepsilon_{pq} &= \left(\int_{\Sigma_p}\int_{D_q}\left\|\nabla_\eta\{\varepsilon\chi_q(\eta)Z_\xi G^\varepsilon_{pq}(\xi-\eta)\}\right\|^2 d\eta\,d\sigma_\xi\right)^{1/2}.\end{aligned}$$

But the trace operator is compact from $H^1(D_q)$ to $L^2(\Sigma_p)$, which implies the compactness of $(\mathbb{T}^\varepsilon_{pq})^*$ in $\mathcal{H}$.

If $p\neq q$, formula (A.5) shows that $C^\varepsilon_{pq}=O(\varepsilon)$, and consequently the same holds for $\|\mathbb{T}^\varepsilon_{pq}\| = \|(\mathbb{T}^\varepsilon_{pq})^*\|$. If $p=q$, the limit operator is given by

$$(\mathbb{T}^0_{pp}\varphi,\psi) = \int_{\Sigma_p} Z\left\{\nabla\varphi_p \overset{D_p}{*} \nabla(\chi_p G_0)\right\}\overline{\psi_p}\,d\sigma,$$

since (A.5) shows in this case (again by Schwarz inequality) that

$$\left|((\mathbb{T}^\varepsilon_{pp}-\mathbb{T}^0_{pp})\varphi,\psi)\right| \le \varepsilon\,C\,\|\nabla\varphi\|_{L^2(D_p)}\,\|\psi\|_{L^2(\Sigma_p)}.$$

Hence $\|\mathbb{T}^\varepsilon_{pp}-\mathbb{T}^0_{pp}\| = O(\varepsilon)$. ∎

Lemma A.1 turns our problem into one of the simplest situations of perturbation theory [11]: the use of the Neumann series readily shows that

$$\|(I+\mathbb{K}^\varepsilon)^{-1}-(I+\mathbb{K}^0)^{-1}\| = O(\varepsilon).$$

It remains to notice that $\|f^\varepsilon - f^0\|$ and $\|g^\varepsilon - g^0\|$ are also of order ε, from which we conclude that

$$\|\varphi^\varepsilon-\varphi^0\| = O(\varepsilon). \tag{A.10}$$

The scattering amplitude. Thanks to formula (2.2), the local convergence expressed by the latter result also provides the far field asymptotics. Here, using our homothetic changes of variables, the scattering amplitude reads

$$A^\varepsilon(\alpha,\beta) = \frac{-\varepsilon}{4\pi}\sum_{p=1,N}\int_{\Gamma_p}\frac{\partial}{\partial\nu}(\varphi^\varepsilon_p(\alpha)-f^\varepsilon_p(\alpha))\,\overline{f^\varepsilon_p(\beta)}\,d\gamma.$$

On one hand, (A.10) implies that $\partial\varphi^\varepsilon_p(\alpha)/\partial\nu$ tends to $\partial\varphi^0_p(\alpha)/\partial\nu = -u^\alpha_I(s_p)\,\partial V_p/\partial\nu$ in $H^{-1/2}(\Gamma_p)$ (recall that V_p is defined in (A.7)). On the other hand, $f^\varepsilon_p(\alpha)$ tends to

the constant function $f_p^0(\alpha) = -u_I^\alpha(s_p)$. Hence

$$A^\varepsilon(\alpha,\beta) \;=\; \frac{-\varepsilon}{4\pi}\sum_{p=1,N} C_p\, u_I^\alpha(s_p)\,\overline{u_I^\beta(s_p)} \;+\; O(\varepsilon^2), \quad \text{where}$$

$$C_p \;=\; \int_{\Gamma_p} \frac{\partial V_p}{\partial \nu}\, d\gamma \;=\; \int_{\mathbb{R}^3\setminus\mathcal{O}_p} |\nabla V_p|^2$$

is referred to as the *capacity* of the obstacle $\mathcal{O}_p$. Proposition 4.2 is thus proved.

Acknowledgments. The authors would like to thank the referees for their valuable comments and suggestions.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Pocketbook of Mathematical Functions, Abridged edition of Handbook of Mathematical Functions*, Verlag Harri Deutsch, Thun, 1984.
[2] C. Bardos and M. Fink, *Mathematical foundations of the time reversal mirror*, Asymptot. Anal., 29 (2002), pp. 157–182.
[3] M. Cheney, *The linear sampling method and the MUSIC algorithm*, Inverse Problems, 17 (2001), pp. 591–595.
[4] M. Cheney, D. Isaacson, and M. Lassas, *Optimal acoustic measurements*, SIAM J. Appl. Math., 61 (2001), pp. 1628–1647.
[5] D. Colton and R. Kress, *Eigenvalues of the far field operator and inverse scattering theory*, SIAM J. Math. Anal., 26 (1995), pp. 601–615.
[6] D. Colton and R. Kress, *Eigenvalues of the far field operator for the Helmholtz equation in an absorbing medium*, SIAM J. Appl. Math., 55 (1995), pp. 1724–1735.
[7] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1992.
[8] Y. Dermenjian and E. Jalade, *Behaviour of* $(-\Delta - k^2 - i0^+)^{-1}$ *outside fading obstacles, independent scattering hypothesis and applications*, Math. Methods Appl. Sci., 26 (2003), pp. 1075–1092.
[9] M. Fink, *Time-reversal in acoustics*, Contemporary Physics, 37 (1996), pp. 95–109.
[10] A. Jami and M. Lenoir, *A variational formulation for exterior problems in linear hydrodynamics*, Comput. Methods Appl. Mech. Engrg., 16 (1978), pp. 341–359.
[11] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[12] A. Kirsch, *The MUSIC-Algorithm and the factorization method in inverse scattering theory for inhomogeneous media*, Inverse Problems, 18 (2002), pp. 1025–1040.
[13] P. D. Lax and R. S. Phillips, *Scattering Theory*, Academic Press, New York, 1967.
[14] T. D. Mast, A. I. Nachman, and R. C. Waag, *Focusing and imaging using the eigenfunctions of the scattering operator*, J. Acoust. Soc. Am., 102 (1997), pp. 715–725.
[15] P. Blomgren, G. Papanicolaou, and H. Zhao, *Super-resolution in time-reversal acoustics*, J. Acoust. Soc. Am., 111 (2002), pp. 230–248.
[16] C. Prada and M. Fink, *Eigenmodes of the time-reversal operator: A solution to selective focusing in multiple-target media*, Wave Motion, 20 (1994), pp. 151–163.
[17] A. G. Ramm, *Wave scattering by small bodies*, Rep. Math. Phys., 21 (1985), pp. 69–77.
[18] A. G. Ramm, *Scattering by Obstacles*, Mathematics and its Applications, D. Reidel, Dordrecht, 1986.
[19] E. M. Stein, *Oscillatory integrals in Fourier analysis*, in Beijing Lectures in Harmonic Analysis, Ann. of Math. Stud. 112, E. M. Stein, ed., Princeton University Press, Princeton, NJ, 1986, pp. 307–355.
[20] A. C. Zaanen, *Linear Analysis. Measure and Integral, Banach and Hilbert Space Linear Integral Equations*, North–Holland, Amsterdam, 1953.

SIAM J. APPL. MATH.
Vol. 64, No. 3, pp. 1077–1097

INTERLAYER EXCHANGE COUPLING FOR FERROMAGNETS THROUGH SPACERS*

KAMEL HAMDACHE† AND MOUHCINE TILIOUA†

Abstract. We consider the interlayer exchange coupling between two ferromagnetic layers mediated by a nonmagnetic spacer layer. We adopt the theoretical model which is based on a macroscopic description using the Landau–Lifshitz equations and the Hoffmann boundary conditions characterizing the interlayer exchange coupling between the interfaces of the ferromagnetic films. An asymptotic study of such a layered system for either small or large spacer thicknesses is presented. The asymptotic problem and the boundary conditions at interfaces both for the magnetization and the magnetic field are characterized.

Key words. thin films, magnetic multilayers, bilinear coupling, Hoffmann's boundary condition, thin and thick nonmagnetic spacers

AMS subject classifications. 35D05, 78A25, 35Q60, 35B40, 82D40

DOI. 10.1137/S0036139901398916

1. Introduction.

The interlayer exchange coupling (IEC) between ferromagnetic layers separated by a nonmagnetic spacer has been the subject of intense research in the past few years both on the experimental side and on the theoretical side due to its great impact on the recording and electronics industries. In the literature, the energy density (per unit surface) used to characterize the IEC is of the form $J_1(1 - m \cdot m') + J_2(1 - (m \cdot m')^2)$, where J_1 and J_2 are parameters describing the kind and the strength of the coupling and m and m' are the magnetization vectors at the inner surfaces of the bilayer stack just facing each other. The magnetization satisfies the saturation constraint $|m(t, x)| = 1$ for all time and in any point of the domain. If J_1 dominates, then the coupling is of ferromagnet or antiferromagnet type for $J_1 > 0$ or $J_1 < 0$, respectively, see, for example, [25], [18]. The first term in the energy is called bilinear, and the second one biquadratic, coupling. We notice, by using the saturation constraint, that these energies may be written as $J_1|m - m'|^2/2$ and $J_2|m - m'|^2|m + m'|^2/4$. The coefficients J_1 and J_2 depend generally on the thickness of the magnetic layers, see, for instance, Barnaś [6], Barnaś and Bulka [5], or Grünberg and Pierce [16]. The interested reader can find complementary information in the various review papers on this subject which have been published recently; see, for example, Stiles [28], [29], Demokritov [14], Bruno [9], [10], Camley [11], Camley and Stamps [12], Hartmann [18], Hubert and Schäfer [23], and the references therein. We also refer for a systematic experimental study to de Vries [32]. Some numerical results can be found in a paper by Labrune and Belliard [24].

In the present work we focus attention on the case in which $J_2 = 0$. Thus we consider only the bilinear coupling. However, recent experimental measurements of J_2 in some multilayered systems predict that J_2 under certain conditions cannot be neglected [7]. In order to simplify the presentation of the model equations, we will neglect the effect of the surface anisotropy energy on thin films; we may refer to [17] for some results including this effect. We restrict ourselves to the case of

*Received by the editors November 28, 2001; accepted for publication (in revised form) October 1, 2003; published electronically April 14, 2004.
http://www.siam.org/journals/siap/64-3/39891.html
†Centre de Mathématiques Appliquées, CNRS UMR 7641 & Ecole Polytechnique, 91128 Palaiseau Cedex, France (hamdache@cmapx.polytechnique.fr, tilioua@cmapx.polytechnique.fr).

magnetic/nonmagnetic multilayered structures, and we discuss the existence theory and the asymptotic behavior of the solutions when the thickness 2ε of the cylinder representing the nonmagnetic spacer is such that ε tends to either 0 or 1. We shall see that for very small spacer thicknesses the magnetic interlayer coupling becomes important. We refer to [19] for an overview in the case where the medium is a stack of magnetic media. Nevertheless, in [33] the authors indicate the main physical and experimental reasons why it is preferable that the coupling should be through an interlayer material (metallic, nonmagnetic, semiconductor, etc.) instead of there being a direct contact between the ferromagnetic slabs.

Before proceeding with the mathematical description of the model, a few clarifying comments are needed with regard to notation. We consider $B \subset \mathbb{R}^2$ a bounded and regular open set representing the cross section of the cylinder $\Omega = B \times (-1,1)$ of $\mathbb{R}^3$. The generic point of $\mathbb{R}^3$ is denoted by $x = (\hat{x}, x_3)$ with $\hat{x} = (x_1, x_2) \in B$. We assume that a ferromagnetic material occupies the domains $\Omega_\varepsilon^- = B \times (-1, -\varepsilon)$ and $\Omega_\varepsilon^+ = B \times (\varepsilon, 1)$ separated by a nonmagnetic spacer of thickness $2\varepsilon > 0$ occupying the domain $\Omega_\varepsilon^0 = B \times (-\varepsilon, \varepsilon)$. In what follows, S^2 represents the unit sphere of $\mathbb{R}^3$, and we set $\Omega_\varepsilon = \Omega_\varepsilon^- \cup \Omega_\varepsilon^+$. The magnetization field is denoted by $M(t,x)$, which belongs to S^2 almost everywhere, and the magnetic polarization is given by $\chi(\Omega_\varepsilon)M$, where $\chi(\Omega_\varepsilon)$ is the characteristic function of Ω_ε while in Ω_ε^0 it vanishes. The motion of the magnetization field M is governed by the Landau–Lifshitz–Gilbert equations; see [1], [8], for example. We have

$$\begin{cases} \dfrac{1}{1+\alpha^2}\Big(\partial_t M - \alpha M \times \partial_t M\Big) = -M \times \mathcal{H}(M) \quad \text{in } \mathbb{R}^+ \times \Omega_\varepsilon, \\ M(0,x) = M_0(x) \quad \text{in } \Omega_\varepsilon, \qquad \dfrac{\partial M}{\partial n} = 0 \quad \text{on } \partial\Omega_\varepsilon \setminus \{x_3 = \pm\varepsilon\}, \end{cases} \tag{1.1}$$

where the symbol $\times$ denotes the vector cross product in $\mathbb{R}^3$. The parameter $\alpha > 0$ depends on the gyromagnetic parameter $\zeta > 0$ and the phenomenological parameter $\beta > 0$. Indeed, the usual Landau–Lifshitz equation $\partial_t M = -\zeta M \times \mathcal{H} - \beta M \times (M \times \mathcal{H})$ may be written in the Landau–Lifshitz–Gilbert equivalent form $\partial_t M = \alpha M \times \partial_t M - \alpha_2 M \times \mathcal{H}$ with $\alpha = \beta/\zeta$ and $\alpha_2 = \zeta(1+\alpha^2)$. In (1.1) we set $\zeta = 1$. The initial magnetization M_0 satisfies the condition $|M_0(x)|^2 = 1$ for almost every $x \in \Omega_\varepsilon$, and the total magnetic excitation $\mathcal{H}(M)$ is given by

$$\mathcal{H}(M) = \operatorname{div}(A \operatorname{grad} M) + \psi(M) + \operatorname{grad}\varphi + H_0. \tag{1.2}$$

The term $M \times \operatorname{div}(A \operatorname{grad} M)$ is well defined if M is regular. We may write it, using the saturation condition $|M(t,x)|^2 = 1$ a.e., in the weaker form $\operatorname{div}(M \times A \operatorname{grad} M)$, which is well defined if $\operatorname{grad} M$ belongs to $\mathbb{L}^2(\Omega)$ with respect to the space variable. The first term on the right-hand side of (1.2) is called the exchange magnetic field, where A is the exchange variable coefficient satisfying the usual ellipticity condition in Ω_ε. The second term is the bulk anisotropy field (which generally is taken as linear with respect to M), and $\operatorname{grad}\varphi$ is the demagnetizing field, satisfying in the magnetostatic approximation of the Maxwell equations

$$\begin{cases} \operatorname{div}(\operatorname{grad}\varphi + \chi(\Omega_\varepsilon)M) = 0 \quad \text{for } x_3 > -1, \\ \dfrac{\partial \varphi}{\partial x_3} + \chi(B) M \cdot \mathbf{u}_3 = 0 \quad \text{on } x_3 = -1, \end{cases} \tag{1.3}$$

where we have set $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ to represent the canonical basis of $\mathbb{R}^3$. Note that the potential φ satisfies the transmission boundary conditions at the interfaces $x_3 = \pm\varepsilon$

and $x_3 = 1$. Indeed, (1.3) is set in the whole space $\mathbb{R}^3$, but if one assumes that our material is set on a substrate represented by the half-space $x_3 < -1$ with infinite conductivity, then the boundary condition satisfied by the potential φ is the one given in (1.3). The transmission boundary conditions mean that φ and $\frac{\partial \varphi}{\partial x_3} + \chi(B)M \cdot \mathbf{u}_3$ are continuous across the interfaces $x_3 = -\varepsilon$, $x_3 = \varepsilon$, and $x_3 = 1$. The last term H_0 on the right-hand side of (1.2) is the applied magnetic field. In what follows, without loss of generality and since we are dealing with the dynamic of the magnetization field, we assume $H_0 = 0$. Equations (1.1) are supplemented by coupling conditions for M at the interfaces $x_3 = \pm\varepsilon$. These conditions couple the layers of the domains Ω_ε^+ and Ω_ε^-. This is given by the so-called Hoffmann interlayer exchange coupling law [20], [21], which can be written as follows:

$$M(\pm\varepsilon) \times \left(\mp A \frac{\partial M(\pm\varepsilon)}{\partial x_3} - JM(\mp\varepsilon) \right) = 0. \tag{1.4}$$

The constant J (previously denoted J_1) is the interlayer exchange constant, which may depend on the thickness 2ε of the nonmagnetic spacer. This boundary condition takes the form

$$\mp A \frac{\partial M(\pm\varepsilon)}{\partial x_3} - JM(\mp\varepsilon) + J(M(\pm\varepsilon) \cdot M(\mp\varepsilon))M(\pm\varepsilon) = 0. \tag{1.5}$$

We wish to point out that the condition (1.4) is also called the modified Rado–Weertman boundary condition. It has been reported by Hoffmann [20]; see also [22], [27].

This paper is organized as follows. In section 2, we discuss the existence of global solutions of the Landau–Lifshitz equation (1.1)–(1.3) with Hoffmann boundary condition (1.4). We follow the classical proof given by [3], [31]. The main difference is related to the interlayer coupling boundary condition satisfied in our model. We give a priori estimates for the solutions and for the energy equality satisfied by the solutions, which plays a crucial role in the next sections.

Section 3 deals with the asymptotic behavior of solutions when the thickness parameter ε tends either to 0 or 1. Let us specify the results obtained. We first introduce the changes of variable which transform the domains $\Omega_\varepsilon^\pm$ and Ω_ε^0 into the domains $\Omega^\pm$ and Ω^0, which are independent of ε, but (1.1)–(1.3) become equations with variable coefficients depending on ε; see (3.9), (3.10), and (3.11). We then analyze the behavior of the model when the thickness parameter $\varepsilon \to 0$. For this behavior we assume that the coefficient J is independent of ε. Passing to the limit in the potential equation, we show that the limit potential satisfies a magnetostatic Maxwell equation set in $\mathbb{R}^2 \times (-1, -1/2)$ and $\mathbb{R}^2 \times (1/2, \infty)$. The solution satisfies a new boundary condition with regularizing effect. This boundary condition couples the interfaces $z = 1/2$ and $z = -1/2$; see Theorem 3.3. In the domains $B \times (-1, -1/2)$ and $B \times (1/2, 1)$ the magnetization field satisfies the Landau–Lifshitz equations together with the Hoffmann interlayer exchange coupling law at the interfaces $z = -1/2$ and $z = 1/2$; see Theorem 3.4.

The second purpose of section 3 is to examine the behavior of thick spacers corresponding to the case where $\varepsilon \to 1$. This means that the thickness $1 - \varepsilon$ of the two ferromagnetic domains tends to 0, and both domains are coupled with a very weak IEC. In this case the coefficient appearing in front of the interlayer exchange energy is $-J/2(1 - \varepsilon)$. To get uniform bounds for the solutions of the problem we assume that the interlayer parameter J is of order $1 - \varepsilon$. More precisely we assume

that $J = (1-\varepsilon)j$, where $j > 0$ is independent of ε. Hence, when $\varepsilon \to 1$, we get two magnetization fields m^+ and m^- satisfying the same initial data and solving Landau–Lifshitz equations in the cross section B of the domain with the effective magnetic field

$$\mathcal{H}^{\pm}(m^{\pm}) = \widehat{\operatorname{div}}(a^{\pm}\widehat{\operatorname{grad}}\, m^{\pm}) + \psi(m^{\pm}) + \widehat{\operatorname{grad}}\,\phi^{\pm} - \chi(B)m^{\pm}\cdot \mathbf{u}_3\mathbf{u}_3 + jm^{\mp}.$$

The potential $\phi^{\pm}$ is given by $\phi^{\pm}(\hat{x}) = \phi(\hat{x}, \pm 1/2)$, where $\phi(\hat{x}, z)$ solves in each slab S^{∞} and S^0 a new magnetostatic equation; see Theorem 3.7. The effect induced by the thin layer behavior for each magnetization is the same as that described first by Gioia and James [15] and later by Ammari, Halpern, and Hamdache [4]; Hamdache and Tilioua [17]; Alicandro and Leone [2] in different frameworks. The Gioia–James effect takes the form $-\chi(B)m^{\pm}\cdot \mathbf{u}_3\mathbf{u}_3$, which penalizes the out-of-plane component of the demagnetizing term $\widehat{\operatorname{grad}}\,\phi^{\pm}$. Moreover, in our result the potential $\phi^{\pm}$ is obtained as the trace at the interface $z = \pm 1/2$ of the potential ϕ. The second effect appearing is due to the Hoffmann IEC law. It takes the form of a coupling bulk anisotropy energy and links the two ferromagnetic films. The magnetizations m^+ and m^- satisfy Landau–Lifshitz equations in the domain B and are associated with the same initial data $m_0(\hat{x})$ and with the magnetic excitation $\mathcal{H}^{\pm}$ containing the bulk anisotropy field $-jm^{\mp}$ due to the interlayer exchange coupling law; see Theorem 3.9. We finally conclude this work with some remarks and comments.

Throughout, we use the following notation: $\mathbb{L}^2(\Omega) = (L^2(\Omega))^3$ an $\mathbb{H}^1(\Omega) = (H^1(\Omega))^3$ are the usual Hilbert spaces equipped with the norm $|\cdot|$ and $\|\cdot\|$, respectively. We denote by $(\cdot;\cdot)$ the scalar product of $\mathbb{L}^2(\Omega)$.

2. Global existence results. In this section, we omit the parameter ε where it is not necessary. We shall solve problem (1.1)–(1.4) by using the nonlinear Galerkin method; see, for example, Lions [26], Tartar [30].

We introduce the following spectral problem: find $(\lambda, U) \in \mathbb{R}^+ \times \mathbb{H}^1(\Omega)$ such that

$$\begin{cases} -\operatorname{div}(A\operatorname{grad} U) + \gamma U = \lambda U \quad \text{in } \Omega, \\ \\ \mp A\dfrac{\partial U(\pm\varepsilon)}{\partial x_3} - JU(\mp\varepsilon) = 0 \quad \text{in } B, \\ \\ A\dfrac{\partial U}{\partial n} = 0 \quad \text{on } \partial\Omega^{\pm} \setminus \{x_3 = \pm\varepsilon\}, \end{cases} \tag{2.1}$$

where $\gamma > 0$ is sufficiently large and $\Omega = \Omega_{\varepsilon}^- \cup \Omega_{\varepsilon}^+$. Recall that the coefficient $A(x)$ satisfies the ellipticity condition

$$0 < a_1 \le A(x) \le a_2 \quad \text{for almost every } x \in \Omega, \tag{2.2}$$

and the parameter J is assumed to be independent of ε and such that

$$J > 0. \tag{2.3}$$

The weak formulation of the problem (2.1) takes, for all $V \in \mathbb{H}^1(\Omega)$, the form

$$\begin{cases} a_{\gamma}(U,V) + \mathcal{I}(U,V) = \lambda(U;V), \text{ with} \\ \\ a_{\gamma}(U,V) = \displaystyle\int_{\Omega} A\operatorname{grad} U \operatorname{grad} V \; dx + \gamma \int_{\Omega} U\cdot V \; dx, \\ \\ \mathcal{I}(U,V) = -J\displaystyle\int_B \Big(U(-\varepsilon)\cdot V(\varepsilon) \; + U(\varepsilon)\cdot V(-\varepsilon)\Big)\, d\hat{x}, \end{cases} \tag{2.4}$$

where $\operatorname{grad} U \operatorname{grad} V = \partial_{x_j} U_i \partial_{x_j} V_i$. A classical estimate shows that for all $U \in \mathbb{H}^1(\Omega)$ we have

$$|\mathcal{I}(U,U)| \le J(\beta|\operatorname{grad} U|^2 + C_\beta |U|^2) \tag{2.5}$$

for all $\beta > 0$ with $C_\beta \to +\infty$ as $\beta \to 0$. We choose $\beta > 0$ and $\gamma > 0$ in order to have $c_1 := a_1 - J\beta > 0$ and $c_2 := \gamma - JC_\beta > 0$. Hence, we get, for all $U \in \mathbb{H}^1(\Omega)$, the estimate

$$a_\gamma(U,U) + \mathcal{I}(U,U) \ge c_1 |\operatorname{grad} U|^2 + c_2 |U|^2. \tag{2.6}$$

Notice that c_1 and c_2 are independent of ε. Moreover, for U and V in $\mathbb{H}^1(\Omega)$ we have for all $\beta > 0$ the following inequality:

$$|\mathcal{I}(U,V)| \le J(\beta|\operatorname{grad} U|^2 + C_\beta |U|^2)^{1/2} (\beta|\operatorname{grad} V|^2 + C_\beta |V|^2)^{1/2}. \tag{2.7}$$

Fixing $\beta > 0$, there exists $c_3 > 0$ independent of ε such that for all $U, V \in \mathbb{H}^1(\Omega)$ we have

$$|a_\gamma(U,V) + \mathcal{I}(U,V)| \le c_3 \|U\| \|V\|. \tag{2.8}$$

The bilinear form $a_\gamma(U,V) + \mathcal{I}(U,V)$ is continuous on $\mathbb{H}^1(\Omega) \times \mathbb{H}^1(\Omega)$ and coercive on $\mathbb{H}^1(\Omega)$ for γ large. Let $\mathcal{A}_{int} : \mathbb{H}^1(\Omega) \to \mathbb{H}^1(\Omega)'$ be the linear operator defined by

$$\langle \mathcal{A}_{int} U, V \rangle_{\mathbb{H}^1(\Omega)' \times \mathbb{H}^1(\Omega)} = a_\gamma(U,V) + \mathcal{I}(U,V),$$

where $\mathbb{H}^1(\Omega)'$ is the dual space of $\mathbb{H}^1(\Omega)$. Then $\mathcal{A}_{int}$ is an isomorphism from $\mathbb{H}^1(\Omega)$ into $\mathbb{H}^1(\Omega)'$, and its inverse $\mathcal{A}_{int}^{-1}$ is a compact operator from $\mathbb{L}^2(\Omega)$ into $\mathbb{L}^2(\Omega)$. Since the bilinear form $a_\gamma(U,V) + \mathcal{I}(U,V)$ is symmetric, the operator $\mathcal{A}_{int}$ also is, and we have

$$\langle \mathcal{A}_{int} U, V \rangle_{\mathbb{H}^1(\Omega)' \times \mathbb{H}^1(\Omega)} = \langle U, \mathcal{A}_{int} V \rangle_{\mathbb{H}^1(\Omega) \times \mathbb{H}^1(\Omega)'}$$

for all $U, V \in \mathbb{H}^1(\Omega)$. Hence, it follows that operator $\mathcal{A}_{int}$ is self-adjoint on its natural domain $D(\mathcal{A}_{int})$. We obtain the following result.

LEMMA 2.1. *There exists an uncountable set of solutions* $(\lambda_k, U_k)_{k\in\mathbb{N}} \subset \mathbb{R}^+ \times \mathbb{H}^1(\Omega)$ *of problem* (2.1) *such that* $c_2 < \lambda_k \to +\infty$ *as* $k \to +\infty$. *Moreover, the eigenvectors* U_k *associated to the eigenvalues* λ_k *form an orthonormal basis of* $\mathbb{L}^2(\Omega)$, *and* $\lambda_k^{-1/2} U_k$ *an orthogonal basis of* $\mathbb{H}^1(\Omega)$.

In the remainder we proceed along the lines of the global existence proof given by Visintin [31] and Alouges and Soyeur [3]; see also Carbou and Fabrie [13]. We briefly describe the main steps of the proof. First we introduce the penalized problem

$$\begin{cases} \dfrac{1}{1+\alpha^2}\Big(\alpha \partial_t M + M \times \partial_t M\Big) = \mathcal{H}(M) + \nu M\Big(1 - |M|^2\Big) & \text{in } \mathbb{R}^+ \times \Omega_\varepsilon^\pm, \\ M(0,x) = M_0(x) \quad \text{in } \Omega_\varepsilon^\pm, \qquad \dfrac{\partial M}{\partial n} = 0 \quad \text{on } \partial\Omega_\varepsilon^\pm \setminus \{x_3 = \pm\varepsilon\}, \\ \mp A \dfrac{\partial M(\pm\varepsilon)}{\partial x_3} - JM(\mp\varepsilon) = 0 \quad \text{in } B, \end{cases} \tag{2.9}$$

where the parameter $\nu > 0$ is fixed. To solve this problem we use the nonlinear Galerkin method (see [26], [30]) by introducing the approximated problem

$$\begin{cases} \dfrac{1}{1+\alpha^2}\Big(\alpha\partial_t M^N + M^N \times \partial_t M^N\Big) \\ \qquad\qquad = \mathcal{H}(M^N) + \nu M^N\Big(1 - |M^N|^2\Big) \quad \text{in } \mathbb{R}^+ \times \Omega_\varepsilon^\pm, \\ M^N(0,x) = M_0^N(x) \quad \text{in } \Omega_\varepsilon^\pm, \end{cases} \tag{2.10}$$

where M_0^N is the projection of M_0 in the space $\mathcal{V}^N = \{V = \sum_{k=1}^N v_k U_k\}$. The solution M^N satisfies the interlayer coupling boundary condition at the interfaces $x_3 = \pm\varepsilon$ and the homogeneous Neumann boundary condition on the remaining part of the boundary of Ω.

Let us give some remarks on the proof. Proceeding as in [31], [3], we deduce that the sequence of solutions M^N converges to M weakly-$\star$ in $L^\infty(\mathbb{R}^+, \mathbb{H}^1(\Omega))$, and $\partial_t M^N$ converges to $\partial_t M$ weakly in $L^2(\mathbb{R}^+, \mathbb{L}^2(\Omega))$. Using the strong convergence of the traces $M^N(\pm\varepsilon)$ in $\mathbb{L}^2_{\text{loc}}(\mathbb{R}^+ \times B)$, we may pass to the limit in the boundary integral $\pm J \int_{\mathbb{R}^+\times B} M^N(\pm\varepsilon) \times M^N(\mp\varepsilon)\phi \, d\hat{x}dt$, where ϕ is a test function. Using the integration by parts, we deduce that M satisfies the interlayer coupling boundary in the following sense: $\int_{\mathbb{R}^+} \langle (\mp A\partial_{x_3} M(\pm\varepsilon) - JM(\mp\varepsilon)), M(\pm\varepsilon) \times \phi\rangle_{\mathbb{H}^{-1/2}(B)\times\mathbb{H}^{1/2}(B)} \, dt = 0$. We get the following existence result.

THEOREM 2.2. *Let $M_0 \in \mathbb{H}^1(\Omega_\varepsilon)$ be such that $|M_0(x)|^2 = 1$ for almost every $x \in \Omega_\varepsilon$, and let $\psi : \mathbb{R}^3 \to \mathbb{R}^3$, $\psi(M) = \operatorname{grad}\Psi(M)$ be a continuous gradient function such that $0 \le \Psi(M) \le \Psi_\infty < \infty$ for all $M \in S^2$. Then there exists a global solution $M \in L^\infty(\mathbb{R}^+, \mathbb{H}^1(\Omega_\varepsilon))$ of problem (1.1)–(1.4) such that $|M(t,x)|^2 = 1$ for almost every $(t,x) \in \mathbb{R}^+ \times \Omega_\varepsilon$, $\partial_t M \in L^2(\mathbb{R}^+, \mathbb{L}^2(\Omega_\varepsilon))$, and $\operatorname{grad}\varphi \in L^\infty(\mathbb{R}^+, \mathbb{L}^2(\mathbb{R}^2 \times (-1,\infty)))$. Moreover, for all $t \ge 0$, the following energy equality holds:*

$$\mathcal{E}(t) + \frac{2\alpha}{1+\alpha^2}\int_0^t \int_{\Omega_\varepsilon} |\partial_t M(s)|^2 \, dxds = \mathcal{E}(0), \tag{2.11}$$

where the total energy $\mathcal{E}(t)$ is expressed by $\mathcal{E}(t) = \mathcal{E}_{exc}(t) + \mathcal{E}_{vol}(t) + \mathcal{E}_{dm}(t) + \mathcal{E}_{int}(t)$ with

$$\begin{aligned} &\mathcal{E}_{exc}(t) = \int_{\Omega_\varepsilon} A|\operatorname{grad} M|^2 \, dx, \qquad \mathcal{E}_{vol}(t) = \int_{\Omega_\varepsilon} \Psi(M) \, dx, \\ &\mathcal{E}_{int}(t) = -J\int_B M(-\varepsilon)\cdot M(\varepsilon) \, d\hat{x}, \qquad \mathcal{E}_{dm}(t) = \int_{\mathbb{R}^2}\int_{-1}^{\infty} |\operatorname{grad}\varphi|^2 \, dx. \end{aligned} \tag{2.12}$$

The magnetization M satisfies the Landau–Lifshitz–Gilbert equation in the sense of distributions where the exchange contribution is written in the weak form $\operatorname{div}(M \times A \operatorname{grad} M)$.

Remark. Notice that if we add to both sides of (2.11) the quantity $J|B|$, where $|B|$ is the Lebesgue measure of the cross section B, then the IEC energy $\mathcal{E}_{int}$ can be replaced by the right interlayer exchange energy $J\int_B (1 - M(-\varepsilon)\cdot M(\varepsilon)) \, d\hat{x} = (J/2)\int_B |M(-\varepsilon) - M(\varepsilon)|^2 \, d\hat{x}$ for all $t \ge 0$.

The initial magnetic field $\operatorname{grad}\varphi_0(x)$ satisfies the compatibility problem

$$\begin{cases} \operatorname{div}(\operatorname{grad}\varphi_0 + \chi(\Omega_\varepsilon)M_0) = 0 \ \text{ for } x_3 > -1, \\ \dfrac{\partial\varphi_0}{\partial x_3} + \chi(B)M_0 \cdot \mathbf{u}_3 = 0 \ \text{ on } x_3 = -1. \end{cases} \tag{2.13}$$

It follows that $\mathcal{E}_{dm}(0)$ is given by

$$\mathcal{E}_{dm}(0) = \int_{\mathbb{R}^2} \int_{-1}^{\infty} |\operatorname{grad} \varphi_0|^2 \, dx = - \int_{\Omega_\varepsilon} M_0 \cdot \operatorname{grad} \varphi_0 \, dx, \tag{2.14}$$

which implies that $|\operatorname{grad} \varphi_0|_{\mathbb{L}^2(\mathbb{R}^2 \times (-1,\infty))} \leq (2(1-\varepsilon)|B|)^{1/2}$, and then the initial demagnetization energy satisfies

$$0 \leq \mathcal{E}_{dm}(0) \leq 2(1-\varepsilon)|B|. \tag{2.15}$$

The initial interlayer exchange energy satisfies

$$|\mathcal{E}_{int}(0)| \leq J|B|. \tag{2.16}$$

One observes that if the initial magnetization is assumed to be such that $M_0(-\varepsilon) \cdot M_0(\varepsilon) \geq 0$, then we get $\mathcal{E}_{int}(0) \leq 0$.

The initial volume anisotropy energy satisfies

$$0 \leq \mathcal{E}_{vol}(0) \leq 2\Psi_\infty(1-\varepsilon)|B|. \tag{2.17}$$

The initial exchange energy is such that

$$0 \leq \mathcal{E}_{exc}(0) < a_2 |\operatorname{grad} M_0|^2_{\mathbb{L}^2(\Omega_\varepsilon)} \leq a_2 \delta, \tag{2.18}$$

where $\delta > 0$ is independent of ε, by assuming, for example, that M_0 is the restriction to Ω_ε of $M_0 \in \mathbb{H}^1(\Omega)$.

Hence the following uniform estimates with respect to ε hold.

COROLLARY 2.3. *Let $M_0 \in \mathbb{H}^1(\Omega)$ be such that $|M_0(x)|^2 = 1$ for almost every $x \in \Omega_\varepsilon$. Then there exists $C > 0$ independent of ε such that*

$$\begin{aligned} &\|\operatorname{grad} M^\varepsilon\|_{L^\infty(\mathbb{R}^+, \mathbb{L}^2(\Omega_\varepsilon))} + \|\partial_t M^\varepsilon\|_{L^2(\mathbb{R}^+, \mathbb{L}^2(\Omega_\varepsilon))} \\ &+ \|\operatorname{grad} \varphi^\varepsilon\|_{L^\infty(\mathbb{R}^+, \mathbb{L}^2(\mathbb{R}^2 \times (-1,\infty)))} \leq C \end{aligned} \tag{2.19}$$

and $|M^\varepsilon(t,x)|^2 = 1$ for almost every $(t,x) \in \mathbb{R}^+ \times \Omega_\varepsilon$.

In the next section we will discuss the behavior of the solutions $(M^\varepsilon, \varphi^\varepsilon)$ in the two cases $\varepsilon \to 0$ and $\varepsilon \to 1$.

3. Convergences. The nonmagnetic spacer occupies the domain Ω^0_ε. We introduce the following change of variables:

$$\begin{cases} z = x_3 \ \text{ if } x_3 \geq 1, \\ z = \dfrac{1}{2(1-\varepsilon)}(x_3 + 1 - 2\varepsilon) \in \left[\dfrac{1}{2}, 1\right] \quad \text{if } \ \varepsilon \leq x_3 \leq 1, \\ z = \dfrac{1}{2\varepsilon} x_3 \in \left[-\dfrac{1}{2}, \dfrac{1}{2}\right] \ \text{ if } \ -\varepsilon \leq x_3 \leq \varepsilon, \\ z = \dfrac{1}{2(1-\varepsilon)}(x_3 - 1 + 2\varepsilon) \in \left[-1, -\dfrac{1}{2}\right] \ \text{ if } \ -1 \leq x_3 \leq -\varepsilon. \end{cases} \tag{3.1}$$

We set

$$\begin{aligned} &G^+ = B \times \left(\frac{1}{2}, 1\right), \ G^- = B \times \left(-1, -\frac{1}{2}\right), \\ &G^0 = B \times \left(-\frac{1}{2}, \frac{1}{2}\right), \qquad G = G^+ \cup G^- \end{aligned} \tag{3.2}$$

and introduce for $t \geq 0$ and $(\hat{x}, z) \in G$ the new functions

$$m^\varepsilon(t,\hat{x},z) = \begin{cases} M(t,\hat{x},2(1-\varepsilon)z + 2\varepsilon - 1) \ \text{ in } \mathbb{R}^+ \times G^+, \\ \\ M(t,\hat{x},2(1-\varepsilon)z - 2\varepsilon + 1) \ \text{ in } \mathbb{R}^+ \times G^-. \end{cases} \tag{3.3}$$

We also define the rescaled exchange coefficient a^ε by

$$a^\varepsilon(\hat{x},z) = \begin{cases} A(\hat{x},2(1-\varepsilon)z + 2\varepsilon - 1) \ \text{ in } G^+, \\ \\ A(\hat{x},2(1-\varepsilon)z - 2\varepsilon + 1) \ \text{ in } G^-. \end{cases} \tag{3.4}$$

We assume that $A \in L^\infty(B, C^0([-1,1]))$. Note that a^ε satisfies the coerciveness condition (2.2).

We also introduce the slabs of $\mathbb{R}^3$,

$$\begin{cases} S^- = \mathbb{R}^2 \times (-1,-1/2),\ S^0 = \mathbb{R}^2 \times (-1/2,1/2),\ S^+ = \mathbb{R}^2 \times (1/2,1), \\ \\ S^\infty = \mathbb{R}^2 \times (1,\infty) \text{ and } S^{+,\infty} = \mathbb{R}^2 \times (1/2,\infty),\ S^{-,\infty} = \mathbb{R}^2 \times (-1,\infty), \end{cases} \tag{3.5}$$

and for $(t,\hat{x},z) \in \mathbb{R}^+ \times \mathbb{R}^3_+$ the functions

$$\phi^\varepsilon(t,\hat{x},z) = \begin{cases} \varphi(t,\hat{x},z) \ \text{ if } z \geq 1, \\ \\ \varphi(t,\hat{x},2(1-\varepsilon)z + 2\varepsilon - 1) \ \text{ if } \dfrac{1}{2} \leq z \leq 1, \\ \\ \varphi(t,\hat{x},2\varepsilon z) \ \text{ if } -\dfrac{1}{2} \leq z \leq \dfrac{1}{2}, \\ \\ \varphi(t,\hat{x},2(1-\varepsilon)z - 2\varepsilon + 1) \ \text{ if } -1 \leq z \leq -\dfrac{1}{2}. \end{cases} \tag{3.6}$$

We finally define $\sigma^\varepsilon(z)$ and $\nu^\varepsilon(z)$ by setting

$$\sigma^\varepsilon(z) = \begin{cases} 1 \ \text{ if } z \geq 1, \\ \\ \dfrac{1}{4(1-\varepsilon)^2} \ \text{ if } \ -1 \leq z \leq -\dfrac{1}{2} \ \text{ or } \ \dfrac{1}{2} < z < 1, \\ \\ \dfrac{1}{4\varepsilon^2} \text{ if } -\dfrac{1}{2} \leq z \leq \dfrac{1}{2} \end{cases} \tag{3.7}$$

and

$$\nu^\varepsilon(z) = \begin{cases} 0 \ \text{ if } \ -\dfrac{1}{2} < z < \dfrac{1}{2} \text{ or } z \geq 1, \\ \\ \dfrac{1}{2(1-\varepsilon)} \ \text{ if } \dfrac{1}{2} \leq z \leq 1 \text{ or } -1 \leq z \leq -\dfrac{1}{2}. \end{cases} \tag{3.8}$$

The potential ϕ^ε satisfies in $\mathbb{R}^+ \times S^{-,\infty}$ the equations

$$\begin{cases} \widehat{\operatorname{div}}\Big(\widehat{\operatorname{grad}}\,\phi^\varepsilon + \chi(G)\widehat{m}^\varepsilon\Big) + \partial_z(\sigma^\varepsilon \partial_z \phi^\varepsilon + \nu^\varepsilon \chi(G) m^\varepsilon \cdot \mathbf{u}_3) = 0, \\ \\ \dfrac{1}{2(1-\varepsilon)} \partial_z \phi^\varepsilon + \chi(B) m^\varepsilon \cdot \mathbf{u}_3 = 0 \quad \text{at } z = -1, \end{cases} \tag{3.9}$$

with the usual transmission boundary conditions at the interfaces $z = 1$ and $z = \pm 1/2$. The operators $\widehat{\mathrm{div}}$ and $\widehat{\mathrm{grad}}$ represent the divergence and the gradient operators, respectively, with respect to the variable $\hat{x}$ and $\widehat{m} = (m_1, m_2, 0)$, where (m_1, m_2) are the first two components of m. The vector $\widehat{\mathrm{grad}}\,\phi$ may also be considered as a 2-D vector or a 3-D vector where the third component is 0.

The magnetic excitation $\mathcal{H}$ defined in G by (1.2) becomes

$$\begin{aligned}\mathcal{H}^\varepsilon(m^\varepsilon) &= \widehat{\mathrm{div}}(a^\varepsilon \widehat{\mathrm{grad}}\, m^\varepsilon) + \frac{1}{4(1-\varepsilon)^2}\partial_z(a^\varepsilon \partial_z m^\varepsilon) \\ &\quad + \psi(m^\varepsilon) + \widehat{\mathrm{grad}}\,\phi^\varepsilon + \frac{1}{2(1-\varepsilon)}\partial_z \phi^\varepsilon \mathbf{u}_3,\end{aligned} \tag{3.10}$$

and the interlayer exchange boundary condition takes the form

$$m^\varepsilon(\pm 1/2) \times \left(\mp \frac{a^\varepsilon}{2(1-\varepsilon)} \partial_z m^\varepsilon(\pm 1/2) - J m^\varepsilon(\mp 1/2) \right) = 0. \tag{3.11}$$

The Landau–Lifshitz equations are rewritten in the same form as (1.1), and the effective magnetic field is denoted by $\mathcal{H}^\varepsilon(m^\varepsilon)$.

Global existence of solutions $(m^\varepsilon, \phi^\varepsilon)$ of the new system is guaranteed by Theorem 2.2 proved in section 2. We shall describe the behavior of such solutions first when $\varepsilon \to 0$ (the thin nonmagnetic case), and then when $\varepsilon \to 1$ (the thick nonmagnetic case).

3.1. Asymptotic behavior with thin nonmagnetic spacers. We investigate here the case when $\varepsilon \to 0$. The energy estimate (2.11) satisfied by the solutions becomes

$$\mathcal{E}^\varepsilon(t) + \frac{2\alpha}{1+\alpha^2} \int_0^t \int_G |\partial_t m^\varepsilon(s)|^2 \, dx ds = \mathcal{E}^\varepsilon(0), \tag{3.12}$$

where the total energy $\mathcal{E}^\varepsilon(t)$ is expressed by $\mathcal{E}^\varepsilon(t) = \mathcal{E}^\varepsilon_{exc}(t) + \mathcal{E}^\varepsilon_{vol}(t) + \mathcal{E}^\varepsilon_{dm}(t) + \mathcal{E}^\varepsilon_{int}(t)$ with

$$\begin{cases} \mathcal{E}^\varepsilon_{exc}(t) = \displaystyle\int_G a^\varepsilon |\widehat{\mathrm{grad}}\, m^\varepsilon|^2 \, dx + \frac{1}{4(1-\varepsilon)^2} \int_G a^\varepsilon |\partial_z m^\varepsilon|^2 \, dx, \\ \mathcal{E}^\varepsilon_{vol}(t) = \displaystyle\int_G \Psi(m^\varepsilon) \, dx, \\ \mathcal{E}^\varepsilon_{int}(t) = -\dfrac{J}{2(1-\varepsilon)} \displaystyle\int_B m^\varepsilon \left(-\frac{1}{2}\right) \cdot m^\varepsilon \left(\frac{1}{2}\right) \, d\hat{x}, \\ \mathcal{E}^\varepsilon_{dm}(t) = \displaystyle\int_{S^{-,\infty}} |\widehat{\mathrm{grad}}\,\phi^\varepsilon|^2 \, dx + \int_{S^{-,\infty}} \sigma^\varepsilon |\partial_z \phi^\varepsilon|^2 \, dx, \end{cases} \tag{3.13}$$

where $\sigma^\varepsilon(z)$ is defined by (3.7) and $a^\varepsilon(\hat{x}, z)$ is defined in G by (3.4).

Let us discuss the uniform boundedness with respect to ε of the initial energy $\mathcal{E}^\varepsilon(0)$. We assume that the initial magnetization m_0^ε is independent of the variable z and is such that $m_0^\varepsilon(\hat{x})$ is uniformly bounded in $\mathbb{H}^1(B)$. It follows that $\partial_z m_0^\varepsilon = 0$ and then $\mathcal{E}^\varepsilon_{exc}(0) = \int_G a^\varepsilon |\widehat{\mathrm{grad}}\, m_0^\varepsilon|^2 \, dx \le C$, where $C > 0$ is independent of ε since, by (2.2), we have $0 < a_1 \le a^\varepsilon(\hat{x}, z) \le a_2$ for almost every $(\hat{x}, z)$. We may weaken this hypothesis if we assume, for example, that $m_0^\varepsilon(\hat{x}, z) = m_0(\hat{x}, (1-\varepsilon)z)$ but, to simplify the presentation, we use our strong hypothesis on m_0^ε. Next, we have $\mathcal{E}^\varepsilon_{vol}(0) \le C$ and

$|\mathcal{E}^\varepsilon_{int}(0)| = \frac{J}{2(1-\varepsilon)}|B| \le C$, where $C > 0$ is independent of ε. It remains to consider the initial potential $\phi^\varepsilon_0 = \phi^\varepsilon_{|t=0}$, which satisfies the equation

$$\begin{cases} \widehat{\operatorname{div}}(\widehat{\operatorname{grad}}\,\phi^\varepsilon_0 + \chi(G)\widehat{m}^\varepsilon_0) + \partial_z(\sigma^\varepsilon\partial_z\phi^\varepsilon_0 + \nu^\varepsilon\chi(G)m^\varepsilon_0\cdot\mathbf{u}_3) = 0 \quad \text{in} \quad S^{-,\infty}, \\ \dfrac{1}{2(1-\varepsilon)}\partial_z\phi^\varepsilon_0 + \chi(B)m^\varepsilon_0\cdot\mathbf{u}_3 = 0 \ \text{ at } \ z=-1. \end{cases} \tag{3.14}$$

In what follows we denote by $C > 0$ various constants which are independent of ε. The energy relation associated with (3.14) is

$$\begin{aligned} &\int_{S^{-,\infty}} |\widehat{\operatorname{grad}}\,\phi^\varepsilon_0|^2\,dx + \int_{S^{-,\infty}} \sigma^\varepsilon|\partial_z\phi^\varepsilon_0|^2\,dx = \\ &\qquad = -\int_G \widehat{m}^\varepsilon_0\cdot\widehat{\operatorname{grad}}\,\phi^\varepsilon_0\,dx - \frac{1}{2(1-\varepsilon)}\int_G m^\varepsilon_0\cdot\mathbf{u}_3\partial_z\phi^\varepsilon_0\,dx, \end{aligned} \tag{3.15}$$

which, by the Cauchy–Schwarz inequality, gives the estimate

$$\mathcal{E}^\varepsilon_{dm}(0) \le 2|G|. \tag{3.16}$$

The above results can be stated as follows.

LEMMA 3.1. *Let $m^\varepsilon_0(\hat{x},z) = m^\varepsilon_0(\hat{x}) \in \mathbb{H}^1(B)$ such that $\|m^\varepsilon_0\|_{\mathbb{H}^1(B)} \le C$ and $|m^\varepsilon_0(\hat{x})|^2 = 1$ a.e. Then there exists a $C > 0$ independent of ε such that the initial energy $\mathcal{E}^\varepsilon(0) = \mathcal{E}^\varepsilon_0$ satisfies the estimate*

$$0 \le \mathcal{E}^\varepsilon_0 \le C. \tag{3.17}$$

From (3.12) and Lemma 3.1 we deduce the following uniform bounds.

LEMMA 3.2. *Under the hypotheses of Lemma* 3.1, *it holds that*

$$\begin{cases} |\operatorname{grad} m^\varepsilon|_{L^\infty(\mathbb{R}^+,\mathbb{L}^2(G))} + |\partial_t m^\varepsilon|_{L^2(\mathbb{R}^+,\mathbb{L}^2(G))} \le C, \\ |\widehat{\operatorname{grad}}\,\phi^\varepsilon|_{L^\infty(\mathbb{R}^+,\mathbb{L}^2(S^{-,\infty}))} + |\partial_z\phi^\varepsilon|_{L^\infty(\mathbb{R}^+,L^2(S^{-,\infty}\setminus S^0))} \le C, \\ |\partial_z\phi^\varepsilon|_{L^\infty(\mathbb{R}^+,L^2(S^0))} \le C\ \varepsilon. \end{cases} \tag{3.18}$$

For a subsequence still denoted $(m^\varepsilon,\phi^\varepsilon)$ we deduce the convergences

$$\begin{cases} m^\varepsilon \rightharpoonup m \text{ weakly}-\star \ \text{ in } L^\infty(\mathbb{R}^+,\mathbb{H}^1(G))\cap\mathbb{L}^\infty(\mathbb{R}^+\times G), \\ \partial_t m^\varepsilon \rightharpoonup \partial_t m \text{ weakly } \ \text{in } L^2(\mathbb{R}^+,\mathbb{L}^2(G)), \\ m^\varepsilon \to m \text{ strongly } \ \text{in } L^2_{\text{loc}}(\mathbb{R}^+,\mathbb{L}^2(G)); \end{cases} \tag{3.19}$$

the last strong convergence is a consequence of the classical use of Aubin's compactness lemma. For the magnetic field we have

$$\begin{cases} \widehat{\operatorname{grad}}\,\phi^\varepsilon \rightharpoonup \widehat{\operatorname{grad}}\,\phi \text{ weakly}-\star \text{ in } L^\infty(\mathbb{R}^+,\mathbb{L}^2(S^{-,\infty})), \\ \partial_z\phi^\varepsilon \rightharpoonup \partial_z\phi \text{ weakly}-\star \ \text{ in } L^\infty(\mathbb{R}^+,L^2(S^{-,\infty}\setminus S^0)), \\ \partial_z\phi^\varepsilon \to \partial_z\phi = 0 \text{ strongly } \ \text{in } L^\infty(\mathbb{R}^+,L^2(S^0)), \\ \Theta^\varepsilon := \dfrac{1}{2\varepsilon}\partial_z\phi^\varepsilon \rightharpoonup 0 \text{ weakly}-\star \ \text{ in } L^\infty(\mathbb{R}^+,L^2(S^0)). \end{cases} \tag{3.20}$$

Indeed, the potential ϕ^ε converges weakly-$\star$ to ϕ in $L^\infty(\mathbb{R}^+, H^1_\rho(S^{-,\infty}))$, where $H^1_\rho(S^{-,\infty}) = \{g \in \mathcal{D}'(S^{-,\infty}),\ \operatorname{grad} g \in \mathbb{L}^2(S^{-,\infty}), \text{and } \rho g \in L^2(S^{-,\infty})\}$ denotes the weighted Sobolev space with $\rho(x) = (1+|\hat{x}|+|z|)^{-1}$. This space describes the behavior, when $|x| \to \infty$, of the potentials ϕ^ε and ϕ. In particular, ϕ^ε converges to ϕ weakly-$\star$ in $L^\infty(\mathbb{R}^+, L^2_{\rm loc}(S^{-,\infty}))$.

We prove the last assertion of (3.20). We multiply (3.9) first by 2ε then by $g \in \mathcal{D}(\mathcal{O})$ with $\mathcal{O} = \mathbb{R}^+ \times S^{-,\infty}$. Integrating by parts, we get

$$\begin{aligned} 2\varepsilon \left(\int_{\mathcal{O}} \widehat{\operatorname{grad}}\, \phi^\varepsilon \widehat{\operatorname{grad}}\, g \; dxdt + \int_{\mathcal{O}\setminus(\mathbb{R}^+\times S^0)} \sigma^\varepsilon \partial_z \phi^\varepsilon \partial_z g \; dxdt \right) + \int_{\mathbb{R}^+\times S^0} \Theta^\varepsilon \partial_z g \; dxdt \\ = -2\varepsilon \int_{\mathbb{R}^+\times G} \Big(\widehat{m}^\varepsilon \cdot \widehat{\operatorname{grad}}\, g + \nu^\varepsilon m^\varepsilon \cdot \mathbf{u}_3 \partial_z g \Big) dxdt. \end{aligned} \tag{3.21}$$

Hence, passing to the limit, we deduce that the weak-$\star$ limit Θ of the sequence Θ^ε satisfies $\partial_z \Theta = 0$. Then Θ is independent of the variable z, and since $\Theta \in L^\infty(\mathbb{R}^+, \mathbb{L}^2(S^0))$ and $S^0 = \mathbb{R}^2 \times (-1/2, 1/2)$, we get $\Theta = 0$. The following theorem gives a characterization of the potential ϕ.

THEOREM 3.3. *Let ϕ be the weak-$\star$ limit in $L^\infty(\mathbb{R}^+, H^1_\rho(S^{-,\infty}))$ of the sequence ϕ^ε. Let ϕ^+ and ϕ^- be the restrictions of ϕ to $\mathbb{R}^+ \times S^{+,\infty}$ and $\mathbb{R}^+ \times S^-$, respectively. Then $(\phi^+, \phi^-) \in L^\infty(\mathbb{R}^+, H^1_\rho(S^{+,\infty}) \times H^1_\rho(S^-))$ satisfies the coupled magnetostatic equations*

$$\begin{cases} \widehat{\operatorname{div}} \left(\widehat{\operatorname{grad}}\, \phi^+ + \chi(G^+) \widehat{m}^+ \right) + \partial_z \left(\sigma(z) \partial_z \phi^+ + \nu(z) \chi(G^+) m^+ \cdot \mathbf{u}_3 \right) = 0 \text{ in } \mathbb{R}^+ \times S^{+,\infty}, \\ \widehat{\operatorname{div}} \left(\widehat{\operatorname{grad}}\, \phi^- + \chi(G^-) \widehat{m}^- \right) + \partial_z \left(\dfrac{1}{4} \partial_z \phi^- + \dfrac{1}{2} \chi(G^-) m^- \cdot \mathbf{u}_3 \right) = 0 \ \text{ in } \mathbb{R}^+ \times S^-, \\ \partial_z \phi^+(1^-) + 2\chi(B) m(1^-) \cdot \mathbf{u}_3 = 4 \partial_z \phi^+(1^+), \\ \partial_z \phi^-(-1) + 2\chi(B) m^-(-1) \cdot \mathbf{u}_3 = 0 \end{cases} \tag{3.22}$$

and the coupling relations

$$\begin{cases} \phi^-(-1/2) = \phi^+(1/2), \\ (\partial_z \phi^+ + 2\chi(B) m^+ \cdot \mathbf{u}_3)_{|z=1/2} - (\partial_z \phi^- + 2\chi(B) m^- \cdot \mathbf{u}_3)_{|z=-1/2} = -4\ \widehat{\Delta} \phi^+(1/2), \end{cases} \tag{3.23}$$

where $m^\pm$ is the restriction of m to $G^\pm$. The restriction ϕ^0 of ϕ to $\mathbb{R}^+ \times S^0$ is independent of the variable z and is given by

$$\phi^0(t, \hat{x}) = \phi^-(t, \hat{x}, -1/2) = \phi^+(t, \hat{x}, 1/2) \text{ a.e.} \tag{3.24}$$

Proof. In $\mathbb{R}^+ \times S^0$ the function Θ^ε defined in (3.20) satisfies

$$\partial_z \Theta^\varepsilon = -2\varepsilon \widehat{\Delta} \phi^\varepsilon. \tag{3.25}$$

It follows that $\partial_z \Theta^\varepsilon$ is uniformly bounded in $L^\infty(\mathbb{R}^+, L^2(-1/2, 1/2; H^{-1}_{\rm loc}(\mathbb{R}^2)))$. Then Θ^ε belongs to the space $L^\infty(\mathbb{R}^+, H^1(-1/2, 1/2; H^{-1}_{\rm loc}(\mathbb{R}^2)))$. Using the trace theorems, it follows that $\Theta^\varepsilon(\pm 1/2^\pm)$ is well defined in $L^\infty(\mathbb{R}^+, H^{-1}_{\rm loc}(\mathbb{R}^2))$. Moreover, thanks to the convergences (3.19)–(3.20), we deduce that $\Theta^\varepsilon \rightharpoonup 0$ weakly-$\star$ in

$L^\infty(\mathbb{R}^+, H^1(-1/2, 1/2; H^{-1}_{\rm loc}(\mathbb{R}^2)))$. Consequently, we obtain $\Theta^\varepsilon(\pm 1/2^\pm) \rightharpoonup 0$ weakly-$\star$. Hence, we get, at least in the sense of distributions, the following convergence:

$$\frac{1}{2\varepsilon}\partial_z\phi^\varepsilon(\pm 1/2^\pm) \rightharpoonup 0. \tag{3.26}$$

Let us consider problem (3.9). Using test functions $g(t,\hat{x},z)$ first in $\mathcal{D}(\mathbb{R}^+\times S^{+,\infty})$ then in $\mathcal{D}(\mathbb{R}^+ \times S^-)$ and taking into account the convergences of (3.19)–(3.20), it is easy to see that ϕ^+ and ϕ^- satisfy, in the sense of distributions, the following equations:

$$\widehat{\rm div}\left(\widehat{\rm grad}\,\phi^+ + \chi(G^+)\widehat{m}^+\right) + \partial_z\left(\sigma(z)\partial_z\phi^+ + \nu(z)\chi(G^+)m^+\cdot\mathbf{u}_3\right) = 0 \tag{3.27}$$

in $\mathbb{R}^+ \times S^{+,\infty}$, where $\sigma(z) = 1$, $\nu(z) = 0$ if $z > 1$ and $\sigma(z) = 1/4$, $\nu(z) = 1/2$ if $1/2 < z < 1$, and

$$\begin{aligned}\widehat{\rm div}\left(\widehat{\rm grad}\,\phi^- + \chi(G^-)\widehat{m}^-\right) + \partial_z\left(\frac{1}{4}\partial_z\phi^- + \frac{1}{2}\chi(G^-)m^-\cdot\mathbf{u}_3\right) = 0 \ \text{ in } \mathbb{R}^+\times S^-,\\ \partial_z\phi^- + 2\chi(B)m^-\cdot\mathbf{u}_3 = 0 \quad \text{at } z=-1.\end{aligned}$$

It remains to couple this set of equations. Let ϕ^0 be the restriction of ϕ to $\mathbb{R}^+\times S^0$. We have shown that ϕ^0 is independent of the variable z in the domain S^0. The transmission boundary conditions satisfied by ϕ^ε at the interfaces $z = \pm 1/2$ and $z = 1$ are the following:

$$\begin{aligned}&\left[\sigma^\varepsilon\partial_z\phi^\varepsilon + \nu^\varepsilon\chi(B)m^\varepsilon\cdot\mathbf{u}_3\right]_{z=\pm 1/2} = \left[\sigma^\varepsilon\partial_z\phi^\varepsilon + \nu^\varepsilon\chi(B)m^\varepsilon\cdot\mathbf{u}_3\right]_{z=1} = 0,\\ &\left[\phi^\varepsilon\right]_{z=\pm 1/2} = \left[\phi^\varepsilon\right]_{z=1} = 0.\end{aligned} \tag{3.28}$$

In (3.28), [ ] denotes the jump across the interfaces. Passing to the limit in the continuity condition of ϕ^ε at the interfaces $z = \pm 1/2$, we get the result

$$\phi^0(t,\hat{x}) = \phi^-(t,\hat{x},-1/2) = \phi^+(t,\hat{x},1/2) \text{ a.e.} \tag{3.29}$$

Since ϕ^0 is independent of the variable z in the domain $\mathbb{R}^+\times S^0$, we use, in the weak formulation of (3.9), test functions $g \in \mathcal{D}(\mathbb{R}^+\times\mathbb{R}^2)$, which are independent of the variable z. An integration by parts gives

$$\begin{aligned}&\int_{\mathbb{R}^+\times S^0}\widehat{\rm grad}\,\phi^\varepsilon\widehat{\rm grad}\,g\ dxdt\\ &= \int_{\mathbb{R}^+\times\mathbb{R}^2}\left(\frac{1}{4\varepsilon^2}\partial_z\phi^\varepsilon(1/2^-) - \frac{1}{4\varepsilon^2}\partial_z\phi^\varepsilon(-1/2^+)\right)g(t,\hat{x})\ d\hat{x}dt.\end{aligned} \tag{3.30}$$

From (3.28) we can write that

$$\begin{aligned}\frac{1}{\varepsilon^2}\partial_z\phi^\varepsilon(1/2^-) &= \frac{1}{(1-\varepsilon)^2}\partial_z\phi^\varepsilon(1/2^+) + \frac{2}{(1-\varepsilon)}\chi(B)m^\varepsilon(1/2^+)\cdot\mathbf{u}_3,\\ \frac{1}{\varepsilon^2}\partial_z\phi^\varepsilon(-1/2^+) &= \frac{1}{(1-\varepsilon)^2}\partial_z\phi^\varepsilon(-1/2^-) + \frac{2}{(1-\varepsilon)}\chi(B)m^\varepsilon(-1/2^-)\cdot\mathbf{u}_3;\end{aligned} \tag{3.31}$$

then we replace (3.31) in (3.30). We pass to the limit in (3.30) by using the fact that $\partial_z\phi^\varepsilon(\pm 1/2^\pm)$ converges to $\partial_z\phi^\pm(\pm 1/2^\pm)$ in the sense of distributions. We get the

equation

$$\int_{\mathbb{R}^+\times\mathbb{R}^2} \widehat{\operatorname{grad}}\,\phi^0\ \widehat{\operatorname{grad}}\,g\ d\hat{x}dt = \int_{\mathbb{R}^+\times\mathbb{R}^2} \left[\left(\frac{1}{4}\partial_z\phi^+(1/2) + \frac{1}{2}\chi(B)m^+(1/2)\cdot\mathbf{u}_3\right)\right.$$
$$\left. - \left(\frac{1}{4}\partial_z\phi^-(-1/2) + \frac{1}{2}\chi(B)m^-(-1/2)\cdot\mathbf{u}_3\right)\right] g\ d\hat{x}dt. \tag{3.32}$$

This result implies, after an integration by parts, the equation

$$\begin{cases} 4\widehat{\Delta}\phi^0 + \Big[(\partial_z\phi^+(1/2) + 2\chi(B)m^+(1/2)) \\ \qquad -(\partial_z\phi^-(-1/2) + 2\chi(B)m^-(-1/2)\cdot\mathbf{u}_3)\Big] = 0 \ \text{ in } \mathbb{R}^+\times\mathbb{R}^2. \end{cases} \tag{3.33}$$

Finally, using (3.29), we get the coupling boundary condition

$$(\partial_z\phi^+(1/2) + 2\chi(B)m^+(1/2)) - (\partial_z\phi^-(-1/2) + 2\chi(B)m(-1/2)\cdot\mathbf{u}_3)$$
$$= -4\widehat{\Delta}\phi^+(1/2) = -4\widehat{\Delta}\phi^-(-1/2). \tag{3.34}$$

We conclude the proof of the theorem by passing to the limit in the transmission boundary condition at the interface $z = 1$. We get

$$\partial_z\phi^+(1^-) + 2\chi(B)m(1^-)\cdot\mathbf{u}_3 = 4\partial_z\phi^+(1^+). \tag{3.35}$$

The proof of the theorem is then complete. □

We now are able to pass to the limit in Landau–Lifshitz equations. Let $F \in \mathcal{D}(Q)$ be a test function where $Q = \mathbb{R}^+\times\Omega$ and $\Omega = B\times(-1,1)$. The following weak formulation is associated with Landau–Lifshitz equations:

$$\begin{cases} \dfrac{1}{1+\alpha^2}\displaystyle\int_Q \chi(G)(\partial_t m^\varepsilon - \alpha m^\varepsilon\times\partial_t m^\varepsilon)F\ dxdt \\ = \displaystyle\int_Q \chi(G)m^\varepsilon\times a^\varepsilon\widehat{\operatorname{grad}}\,m^\varepsilon\widehat{\operatorname{grad}}\,F\ dxdt \\ + \dfrac{1}{4(1-\varepsilon)^2}\displaystyle\int_Q \chi(G)m^\varepsilon\times a^\varepsilon\partial_z m^\varepsilon\partial_z F\ dxdt \\ - \displaystyle\int_Q \chi(G)m^\varepsilon\times\Big(\psi(m^\varepsilon) + \widehat{\operatorname{grad}}\,\phi^\varepsilon + \dfrac{1}{2(1-\varepsilon)}\partial_z\phi^\varepsilon\mathbf{u}_3\Big)F\ dxdt \\ - \dfrac{J}{2(1-\varepsilon)}\displaystyle\int_{\mathbb{R}^+\times B} m^\varepsilon(1/2^+)\times m^\varepsilon(-1/2^-)\Big(F(1/2^+) - F(-1/2^-)\Big)\ d\hat{x}dt, \end{cases} \tag{3.36}$$

where we set $G = G^+\cup G^-$. By the definition of a^ε (see (3.4)) we have the convergence

$$a^\varepsilon \to a \quad \text{a.e.}, \tag{3.37}$$

where

$$a(\hat{x},z) = \begin{cases} A(\hat{x},2z+1) & \text{in } G^-, \\ A(\hat{x},2z-1) & \text{in } G^+. \end{cases} \tag{3.38}$$

This result and (3.19)–(3.20) give the following result.

THEOREM 3.4. *Let $(m^\varepsilon, \phi^\varepsilon)$ be a global solution of problem* (1.1) *with* (3.9)–(3.10) *satisfying the interlayer exchange boundary condition* (3.11). *Let (m, ϕ) be the weak-$\star$ limit of a subsequence of $(m^\varepsilon, \phi^\varepsilon)$. Then (m, ϕ) satisfies in $\mathbb{R}^+ \times G$ the equations*

$$\begin{cases} \partial_t m - \alpha m \times \partial_t m = -(1+\alpha^2) m \times \mathcal{H}_0(m) \ \text{ in } \mathbb{R}^+ \times (G^- \cup G^+), \\ m(0,x) = m_0(x), \qquad \dfrac{\partial m}{\partial n} = 0 \text{ on } \partial\Omega \setminus \{z = \pm 1/2\}, \\ m(\pm 1/2) \times \Big(\mp a \partial_z m(\pm 1/2) - 2Jm(\mp 1/2) \Big) = 0 \text{ in } B, \\ \mathcal{H}_0(m) = \widehat{\operatorname{div}}(a\ \widehat{\operatorname{grad}}\, m) + \dfrac{1}{4}\partial_z(a\ \partial_z m) + \psi(m) + \widehat{\operatorname{grad}}\, \phi + \dfrac{1}{2}\partial_z \phi\ \mathbf{u}_3 \ \text{ in } G, \end{cases}$$

where ϕ is the solution of the magnetostatic equations (3.22)–(3.23).

Remark. Notice that the potential ϕ solves the magnetostatic equations in the disjoint domains S^- and $S^{+,\infty}$. These domains are coupled together by a Ventcel-type boundary condition given in Theorem 3.3 linking the interfaces $z = -1/2$ and $z = 1/2$. The magnetization m satisfies the Hoffmann IEC law with the coefficient $2J$ instead of J due to the change of variable used.

3.2. Asymptotic behavior with large nonmagnetic spacers. The aim of this subsection is to discuss the behavior of the problem when $\varepsilon \to 1$. The estimates satisfied by the solutions $(m^\varepsilon, \phi^\varepsilon)$ are given by the energy estimate (3.12)–(3.13). Let us discuss the admissibility criterion for the initial data m_0^ε. The compatibility condition for ϕ_0^ε is given by (3.14). The condition that m_0^ε is independent of the variable $z \in G$ ensures that $\mathcal{E}_{exc}^\varepsilon(0) \le C$ and $\mathcal{E}_{int}^\varepsilon(0) = -\frac{J}{2(1-\varepsilon)}|B|$. Since we have $\mathcal{E}_{vol}^\varepsilon(0) \le C$, it remains to show that under this condition on m_0^ε we have $\mathcal{E}_{dm}^\varepsilon(0) \le C$ uniformly with respect to ε. This follows from estimates (3.15)–(3.16). As in subsection 3.1, $C > 0$ denotes various constants which are independent of ε.

LEMMA 3.5. *Let $m_0^\varepsilon \in \mathbb{H}^1(B)$ such that m_0^ε is uniformly bounded in $\mathbb{H}^1(B)$ and satisfies $|m_0^\varepsilon(\hat{x})|^2 = 1$ a.e. in B. Then we have*

$$\mathcal{E}_{exc}^\varepsilon(0) + \mathcal{E}_{vol}^\varepsilon(0) + \mathcal{E}_{dm}^\varepsilon(0) \le C, \qquad |\mathcal{E}_{int}^\varepsilon(0)| \le C\frac{J}{1-\varepsilon}. \tag{3.39}$$

Moreover, the solutions $(m^\varepsilon, \phi^\varepsilon)$ associated with m_0^ε satisfy, for all $t \ge 0$, the estimates

$$0 \le \mathcal{E}_{exc}^\varepsilon(t) + \mathcal{E}_{vol}^\varepsilon(t) + \mathcal{E}_{dm}^\varepsilon(t) \le C\left(1 + \frac{J}{1-\varepsilon}\right), \qquad |\mathcal{E}_{int}^\varepsilon(t)| \le C\left(1 + \frac{J}{1-\varepsilon}\right). \tag{3.40}$$

In what follows this subsection, we assume that the interlayer exchange coefficient J satisfies the hypothesis

$$J = j(1-\varepsilon),\ j > 0, \tag{3.41}$$

where j is independent of ε.

The energy estimate (3.12) and (3.13) imply the following uniform bounds.

LEMMA 3.6. *Under the hypotheses of Lemma* 3.5 *and* (3.40), *the solutions*

$(m^\varepsilon, \phi^\varepsilon)$ satisfy the following estimates:

$$\begin{cases} |\widehat{\operatorname{grad}} m^\varepsilon|_{L^\infty(\mathbb{R}^+,\mathbb{L}^2(G))} + |\partial_t m^\varepsilon|_{L^2(\mathbb{R}^+,\mathbb{L}^2(G))} \le C, \\ |\widehat{\operatorname{grad}} \phi^\varepsilon|_{L^\infty(\mathbb{R}^+,\mathbb{L}^2(S^{-,\infty}))} + |\partial_z \phi^\varepsilon|_{L^\infty(\mathbb{R}^+,L^2(S^0\cup S^\infty))} \le C, \\ |\partial_z m^\varepsilon|_{L^\infty(\mathbb{R}^+,\mathbb{L}^2(G))} + |\partial_z \phi^\varepsilon|_{L^\infty(\mathbb{R}^+,L^2(S^-\cup S^+))} \le C(1-\varepsilon). \end{cases} \tag{3.42}$$

There exists a subsequence still denoted $(m^\varepsilon, \phi^\varepsilon)$ such that the following convergences hold:

$$\begin{cases} m^\varepsilon \rightharpoonup m \text{ weakly}-\star \text{ in } L^\infty(\mathbb{R}^+, \mathbb{H}^1(G)) \cap \mathbb{L}^\infty(\mathbb{R}^+ \times G), \\ \partial_t m^\varepsilon \rightharpoonup \partial_t m \text{ weakly in } L^2(\mathbb{R}^+, \mathbb{L}^2(G)), \\ \partial_z m^\varepsilon \to 0 \text{ strongly in } L^\infty(\mathbb{R}^+, \mathbb{L}^2(G)), \\ m^\varepsilon \to m \text{ strongly in } L^2_{\mathrm{loc}}(\mathbb{R}^+, \mathbb{L}^2(G)) \end{cases} \tag{3.43}$$

and

$$\begin{cases} \widehat{\operatorname{grad}} \phi^\varepsilon \rightharpoonup \widehat{\operatorname{grad}} \phi \text{ weakly}-\star \text{ in } L^\infty(\mathbb{R}^+, \mathbb{L}^2(S^{-,\infty})), \\ \partial_z \phi^\varepsilon \to 0 \text{ strongly in } L^\infty(\mathbb{R}^+, L^2(S^- \cup S^+)), \\ \partial_z \phi^\varepsilon \rightharpoonup \partial_z \phi \text{ weakly}-\star \text{ in } L^\infty(\mathbb{R}^+, L^2(S^0 \cup S^\infty)). \end{cases} \tag{3.44}$$

Let us first consider the behavior of the potential ϕ^ε when $\varepsilon \to 1$. We have the following characterization of the potential limit ϕ.

We denote by ϕ^-, ϕ^0, ϕ^+, and ϕ^∞ the restrictions of ϕ to the domains $\mathbb{R}^+ \times S^-$, $\mathbb{R}^+ \times S^0$, $\mathbb{R}^+ \times S^+$, and $\mathbb{R}^+ \times S^\infty$, respectively, and $m^\pm$ the restriction of m to $\mathbb{R}^+ \times G^\pm$. Notice that $\phi^\pm$ and $m^\pm$ are independent of the variable z.

THEOREM 3.7. *We assume that the hypotheses of Lemma* 3.6 *are fulfilled. Let ϕ^ε be the solution of* (3.9) *and ϕ its weak-$\star$ limit. Then we have*

$$\begin{cases} \phi^+(\hat{x}) = \phi^\infty(\hat{x}, 1) = \phi^0(\hat{x}, 1/2) \text{ a.e.}, \\ \phi^-(\hat{x}) = \phi^0(\hat{x}, -1/2) \text{ a.e.} \end{cases} \tag{3.45}$$

Moreover, the couple (ϕ^∞, ϕ^0) satisfies, for all $t \ge 0$, the problem

$$\begin{cases} \widehat{\Delta} \phi^\infty + \partial_z^2 \phi^\infty = 0 \text{ in } \mathbb{R}^+ \times S^\infty, \\ \widehat{\Delta} \phi^0 + \dfrac{1}{4} \partial_z^2 \phi^0 = 0 \text{ in } \mathbb{R}^+ \times S^0, \end{cases} \tag{3.46}$$

with the coupling boundary conditions

$$\begin{cases} \phi^\infty(1) = \phi^0(1/2), \\ \partial_z \phi^\infty(1) - \dfrac{1}{4} \partial_z \phi^0(1/2) = -\dfrac{1}{2} \widehat{\operatorname{div}}(\widehat{\operatorname{grad}} \phi^\infty(1) + \chi(B) \widehat{m}^+), \\ \partial_z \phi^0(-1/2) = -2 \widehat{\operatorname{div}}(\widehat{\operatorname{grad}} \phi^0(-1/2) + \chi(B) \widehat{m}^-). \end{cases} \tag{3.47}$$

Furthermore, setting $\Theta_\varepsilon^\pm = \frac{1}{2(1-\varepsilon)}\partial_z\phi^\varepsilon_{|S^\pm}$ *the restriction to* $\mathbb{R}^+ \times S^\pm$*, we have the convergence*

$$\Theta_\varepsilon^\pm \rightharpoonup -\chi(G^\pm)m^\pm \cdot \mathbf{u}_3 \text{ weakly}-\star \text{ in } L^\infty(\mathbb{R}^+, L^2(S^\pm)). \tag{3.48}$$

Proof. Estimates (3.43) and (3.44) show that the restriction of ϕ to $\mathbb{R}^+ \times S^\pm$ and the magnetization m are independent of the variable z. We shall pass to the limit in each slab where the potential ϕ^ε is defined.

In the domains $\mathbb{R}^+ \times S^\pm$, the potential ϕ^ε satisfies the equation

$$\partial_z\left(\frac{1}{2(1-\varepsilon)}\partial_z\phi^\varepsilon + \chi(G^\pm)m^\varepsilon \cdot \mathbf{u}_3\right) = -2(1-\varepsilon)\ \widehat{\operatorname{div}}\Big(\widehat{\operatorname{grad}}\,\phi^\varepsilon + \chi(G^\pm)\widehat{m}^\varepsilon\Big). \tag{3.49}$$

Let $\Theta_\varepsilon^\pm = \frac{1}{2(1-\varepsilon)}\partial_z\phi^\varepsilon_{|S^\pm}$ be defined as previously. By Lemma 3.6, we have $\Theta_\varepsilon^\pm \rightharpoonup \Theta^\pm$ weakly-$\star$ in $L^\infty(\mathbb{R}^+, L^2(S^\pm))$. Passing to the limit in (3.49) by using the estimates given in Lemma 3.6, one deduces that $\Theta^\pm$ satisfies the equation

$$\partial_z(\Theta^\pm + \chi(G^\pm)m(\hat{x}) \cdot \mathbf{u}_3) = 0 \text{ in } \mathbb{R}^+ \times S^\pm. \tag{3.50}$$

Consequently, we have $\Theta^\pm + \chi(G^\pm)m(\hat{x}) \cdot \mathbf{u}_3 = C^\pm(t, \hat{x})$, where the unknown functions $C^\pm$ are independent of the variable z. Next, since the transmission boundary conditions at the interfaces $z = \pm 1/2^\pm$ take the form

$$\Theta_\varepsilon^\pm(\pm 1/2^\pm) + \chi(B)m^\varepsilon(\pm 1/2^\pm) \cdot \mathbf{u}_3 = \frac{1-\varepsilon}{2\varepsilon^2}\partial_z\phi^\varepsilon(\pm 1/2^\mp), \tag{3.51}$$

we deduce, by using the convergences given in (3.44), that $\partial_z\phi^\varepsilon(\pm 1/2^\mp) \rightharpoonup \partial_z\phi(\pm 1/2^\mp)$ and $\Theta^\varepsilon(\pm 1/2^\pm) \rightharpoonup \Theta(\pm 1/2^\pm)$ in the sense of distributions. Hence, we get

$$\Theta^\pm(\pm 1/2) + \chi(B)m(\pm 1/2) \cdot \mathbf{u}_3 = 0, \tag{3.52}$$

and finally (3.50) gives the result

$$\Theta^\pm(t, \hat{x}, z) = -\chi(G^\pm)\ m(t, \hat{x}) \cdot \mathbf{u}_3 \text{ in } \mathbb{R}^+ \times S^\pm. \tag{3.53}$$

Now we are dealing with the equation satisfied by $\phi^\pm$ in $\mathbb{R}^+ \times S^\pm$. (Recall that $\phi^\pm$ is independent of z.) Let $g \in \mathcal{D}(\mathbb{R}^+ \times \mathbb{R}^2)$ be a test function which is independent of the variable z. Multiplying (3.49) by g and integrating by parts, we get

$$\begin{cases} \displaystyle\int_{\mathbb{R}^+\times S^+} (\widehat{\operatorname{grad}}\,\phi^\varepsilon + \chi(G^+)\widehat{m}^\varepsilon)(\widehat{\operatorname{grad}}\,g)\ d\hat{x}dzdt \\ \displaystyle= \int_{\mathbb{R}^+\times\mathbb{R}^2}\left(\frac{1}{(2(1-\varepsilon))^2}\partial_z\phi^\varepsilon + \frac{1}{2(1-\varepsilon)}\chi(B)m^\varepsilon \cdot \mathbf{u}_3\right)(1^-)g\ d\hat{x}dt \\ \displaystyle- \int_{\mathbb{R}^+\times\mathbb{R}^2}\left(\frac{1}{(2(1-\varepsilon))^2}\partial_z\phi^\varepsilon + \frac{1}{2(1-\varepsilon)}\chi(B)m^\varepsilon \cdot \mathbf{u}_3\right)(1/2^+)g\ d\hat{x}dt. \end{cases} \tag{3.54}$$

Using the transmission boundary conditions at the interfaces $z = 1$ and $z = 1/2$, (3.54) then becomes

$$\begin{aligned} &\int_{\mathbb{R}^+\times S^+} (\widehat{\operatorname{grad}}\,\phi^\varepsilon + \chi(G^+)\widehat{m}^\varepsilon)(\widehat{\operatorname{grad}}\,g)\ d\hat{x}dzdt \\ &= \int_{\mathbb{R}^+\times\mathbb{R}^2} \partial_z\phi^\varepsilon(1^+)g\ d\hat{x}dt - \int_{\mathbb{R}^+\times\mathbb{R}^2} \frac{1}{(2\varepsilon)^2}\partial_z\phi^\varepsilon(1/2^-)g\ d\hat{x}dt. \end{aligned} \tag{3.55}$$

Passing to the limit by using Lemma 3.6, we get the equation

$$\begin{aligned}&\frac{1}{2}\int_{\mathbb{R}^+\times\mathbb{R}^2}(\widehat{\operatorname{grad}}\,\phi^+ + \chi(B)\widehat{m}^+)(\widehat{\operatorname{grad}}\,g)\ d\hat{x}dt\\ &=\int_{\mathbb{R}^+\times\mathbb{R}^2}\partial_z\phi^\infty(1)g\ d\hat{x}dt-\int_{\mathbb{R}^+\times\mathbb{R}^2}\frac{1}{4}\partial_z\phi^0(1/2)g\ d\hat{x}dt.\end{aligned}\tag{3.56}$$

Using the continuity of the potential across the interfaces $z=1$, $z=\pm 1/2$, we deduce that ϕ satisfies the continuity property

$$\phi^+(\hat{x},1)=\phi^\infty(\hat{x},1),\ \phi^0(\hat{x},1/2)=\phi^+(\hat{x},1/2),\ \phi^0(\hat{x},1/2)=\phi^-(\hat{x},1/2)\ \text{a.e.}\tag{3.57}$$

Integrating by parts in (3.55) and recalling that m is independent of the variable z in $\mathbb{R}^+\times G^+$, we get the transmission boundary condition coupling the interfaces $z=1$ and $z=1/2$:

$$\partial_z\phi^\infty(1)-\frac{1}{4}\partial_z\phi^0(1/2)=-\frac{1}{2}\widehat{\operatorname{div}}(\widehat{\operatorname{grad}}\,\phi^\infty(1)+\chi(B)\widehat{m}^+).\tag{3.58}$$

Let us consider the convergence of the restriction of ϕ^ε to the domain $\mathbb{R}^+\times S^-$. We proceed as we did previously in the domain $\mathbb{R}^+\times S^+$ and recall that ϕ^ε satisfies at $z=-1$ the Neumann homogeneous boundary condition. We get

$$\begin{aligned}&\int_{\mathbb{R}^+\times S^-}(\widehat{\operatorname{grad}}\,\phi^\varepsilon+\chi(G^-)\widehat{m}^\varepsilon)(\widehat{\operatorname{grad}}\,g)\ d\hat{x}dzdt\\ &=\int_{\mathbb{R}^+\times\mathbb{R}^2}\left(\frac{1}{(2(1-\varepsilon))^2}\partial_z\phi^\varepsilon+\frac{1}{2(1-\varepsilon)}\chi(B)m^\varepsilon\cdot\mathbf{u}_3\right)(-1/2^-)g\ d\hat{x}dt,\end{aligned}\tag{3.59}$$

which gives, by using the transmission condition at the interface $z=-1/2$,

$$\begin{aligned}&\int_{\mathbb{R}^+\times S^-}(\widehat{\operatorname{grad}}\,\phi^\varepsilon+\chi(G^-)\widehat{m}^\varepsilon)(\widehat{\operatorname{grad}}\,g)\ d\hat{x}dzdt\\ &=\int_{\mathbb{R}^+\times\mathbb{R}^2}\frac{1}{(2\varepsilon)^2}\partial_z\phi^\varepsilon(-1/2^+)g\ d\hat{x}dt.\end{aligned}\tag{3.60}$$

We pass to the limit in the equation. To do that we use the fact that ϕ and m are independent of the variable z in S^-, and we use the continuity of the potential ϕ at the interface $z=-1/2$. We get the equation

$$2\int_{\mathbb{R}^+\times\mathbb{R}^2}(\widehat{\operatorname{grad}}\,\phi^0+\chi(B)\widehat{m}^-)(\widehat{\operatorname{grad}}\,g)\ d\hat{x}dt=\int_{\mathbb{R}^+\times\mathbb{R}^2}\partial_z\phi^0(-1/2^+)g\ d\hat{x}dt\tag{3.61}$$

or, equivalently, the boundary condition at the interface $z=-1/2$, by using the fact that $\phi^-(\hat{x})=\phi^0(\hat{x},-1/2)$:

$$\partial_z\phi^0(-1/2^+)=-2\widehat{\operatorname{div}}(\widehat{\operatorname{grad}}\,\phi^0+\chi(B)\widehat{m}^-).\tag{3.62}$$

Next, we pass to the limit in the equations satisfied by the potential ϕ^ε in $\mathbb{R}^+\times S^0$ and $\mathbb{R}^+\times S^\infty$, which satisfies the equations

$$\begin{cases}\widehat{\operatorname{div}}(\widehat{\operatorname{grad}}\,\phi^\varepsilon)+\dfrac{1}{4\varepsilon^2}\partial_z^2\phi^\varepsilon=0\ \text{ in }\ \mathbb{R}^+\times S^0,\\ \widehat{\operatorname{div}}(\widehat{\operatorname{grad}}\,\phi^\varepsilon)+\partial_z^2\phi^\varepsilon=0\ \text{ in }\ \mathbb{R}^+\times S^\infty.\end{cases}\tag{3.63}$$

It follows that the potential ϕ satisfies the equations (recall the interface continuity equalities (3.57))

$$\begin{cases} \widehat{\Delta}\phi^0 + \dfrac{1}{4}\partial_z^2\phi^0 = 0 \ \text{ in } \mathbb{R}^+ \times S^0, \\ \widehat{\Delta}\phi^\infty + \partial_z^2\phi^\infty = 0 \ \text{ in } \mathbb{R}^+ \times S^\infty, \\ \phi^\infty(1) = \phi^0(1/2). \end{cases} \tag{3.64}$$

Hence, the theorem is proved. □

The local magnetic excitation $\mathcal{H}^\varepsilon(m^\varepsilon)$ involves the magnetic field

$$\widehat{\operatorname{grad}}\,\phi^\varepsilon + \frac{1}{2(1-\varepsilon)}\partial_z\phi^\varepsilon \mathbf{u}_3$$

in $S^\pm$. When $\varepsilon \to 1$, we get the following convergence result.

LEMMA 3.8. *The magnetic field satisfies the convergence*

$$\left(\widehat{\operatorname{grad}}\,\phi^\varepsilon + \frac{1}{2(1-\varepsilon)}\partial_z\phi^\varepsilon \mathbf{u}_3\right)_{|\mathbb{R}^+\times S^\pm} \rightharpoonup (\widehat{\operatorname{grad}}\,\phi^\pm - \chi(G^\pm) m\cdot \mathbf{u}_3\mathbf{u}_3)_{|\mathbb{R}^+\times S^\pm} \tag{3.65}$$

weakly-$\star$ in $L^\infty(\mathbb{R}^+, \mathbb{L}^2(S^\pm))$.

Now, we may pass to the limit in the Landau–Lifshitz equations. Arguing that, in the limit, $m^\pm$ is independent of z, we use test functions of the type $F^\varepsilon(t,\hat{x},z) = F(t,\hat{x}) + 2(1-\varepsilon)F_0(t,\hat{x})h(2(1-\varepsilon)z)$, where $h \in \mathcal{D}([-1,1])$ and $F, F_0 \in \mathcal{D}(B)$. The weak formulation of the problem reads in each domain $Q^\pm = \mathbb{R}^+ \times G^\pm$ as follows:

$$\begin{cases} \dfrac{1}{1+\alpha^2}\displaystyle\int_{Q^\pm} (\partial_t m^\varepsilon - \alpha m^\varepsilon \times \partial_t m^\varepsilon)\ (F + h^\varepsilon(z)F_0)dxdt \\ = \displaystyle\int_{Q^\pm} m^\varepsilon \times (a^\varepsilon \widehat{\operatorname{grad}}\, m^\varepsilon)(\widehat{\operatorname{grad}}\, F + h^\varepsilon(z)\widehat{\operatorname{grad}}\, F_0)\ dxdt \\ + \displaystyle\int_{Q^\pm} m^\varepsilon \times (a^\varepsilon \partial_z m^\varepsilon)F_0 h'(2(1-\varepsilon)z)\ dxdt \\ - \displaystyle\int_{Q^\pm} m^\varepsilon \times \left(\psi(m^\varepsilon) + \widehat{\operatorname{grad}}\,\phi^\varepsilon + \frac{1}{2(1-\varepsilon)}\partial_z\phi^\varepsilon \mathbf{u}_3\right)(F + h^\varepsilon(z)F_0)\ dxdt \\ - \dfrac{j}{2}\displaystyle\int_{\mathbb{R}^+\times B} m^\varepsilon(\pm 1/2) \times m^\varepsilon(\mp 1/2)\Big(F + h^\varepsilon(\pm(1-\varepsilon))F_0\Big)\ d\hat{x}dt, \end{cases} \tag{3.66}$$

where we have set $h^\varepsilon(z) = 2(1-\varepsilon)h(2(1-\varepsilon)z)$ and used the IEC boundary condition at the interfaces $z = \pm 1/2$:

$$m^\varepsilon(\pm 1/2) \times \left(\mp\frac{a^\varepsilon}{2(1-\varepsilon)}\partial_z m^\varepsilon(\pm 1/2) - j(1-\varepsilon)m^\varepsilon(\mp 1/2)\right) = 0. \tag{3.67}$$

We pass to the limit in each term of the weak formulation (3.66). By using

Lemmas 3.6 and 3.8, we get

$$
\begin{cases}
\dfrac{1}{1+\alpha^2}\displaystyle\int_{\mathbb{R}^+\times B}\left(\partial_t m^\pm - \alpha m^\pm\times\partial_t m^\pm\right)\ F d\hat{x}dt \\
=\displaystyle\int_{\mathbb{R}^+\times B} m^\pm\times(a^\pm\widehat{\operatorname{grad}}\, m^\pm)\widehat{\operatorname{grad}}\, F\ d\hat{x}dt \\
-\displaystyle\int_{\mathbb{R}^+\times B} m^\pm\times\left(\psi(m^\pm)+\widehat{\operatorname{grad}}\,\phi^\pm-\chi(B)(m^\pm\cdot\mathbf{u}_3)\mathbf{u}_3\right)\ F\ d\hat{x}dt \\
-j\displaystyle\int_{\mathbb{R}^+\times B} m^\pm\times m^\mp\ F\ d\hat{x}dt,
\end{cases}
\tag{3.68}
$$

where we used the strong convergence

$$
a^\varepsilon_{|G^\pm}\to a^\pm(\hat{x})=A(\hat{x},\pm 1)\quad\text{a.e.}
\tag{3.69}
$$

Gathering all results of this subsection, we get the following convergence theorem.

THEOREM 3.9. *Let $m^\pm$ be the weak-$\star$ limit in $L^\infty(\mathbb{R}^+,\mathbb{H}^1(G^\pm))$ of a subsequence of $m^\varepsilon_{|G^\pm}$. Then the couple (m^+,m^-) is independent of the variable z and satisfies $|m^\pm(t,\hat{x})|^2=1$ a.e. Moreover, (m^+,m^-) satisfies in $\mathbb{R}^+\times B$ the Landau–Lifshitz–Gilbert equations*

$$
\begin{cases}
\partial_t m^\pm-\alpha m^\pm\times\partial_t m^\pm=-(1+\alpha^2)m^\pm\times\mathcal{H}^\pm(m^\pm), \\
\\
m^\pm(0,\hat{x})=m_0(\hat{x}),\qquad \dfrac{\partial m^\pm}{\partial n}=0\ \text{on}\ \partial B,
\end{cases}
\tag{3.70}
$$

where the total magnetic field $\mathcal{H}^\pm(m^\pm)$ is given by

$$
\mathcal{H}^\pm(m^\pm)=\widehat{\operatorname{div}}(a^\pm\ \widehat{\operatorname{grad}}\, m^\pm)+\psi(m^\pm)+\widehat{\operatorname{grad}}\,\phi^\pm-\chi(B)m^\pm\cdot\mathbf{u}_3\ \mathbf{u}_3+jm^\mp.
\tag{3.71}
$$

Furthermore, the potential $\phi^\pm(t,\hat{x})$ is given by Theorem 3.7.

4. Concluding remarks. The results obtained can be applied without difficulty to the case of the full Maxwell system. In this study we have considered a plane interface between layers. It would be very interesting to extend this analysis to a more general case of interfacial roughness, especially in films with compensated interfaces, which play an important role in giant magnetoresistance (GMR) [18], [23]. Note that the magnetic multilayers with GMR have attracted a lot of interest due to their high density information storage and retrieval capacities. Also, as already said in the introduction, we have limited ourselves to bilinear coupling. Thus one may consider a more general energy coupling density which takes into account the biquadratic effect recently discovered in layered magnetic systems [14]. It is mostly expressed [18], [23] (see also the introduction) as $J_1(1-m^+\cdot m^-)+J_2(1-(m^+\cdot m^-)^2)$, where m^+ and m^- are the magnetization vectors at the inner surfaces of the first and the second magnetic slabs, respectively. To take into account that energy, we use the following boundary condition at the interfaces $+$ and $-$ facing each other:

$$
m^\pm\times\left(A\frac{\partial m^\pm}{\partial n^\pm}-J_1m^\mp-J_2(m^+\cdot m^-)m^\mp\right)=0.
$$

Considering the problem with this energy and the roughness of interfaces, the asymptotic behavior of the interlayer coupling will be difficult to approach and will require more detailed studies.

Acknowledgments. We greatly appreciate the remarks and comments of the anonymous referee. We thank her/him for a careful reading of the first draft of this paper and for many helpful suggestions for improving the overall presentation.

REFERENCES

[1] A. Aharoni, *Introduction to the Theory of Ferromagnetism*, Oxford University Press, London, 1996.
[2] R. Alicandro and C. Leone, *3D-2D asymptotic analysis for micromagnetic thin films*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 489–498.
[3] F. Alouges and A. Soyeur, *On global weak solutions for Landau–Lifshitz equations: Existence and non uniqueness*, Nonlinear Anal., 18 (1992), pp. 1071–1084.
[4] H. Ammari, L. Halpern, and K. Hamdache, *Asymptotic behaviour of thin ferromagnetic films*, Asymptot. Anal., 24 (2000), pp. 277–294.
[5] J. Barnaś and B. Bulka, *Oscillations in magnetoresistance and interlayer coupling in magnetic sandwich structures*, Acta Phys. Polon. A, 91 (1997), pp. 253–256.
[6] J. Barnaś, *Dependence of bilinear and biquadratic interlayer coupling on thickness of magnetic films*, Acta Phys. Polon. A, 91 (1997), pp. 257–260.
[7] J. Barnaś, *Spin waves in a bilayer with biquadratic interlayer coupling*, Phys. Status Solidi (b), 203 (1997), pp. 221–228.
[8] W. F. Brown, *Micromagnetics*, Interscience Publishers, New York, 1963.
[9] P. Bruno, *Theory of interlayer magnetic coupling*, Phys. Rev. B, 52 (1995), pp. 411–439.
[10] P. Bruno, *Theory of interlayer exchange coupling*, in "30. Ferienkurs des Instituts für Festkörperforschung: Magnetische Schichtsysteme," P. H. Dederichs and P. Grünberg, eds., Forschungszentrum Jülich, preprint, 1999.
[11] R. E. Camley, *Magnetization dynamics in thin films and multilayers*, J. Magn. Magn. Mater., 200 (1999), pp. 583–597.
[12] R. E. Camley and R. L. Stamps, *Magnetic multilayers: Spin configurations, excitations and giant magnetoresistance*, J. Phys.: Condensed Matter, 5 (1993), pp. 3727–3786.
[13] G. Carbou and P. Fabrie, *Time average in micromagnetism*, J. Differential Equations, 147 (1998), pp. 383–409.
[14] S. O. Demokritov, *Biquadratic interlayer coupling in layered magnetic systems*, J. Phys. D: Appl. Phys., 31 (1998), pp. 925–941.
[15] G. Gioia and R. D. James, *Micromagnetics of very thin films*, Proc. Roy. Soc. London. A, 453 (1997), pp. 213–223.
[16] P. Grünberg and D. T. Pierce, *Interlayer exchange coupling*, in Encyclopedia of Materials: Science and Technology, Elsevier Science, New York, 2001, pp. 5883–5888.
[17] K. Hamdache and M. Tilioua, *On the zero thickness limit of thin ferromagnetic films with surface anisotropy energy*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1469–1490.
[18] U. Hartmann, ed., *Magnetic Multilayers and Giant Magnetoresistance. Fundamentals and Industrial Applications*, Springer-Verlag, Berlin, 2000.
[19] B. Hillebrands, *Spin-wave calculations for multilayered structures*, Phys. Rev. B, 41 (1990), pp. 530–540.
[20] F. Hoffmann, *Dynamic pinning induced by nickel layers on permalloy films*, Phys. Status Solidi, 41 (1970), pp. 807–813.
[21] F. Hoffmann, *Ondes de Spin Stationnaires dans les Couches Couplées*, Thèse de Doctorat d'Etat, Sciences physiques, Université Paris XI-Orsay, Orsay, France, 1971.
[22] F. Hoffmann, A. Stankoff, and H. Pascard, *Evidence for an exchange coupling at the interface between two ferromagnetic films*, J. Appl. Phys., 41 (1970), pp. 1022–1023.
[23] A. Hubert and R. Schäfer, *Magnetic Domains. The Analysis of Magnetic Microstructures*, Springer-Verlag, New York, Berlin, 1998.
[24] M. Labrune and L. Belliard, *Stripe domains in multilayers: Micromagnetic simulations*, Phys. Status Solidi (a), 174 (1999), pp. 483–489.
[25] M. Labrune and J. Miltat, *Wall structures in ferro/antiferromagnetic exchange-coupled bilayers: A numerical micromagnetic approach*, J. Magn. Magn. Mater., 151 (1995), pp. 231–245.
[26] J.-L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod & Gauthier-Villars, Paris, 1969.
[27] Y. Roussigné, F. Ganot, C. Dugautier, P. Moch, and D. Renard, *Brillouin scattering in Co/Cu/Co and Co/Au/Co trilayers: Anisotropy fields and interlayer magnetic exchange*, Phys. Rev. B, 52 (1995), pp. 350–360.

[28] M. D. Stiles, *Exchange coupling in magnetic heterostructures*, Phys. Rev. B, 48 (1993), pp. 7238–7258.
[29] M. D. Stiles, *Interlayer exchange coupling*, J. Magn. Magn. Mater., 200 (1999), pp. 322–337.
[30] L. Tartar, *Topics in Nonlinear Analysis*, technical report, Publications Mathématiques d'Orsay 78.13, Université de Paris-Sud, Paris, 1978.
[31] A. Visintin, *On the Landau–Lifshitz equation for ferromagnetism*, Japan J. Appl. Math., 2 (1985), pp. 69–84.
[32] J. J. de Vries, *Interlayer Exchange Coupling in Magnetic Multilayers*, Ph.D. thesis, Department of Physics, TU Eindhoven, Eindhoven, The Netherlands, 1996.
[33] A. Yoshihara, J. T. Wang, K. Takanashi, K. Himi, Y. Kawazoe, H. Fujimori, and P. Grünberg, *Interlayer exchange coupling in fine-layered Fe/Au superlattices*, Phys. Rev. B, 63 (2001), 100405(R).

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1099–1124

CURVATURE-INDUCED DISPERSION IN ELECTRO-OSMOTIC SERPENTINE FLOWS*

EHUD YARIV†, HOWARD BRENNER‡, AND SANGTAE KIM§

Abstract. Flow and transport phenomena occurring within serpentine microchannels are analyzed for both two- and three-dimensional curvilinear configurations. The microfluidic conduit is modeled as a spatially periodic "thin" channel, enabling asymptotic expansions of the pertinent transport fields in terms of a small parameter ϵ, representing the ratio of channel (half-)width to curvilinear channel length per serpentine period. The electric potential distribution, as well as the attendant electro-osmotic flow field, is calculated for the limiting case where the Debye layer thickness is small relative to the channel width. Generalized Taylor–Aris dispersion theory is employed to calculate the serpentine-scale velocity and dispersivity of a charged point-size colloidal Brownian particle ("molecule") entrained in the solvent Stokes flow engendered by the electrokinetic forces. These respective macrotransport coefficients are expressed, inter alia, in terms of quadratures of the local curvature within a unit cell of the serpentine device.

Key words. electro-osmosis, asymptotic expansions, Taylor dispersion

AMS subject classifications. 76W05, 76D07, 76D08, 41A60

DOI. 10.1137/S003613990342284X

1. Introduction.

Conventional chromatographic separation schemes exploit the differences in mean solute velocity with which the various species move throughout the chromatograph (into which they were introduced simultaneously). As the average separation distance between species increases linearly with channel length, it is desirable to maximize the available length between the entrance and exit of a solute molecule within the device. In microfluidic devices, constrained by chip-size limitations, improved chromatographic efficiency can therefore be achieved by folding the otherwise straight separation channel into the shape of a serpentine channel [12, 13, 5]. In attempting to optimize chip layout design it is obviously useful to qualitatively understand the transport factors which distinguish curvilinear channels from their rectilinear counterparts, whose chromatographic properties are well understood [3]. This is the goal of the present paper.

The major drawback of all chromatographic separation devices lies in the increased band-broadening experienced by the solute sample over and above that due to purely molecular diffusion alone. This increased solute spread results from the Taylor-dispersion mechanism, whereby Brownian solute particles sample different streamlines (in particular their concomitant axial velocities) within the channel cross section. The cross-sectional velocity variance is generally greater in curved channels than in comparably straight channels [2, 8]. Thus, while Taylor effects also exist in straight channels (especially in pressure-driven flows), channel curvature could significantly modify these convective-dispersion effects. Accordingly, one of the main foci of the present

*Received by the editors February 21, 2003; accepted for publication (in revised form) July 17, 2003; published electronically April 21, 2004. This research was supported by the Lilly Research Laboratories.

http://www.siam.org/journals/siap/64-4/42284.html

†Faculty of Mechanical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (yarive@tx.technion.ac.il).

‡Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02478 (hbrenner@mit.edu).

§Department of Mechanical Engineering and Department of Chemical Engineering, Purdue University, 585 Purdue Mall, West Lafayette, IN 47907 (kim55@purdue.edu).

paper lies in examining the relevant enhancement of solute dispersion associated with the curvature of the serpentine channel.

A systematic analysis of solute dispersion accompanying pressure-driven flows in serpentine channels was recently carried out by [22], using macrotransport theory [3] for spatially periodic systems. All other things being equal, the mean axial solute velocity $\overline{U}^*$ and dispersivity $\overline{D}^*$ in the direction of net flow were found to be proportional to the corresponding quantities for the straight unfolded channel, the constants of proportionality for these two macrotransport parameters being L/l (< 1) and $(L/l)^2$, respectively. The factor L/l constitutes the inverse of the "tortuosity," namely the ratio of serpentine channel length, l, to its projected (rectilinear) channel length in the direction of net flow, L. These results are easily rationalized in terms of the combined effects arising from the extended total serpentine channel length and the decreased effective cross-sectional area of the curved channel. Since only leading-order terms were evaluated by [22], their results are based upon a Poiseuillian velocity profile, with no dependence upon curvature-induced deviations from it.

Practical microfluidic separation devices employ ionic solutions as the solvent phase, acting as a buffer for the chemically sensitive solute molecules. As the boundaries of the channel are usually charged (as a consequence of the presence of electrolytes), it is common practice to use external electric fields to drive the bulk flow. The resulting electro-osmotic flow represents a balance between the animating electrical body forces, concentrated within the Debye double-layer, and the retarding viscous stresses. Use of electro-osmotic flows eliminates the need for inefficient micropumps in favor of electrodes. Moreover, the length-scale on which the velocity field varies cross-sectionally across the double-layer is extremely small compared with typical channel widths. Thus, the velocity profile is substantially uniform over the major portion of the channel cross section. For such "plug flows" the only source of solute dispersion is axial molecular diffusion. We seek, inter alia, to calculate the additional Taylor–Aris dispersion arising in curvilinear channels (over that occurring in straight channels), wherein the velocity field deviates from a plug flow owing to channel curvature.

Furthermore, if the solute particles are charged, the electric field operates as an external force field (albeit indirectly, owing to the presence of a Debye double-layer around the particles). As a result of the lateral (and perhaps longitudinal) variations of the electric field in a curved channel, the electrophoretic particle velocity (relative to the already nonuniform velocity field) depends upon the local position of the particle within the channel. (This electrophoretic velocity is related to the electric field through the particle's electrophoretic mobility.) This constitutes yet another source of solute Taylor dispersion, which, similarly to that arising from the nonuniform solvent velocity field, is absent in electro-osmotic flows in straight channels. As nonuniformities of both velocity and electric field occur in electro-osmotic flows in curved channels, each may individually contribute to the resulting Taylor dispersion. It is the purpose of this paper to provide the required analysis for such phenomena, at least for solute particles that can be regarded as being effectively "point-size" compared with the channel width.

This paper addresses the transport of a charged solute Brownian particle entrained in an electro-osmotic flow occurring within a curved serpentine channel of uniform width, modeled as a two-dimensional spatially periodic conduit. (The periodicity assumption, reflecting the configurationally serpentine character of the device, removes the need to deal with finite size "end effects" in real microfluidic devices, at least for channels consisting of a sufficiently large number of turns.) In contrast to the

specific conduit geometries (toroids and helices) analyzed to date for pressure-driven flows in curvilinear channels of circular cross section [2, 8], we consider here the general case of an arbitrary nonuniform local curvature—that is, where channel curvature varies with distance along the center-line of the serpentine device. Attention is focused upon situations where the channel width is small compared with the characteristic local radius of curvature,[1] which circumstances are assumed (either explicitly or implicitly) in all previous attempts to analyze turn-induced dispersion [5, 10, 16]. The present contribution entails a rigorous description of the generic channel geometry via channel-fixed curvilinear coordinates. This allows for the systematic use of regular asymptotic expansions of the pertinent transport fields, enabling a straightforward perturbation solution of the coupled flow-electrostatics problem.

Using the solvent velocity and electric fields thereby derived, generalized Taylor–Aris dispersion theory [3] is invoked to evaluate the effective macrotransport coefficients $\overline{U}^*$ and $\overline{D}^*$ serving to quantify the net unidirectional global solute transport through the serpentine device as a whole. Since Taylor dispersion is absent during electro-osmotic "plug flow" in a straight channel (at least in circumstances where the Debye layer is thin compared with the channel width and where wall effects are negligible), the counterparts of the "zeroth-order" tortuosity-induced effects, analyzed by [22] for pressure-driven flows, are trivial in the present case. Thus, leading-order Taylor dispersion is associated with (first-order) curvature-induced velocity deviations from a plug flow. Owing to the dependence of curvature upon curvilinear position along the channel, both $\overline{U}^*$ and $\overline{D}^*$ are expressed as quadratures of the local channel curvature $k(s)$ within a unit cell of the serpentine device, s being a length parameter measured along the (center-line of the) curved channel. Inasmuch as typical solute particles (such as DNA molecules) encountered in microfluidic devices are small compared with the device's channel width [24], we concentrate—in this initial communication—on the case of "point-size" particles. This furnishes the dominant effect upon $\overline{D}^*$. Wall effects upon both $\overline{U}^*$ and $\overline{D}^*$, arising from the finite size of particles relative to channel width, will be discussed in a subsequent publication.

Section 2 deals with the specification of a general curvilinear channel geometry, one which lends itself to analytic manipulation. Explicitly, channel-fixed tangent-normal coordinates are constructed, and the required scalar and vector operators evaluated. Description of the physical problem governing both the flow and electric potential fields, as well as the accompanying asymptotic analysis, is outlined in section 3 for an arbitrary two-dimensional curved channel geometry. This analysis provides both the electric and vector velocity fields, expressed in the respective forms of regular perturbation expansions in the small parameter ϵ, representing the dimensionless curvilinear channel width. Using these data, in section 4 we perform the requisite macrotransport analyses, resulting in evaluation of the macrotransport coefficients $\overline{U}^*$ and $\overline{D}^*$ in terms of the specified global parameters characterizing the system as a whole. In section 5 we generalize the preceding two-dimensional analysis to a three-dimensional configuration, wherein the serpentine channel possesses a circular cross section of uniform radius. The relationship between the present macrotransport results and the single-turn models which appear in the literature is discussed in section 6. Results are summarized at section 7.

[1] Cases lacking this fundamental disparity in scales are amenable only to numerical analysis, which cannot be expected to provide qualitative insights comparable to those resulting from the present analytical investigation.

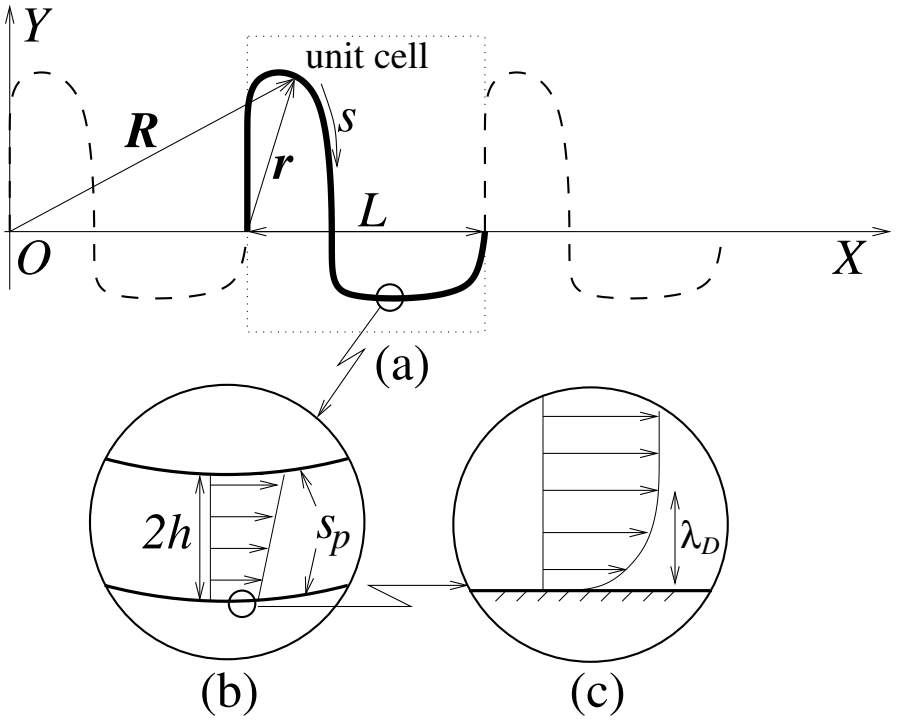


FIG. 1. *Serpentine geometry:* (a) *L-scale;* (b) *h-scale; and* (c) *Debye layer scale.*

2. Channel geometry. The present analysis addresses electro-osmotic transport processes occurring in a spatially periodic two-dimensional curved channel of uniform width, depicted in Figure 1. The unidirectional periodicity of the channel is captured by the presence of a repetitive "unit cell," characterized by the lattice vector $L\hat{\boldsymbol{X}}$, with $\hat{\boldsymbol{X}}$ a unit vector in the X-direction and L the unit-cell rectilinear length, measured along X. The curvilinear arc-length of the channel along its center-line for one period is denoted by l ($l \geq L$, with equality holding only for a straight channel). The "global" position vector $\boldsymbol{R}_n$ of the nth unit cell ($-\infty < n < \infty$) relative to an arbitrary origin O (situated, say, at the entrance to the cell labeled $n = 0$) is given by $\boldsymbol{R}_n = nL\hat{\boldsymbol{X}}$. The vector $\boldsymbol{r}$ is used to denote the "local" position within the fluid domain of this particular cell. The position vector $\boldsymbol{R} = (X, Y)$ of a point lying within the nth cell may thus be represented as $\boldsymbol{R}_n + \boldsymbol{r}$.

In order to define the channel geometry unambiguously it is necessary to specify the configuration of the boundaries (in some parametric form) of a single unit cell. We begin with a plane curve Γ, serving to define the center-line of the channel. This curve is given by the following parametric description of the local position vector

$$\boldsymbol{r} = \boldsymbol{r}_c(s), \tag{2.1}$$

wherein the parameter s denotes arc-length measured along Γ ($0 \leq s \leq l$). The unit vector normal to the osculating plane (pointing out of the page) is denoted by $\hat{\boldsymbol{k}}$, whereas the unit vector tangent to Γ is given by

$$\hat{\boldsymbol{s}} = \frac{d\boldsymbol{r}_c}{ds}. \tag{2.2}$$

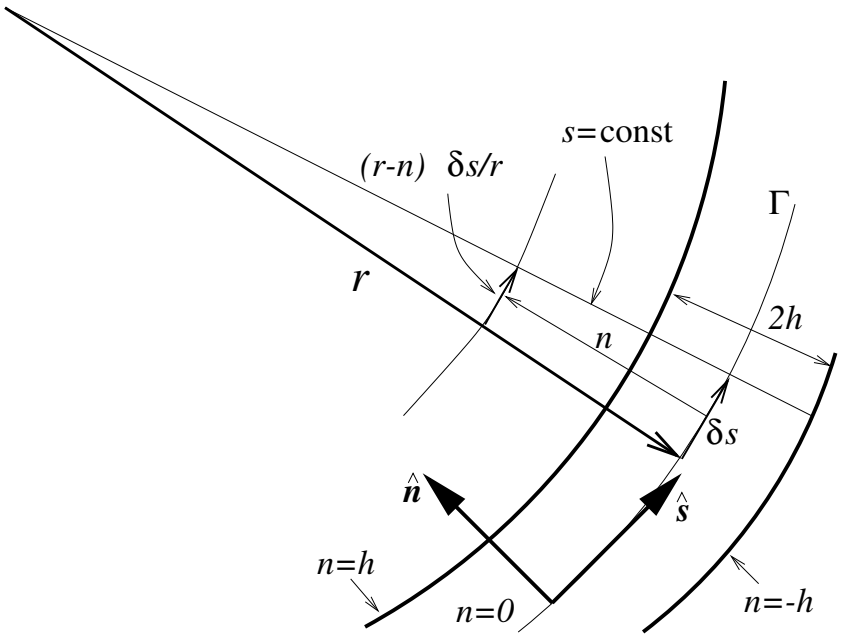


FIG. 2. *Tangent-normal local curvilinear orthogonal coordinate system.*

The principal unit normal to Γ, namely $\hat{\boldsymbol{n}}$, is defined by the requirement that the triad $(\hat{\boldsymbol{s}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{k}})$, in that order, form a right-handed orthonormal system. The unit vectors $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{n}}$ are related by the Serret–Frenet relations [1],

$$\frac{d\hat{\boldsymbol{s}}}{ds} = k\hat{\boldsymbol{n}}, \tag{2.3a}$$

$$\frac{d\hat{\boldsymbol{n}}}{ds} = -k\hat{\boldsymbol{s}}, \tag{2.3b}$$

wherein $k(s)$ is the (algebraically signed) curvature.

The lateral boundaries of the serpentine channel, denoted by s_p, are defined as the locus of points situated at a distance h from Γ. Every point on Γ is thus associated with two boundary points, located at distances h in the $\hat{\boldsymbol{n}}$ and $-\hat{\boldsymbol{n}}$ directions, respectively.[2] The length $2h$ is thus to be interpreted as the channel "width." Moreover, this procedure may be generalized to construct a local system of orthogonal, "tangent-normal," coordinates (s, n). This is achieved by applying the same type of translation from points on Γ, but at an arbitrary, algebraically signed distance, say n, normal to Γ. Every point $\boldsymbol{r}$ in the vicinity of the center-line may be identified with a pair of coordinates s (signifying distance traversed along the center-line) and n. These local coordinates are defined only within the fluid domain, say τ_0, of a single unit cell $(0 \le s \le l,\ -h \le n \le h)$. The geometry of the channel is schematically presented in Figure 2 (for a positive curvature).

The family of curves $s = \text{const}$ constitutes a collection of straight lines, which

[2]These boundaries are well defined only for $h < |k|^{-1}$. In the following analysis, however, we concentrate on the asymptotic limit $h \ll |k|^{-1}$. In that case the straight segments normal to Γ do not intersect. As such, any ambiguity is appropriately removed to a degree consistent with our eventual asymptotic solutions.

is atypical for coordinate systems normally used to describe curvilinear geometries. The appearance of such straight coordinate curves in the present context is related to the fact that only lateral "width" constraints are involved in the specification of the channel boundaries. As the orthogonal pair of tangent and normal unit vectors, say, $\hat{\boldsymbol{s}}(s,n)$ and $\hat{\boldsymbol{n}}(s,n)$, are independent of n, each is identical to its corresponding center-line value, namely, $\hat{\boldsymbol{s}}(s)$ and $\hat{\boldsymbol{n}}(s)$.

By definition, the local position vector at the point (s,n) is given by

$$\boldsymbol{r}(s,n) = \boldsymbol{r}_c(s) + n\,\hat{\boldsymbol{n}}(s). \tag{2.4}$$

Use of (2.3) yields the differential vector displacement between the neighboring points (s,n) and $(s+ds, n+dn)$, namely,

$$d\boldsymbol{r} = \hat{\boldsymbol{s}}(1-nk)\,ds + \hat{\boldsymbol{n}}\,dn. \tag{2.5}$$

The metrical coefficient, $|d\boldsymbol{r}|/ds$, associated with the coordinate s is thus given by $1-nk$. Since the (algebraically signed) radius of curvature is given by $r = 1/k$, this coefficient may be expressed as $(r-n)/r$, a result which was to be anticipated (see Figure 2). The gradient operator appropriate to the tangent-normal coordinate system is thus given by

$$\boldsymbol{\nabla} = \hat{\boldsymbol{n}}\frac{\partial}{\partial n} + \hat{\boldsymbol{s}}\frac{1}{1-nk}\frac{\partial}{\partial s}. \tag{2.6}$$

This result may also be obtained from the invariant definition of the gradient of a generic function f, namely,

$$\boldsymbol{\nabla} f \triangleq \lim_{V\to 0} \frac{\int_{\partial V} dA\; \boldsymbol{n}\, f}{V},$$

wherein $\boldsymbol{n}$ is a unit vector normal to the boundary ∂V of V; see the elementary "volume" element depicted in Figure 3. Use of (2.6) in conjunction with (2.3) yields [11] the following expression for the Laplacian of a generic scalar field $f(n,s)$,

$$\nabla^2 f = \frac{\partial^2 f}{\partial n^2} - \frac{k}{1-nk}\frac{\partial f}{\partial n} + \frac{1}{(1-nk)^2}\frac{\partial^2 f}{\partial s^2} + \frac{n}{(1-nk)^3}\frac{dk}{ds}\frac{\partial f}{\partial s}. \tag{2.7}$$

Note that the generic geometric construction outlined in this section is, in fact, not restricted to periodic configurations. In particular, if k is uniform along the channel (i.e., s-independent), the curve Γ forms a circular arc, possessing a radius $r = 1/k$. The operators (2.6)–(2.7) then degenerate to their respective polar-coordinate counterparts, wherein the radial and azimuthal coordinates, say ρ and θ, are given by $r-n$ and s/r, respectively. (The latter operators are, obviously, independent of k.)

3. Electrokinetics. Consider a periodic electrolyte-filled serpentine channel, whose nonconducting boundary is assumed to possess a uniform surface charge density. Electric potential difference is applied across the two ends of the channel, resulting in the establishment of an electric field, $\boldsymbol{E} = -\boldsymbol{\nabla}\phi$.

The presence of different ionic species, combined with the surface charge, results in regions near the boundaries displaying sharp variations in electric field intensity and ionic species concentrations. This region constitutes the Debye double-layer. The length-scale, say, λ_D (see Figure 1), associated with this variation is of the order of

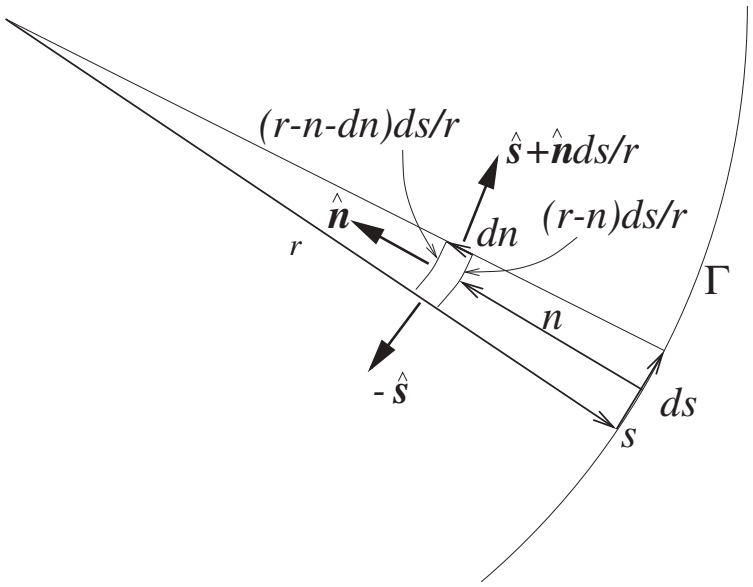


FIG. 3. *Elementary "volume" (areal) element centered about* (n, s).

nanometers for typical values of buffer ionic concentration [21]. This length is extremely small compared with typical channel widths, the latter being of the order of tens of microns [5]. We therefore assume, henceforth, that $\lambda_D/h \to 0$. The matching conditions for the ensuing λ_D-scale "double-layer" fields provide the pertinent boundary conditions for the h-scale analysis [14]. The first is that the electric field normal to the boundary vanishes:

$$\frac{\partial \phi}{\partial n} = 0; \tag{3.1}$$

the second is the well-known Smoluchowski equation,

$$\boldsymbol{v} = \frac{\varepsilon_{el}\zeta}{\mu}\boldsymbol{\nabla}\phi, \tag{3.2}$$

relating the (local) values of the fluid velocity $\boldsymbol{v}$ to the electric field. In the above, ε_{el} is the electrolyte permittivity and μ the (presumably uniform) fluid viscosity. The "zeta potential" ζ is the (uniform) excess potential of the charged surface relative to the (local) value of the potential outside the Debye layer. Equation (3.2) serves as a "slip-condition" imposed upon the velocity field.

The flow domain in the present problem thus corresponds to the "outer" region of the double-layer, wherein deviations from charge neutrality are exponentially small. In this region the Poisson equation degenerates to the Laplace equation,

$$\nabla^2 \phi = 0, \tag{3.3}$$

and electrical body forces are absent from the Navier–Stokes equations. For typical Reynolds numbers encountered in microfluidic devices it is common to assume Stokes flow [24], whence the pertinent flow equations become

$$\boldsymbol{\nabla} \cdot \boldsymbol{v} = 0, \tag{3.4}$$

$$\mu \nabla^2 \boldsymbol{v} = \boldsymbol{\nabla} p, \tag{3.5}$$

wherein $p(\boldsymbol{r})$ is the pressure field.

Owing to the periodicity of the channel, the electric field and local pressure gradient are both spatially periodic, resulting in the following conditions:

$$\boldsymbol{\nabla}\phi(\boldsymbol{R}+L\hat{\boldsymbol{X}})-\boldsymbol{\nabla}\phi(\boldsymbol{R})=\boldsymbol{0},$$
$$\boldsymbol{\nabla}p(\boldsymbol{R}+L\hat{\boldsymbol{X}})-\boldsymbol{\nabla}p(\boldsymbol{R})=\boldsymbol{0},$$

which are each valid for all points $\boldsymbol{R}$ in the fluid domain of the serpentine channel. The former condition is equivalent to the requirement that

$$\phi(\boldsymbol{R})=\tilde{\phi}(\boldsymbol{R})+V\frac{X}{L},$$

where $\tilde{\phi}(\boldsymbol{R})$ is a spatially periodic function (possessing a period of length L in the $\hat{\boldsymbol{X}}$ direction), and V is a constant, representing the voltage difference across a unit cell. The corresponding "jump" condition, expressed in the local coordinate system, is simply

$$\phi(n,s=l)-\phi(n,s=0)=V \qquad \forall n\in[-h,h]\,. \tag{3.6}$$

The analogous condition imposed upon p, reflecting the absence of an externally applied pressure difference, is

$$p(n,s=l)-p(n,s=0)=0 \qquad \forall n\in[-h,h]\,. \tag{3.7}$$

Owing to the periodicity of the serpentine geometry, we expect on the basis of (3.7) that the resulting flow field will be periodic; that is,

$$\boldsymbol{v}(n,s=l)-\boldsymbol{v}(n,s=0)=\boldsymbol{0} \qquad \forall n\in[-h,h]\,. \tag{3.8}$$

This, however, does not constitute an additional condition, inasmuch as the problem is already uniquely defined by (3.1)–(3.7). As such, (3.8) will be satisfied automatically.

Inasmuch as any irrotational flow field identically satisfies the Stokes equations (with no pressure gradients), it is obvious that the electro-osomtic velocity field is simply proportional to the electric field, namely,

$$\boldsymbol{v}\equiv\frac{\varepsilon_{el}\zeta}{\mu}\boldsymbol{\nabla}\phi.$$

The similarity between the two fields in the present problem arises from: (i) the lack of any pressure differences in the system (cf. (3.7)) and (ii) the uniformity of the zeta potential over the channel walls. This so-called "similitude" was discussed by [19] for the case of electro-osmosis and by [18] for the case of electrophoresis; it reduces the present electrokinetic analysis to the solution of the Neumann-type boundary-value problem governing the electric potential.

3.1. Normalization. We normalize length variables with h, velocities with the characteristic electro-osmotic velocity,

$$v_0=\frac{\varepsilon_{el}\zeta}{\mu}\frac{V}{l}, \tag{3.9}$$

and the electric potential with V. The dimensionless unit-cell potential problem thereby posed is thus governed by Laplace's equation in the fluid domain,

$$\nabla^2\phi=0, \tag{3.10}$$

the Neumann-type boundary condition,

$$\hat{\boldsymbol{n}} \cdot \boldsymbol{\nabla}\phi = 0 \quad \text{at} \quad n = \pm 1, \tag{3.11}$$

and the periodicity conditions,

$$\phi(n, s = \epsilon^{-1}) - \phi(n, s = 0) = 1 \qquad \forall n \in [-1, 1]\,. \tag{3.12}$$

In the above,

$$\epsilon \triangleq \frac{h}{l} \tag{3.13}$$

is the dimensionless "aspect-ratio" of the unit cell. Once this problem is solved, the fluid velocity field is given by

$$\boldsymbol{v} = \frac{1}{\epsilon}\boldsymbol{\nabla}\phi. \tag{3.14}$$

Consider the asymptotic limit $\epsilon \ll 1$, corresponding to the case of "thin" channels. It is obvious in such circumstances (see (3.12)) that the appropriate choice of a dimensionless center-line coordinate is $S = \epsilon s$, rather[3] than s. We focus on the case where the center-line radius of curvature is of order l; we thus normalize the dimensional curvature (now a function of S) by l^{-1}. The problem reformulation, wherein all fields appearing therein are functions of n, S, and ϵ, consists of the following: (i) Laplace's equation,

$$\frac{\partial^2\phi}{\partial n^2} - \frac{\epsilon k}{1 - \epsilon n k}\frac{\partial\phi}{\partial n} + \frac{\epsilon^2}{(1 - \epsilon n k)^2}\frac{\partial^2\phi}{\partial S^2} + \frac{\epsilon^3 n}{(1 - \epsilon n k)^3}\frac{dk}{dS}\frac{\partial\phi}{\partial S} = 0; \tag{3.15}$$

(ii) the channel walls boundary conditions (at $n = \pm 1$),

$$\frac{\partial\phi}{\partial n} = 0; \tag{3.16}$$

and (iii) the serpentine periodicity condition

$$\phi(n, S = 1; \epsilon) - \phi(n, S = 0; \epsilon) = 1 \qquad \forall n \in [-1, 1]\,. \tag{3.17}$$

3.2. Asymptotic solutions. Expand the ϵ-dependent terms in the Laplace equation (3.15) into Taylor series in ϵ, so as to obtain the expression

$$\begin{aligned}\frac{\partial^2\phi}{\partial n^2} - \epsilon k\frac{\partial\phi}{\partial n} + \epsilon^2\left(\frac{\partial^2\phi}{\partial S^2} - nk^2\frac{\partial\phi}{\partial n}\right) + \epsilon^3\left(2nk\frac{\partial^2\phi}{\partial S^2} - n^2k^3\frac{\partial\phi}{\partial n} + n\frac{dk}{dS}\frac{\partial\phi}{\partial S}\right)\\ + \epsilon^4\left(3n^2k\frac{\partial^2\phi}{\partial S^2} - n^3k^4\frac{\partial\phi}{\partial n} + 3n^2k\frac{dk}{dS}\frac{\partial\phi}{\partial S}\right) + \cdots = 0.\end{aligned} \tag{3.18}$$

Next, expand ϕ into the asymptotic power series

$$\phi(n, S; \epsilon) \sim \phi^{(0)}(n, S) + \epsilon\phi^{(1)}(n, S) + \epsilon^2\phi^{(2)}(n, S) + \cdots\,. \tag{3.19}$$

[3]In general the coordinate S should appear as an additional "slow" variable. However, the emerging solvability conditions (required to eliminate the secular terms appearing in such a multiple-scale approach) result in the absence of any dependence upon the "fast" variable s. This is expected, as no mechanism exists in the present context for pronounced longitudinal variations along the length of the channel.

Substitution of (3.19) into (3.18) yields a system of equations which can be solved recursively in conjunction with use of the various respective orders of the boundary condition (3.16), namely,

$$\frac{\partial \phi^{(i)}}{\partial n} = 0 \quad \text{at} \quad n = \pm 1 \qquad (i = 0, 1, 2, \dots), \tag{3.20}$$

together with the "decomposed" periodicity conditions,

$$\left.\begin{array}{c} \phi^{(0)}(n, S=1) - \phi^{(0)}(n, S=0) = 1, \\ \phi^{(i)}(n, S=1) - \phi^{(i)}(n, S=0) = 0 \qquad (i = 1, 2, \dots) \end{array}\right\} \qquad \forall n \in [-1, 1]\,. \tag{3.21}$$

Owing to the presence of the "slow" variable S, the ensuing solution possesses a straightforward multiple-scale structure. The "solvability conditions" for the S-dependence of each of the functions $\phi^{(i)}$ naturally result upon application of the periodicity condition (3.21) at the ϵ^{i+2}-order problem. The potential ϕ was evaluated to $O\left(\epsilon^4\right)$, which is sufficient for the present macrotransport analysis. Details of the calculations are presented in the appendix. The resulting solution is

$$\begin{aligned} \phi \sim S + \frac{\epsilon^2}{3}\left[S\int_0^1 k^2(x)\,dx - \int_0^S k^2(x)\,dx\right] \\ + \epsilon^3\left[\frac{3n-n^3}{6}\frac{dk}{dS} + C(S)\right] + \epsilon^4\left[\frac{7}{24}(2n^2-n^4)k\frac{dk}{dS} + D(S)\right] + \mathcal{O}(\epsilon^5), \end{aligned} \tag{3.22}$$

wherein $C(S)$ and $D(S)$ are arbitrary periodic functions of S, which do not affect the leading- and first-order macrotransport coefficients. Their S-dependence, which is required if higher-order corrections are sought, may be obtained from the respective analyses of the $\mathcal{O}(\epsilon^5)$ and $\mathcal{O}(\epsilon^6)$ problems. Since the entire unit cell is identified with a line in the leading-order approximation, the corresponding zeroth-order solution for the Laplace equation, appearing in the leading-order term above, was to be expected. It obviously corresponds to a uniform electric field along the channel. The $\mathcal{O}(\epsilon^2)$ term embodies longitudinal nonuniformities in the electric field, resulting from curvature effects. Lateral variations in the electric potential are manifested only at the $\mathcal{O}(\epsilon^3)$ level of approximation.

Substitution of (3.22) into (3.14) provides the following velocity field:

$$u(n, s; \epsilon) \sim \epsilon^2 \frac{1-n^2}{2}\frac{dk}{dS} + \frac{7}{6}\epsilon^3 n(1-n^2)k\frac{dk}{dS} + \mathcal{O}(\epsilon^4), \tag{3.23a}$$

$$v(n, s; \epsilon) \sim 1 + \epsilon n k(S) + \epsilon^2\left[\left(n^2 - \frac{1}{3}\right)k^2(S) + \frac{1}{3}\int_0^1 k^2(x)\,dx\right] + \mathcal{O}(\epsilon^3). \tag{3.23b}$$

The leading-order solution constitutes an electro-osmotic plug flow, corresponding to curvature-free geometry. The first correction to this uniform field (cf. [10]) is a locally homogeneous shear (as in Couette flow), with a positive tangential velocity existing near to the inner boundary.

4. Macrotransport analysis. Following the preceding calculation of the electro-osmotic velocity field, macrotransport theory for spatially periodic systems [3] is used to evaluate the phenomenological coefficients characterizing the transport of a colloidal Brownian particle undergoing convection, diffusion, and electromigration processes. The two solute particle transport coefficients of interest are, respectively,

its mean particle velocity vector $\boldsymbol{U}^*$ and dispersivity dyadic $\overline{\mathbf{D}}^*$. The latter quantity, stochastically characterizing the variance in the actual instantaneous particle position relative to its mean position $\overline{\boldsymbol{U}}^* t$, results from the (longitudinal) molecular diffusion of the particle (characterized by the diffusion coefficient D) as well as from the convective Taylor dispersion mechanism. This dispersion arises from the lateral (and perhaps longitudinal) flow and electric field nonuniformities in conjunction with the molecular diffusion. While the Taylor mechanism may dominate for certain convective-diffusive processes, it is obviously absent during curvature-free electro-osmotic plug flow occurring in straight channels. In the present case, where net (i.e., "global") solute motion occurs only in the X-direction, and where $\hat{\boldsymbol{X}}$ is the only vector involved in the global specification of the serpentine geometry, it is anticipated (and confirmed; cf. (4.31), (4.43)) that $\overline{\boldsymbol{U}}^*$ and $\overline{\mathbf{D}}^*$ possess the respective "unidirectional" forms

$$\overline{\boldsymbol{U}}^* = \hat{\boldsymbol{X}}\overline{U}^*, \quad \overline{\mathbf{D}}^* = \hat{\boldsymbol{X}}\hat{\boldsymbol{X}}\overline{D}^*. \tag{4.1}$$

Taylor–Aris macrotransport analyses are asymptotic in nature, valid only when the diffusive unit-cell sampling time-scale, l^2/D, is small compared with the "global" solute holdup time, $\mathcal{O}(Nl/v_0)$, occurring within the entire serpentine system (of projected length NL, where N denotes the number of periods comprising the finite-size chip). As such, these analyses naturally conform with the present model of a serpentine device, consisting of numerous turns, $N \gg 1$. The quantities $\overline{\boldsymbol{U}}^*$ and $\overline{\mathbf{D}}^*$, which characterize the transient global transport of solute through the serpentine system, are evaluated using the local solutions of a pair of steady-state equations within the unit-cell domain τ_0, these being defined in the subsequent paragraphs. In that context we temporarily abandon the dimensionless notation previously used.

Evaluation of $\overline{\boldsymbol{U}}^*$, using the macrotransport scheme of [3] for periodic systems, requires calculating the long-time probability flux density vector field $\boldsymbol{J}_0^\infty(\boldsymbol{r})$, governed by the steady-state conservation equation

$$\boldsymbol{\nabla} \cdot \boldsymbol{J}_0^\infty = 0. \tag{4.2}$$

Constitutively, this flux vector possesses the convective-diffusive form,

$$\boldsymbol{J}_0^\infty = \boldsymbol{U}(\boldsymbol{r}) P_0^\infty - \mathbf{D}(\boldsymbol{r}) \cdot \boldsymbol{\nabla} P_0^\infty, \tag{4.3}$$

wherein $P_0^\infty(\boldsymbol{r})$ is the long-time intracellular probability density field, $\mathbf{D}$ is the local molecular diffusivity dyadic, and $\boldsymbol{U}$ is the local velocity vector of the solute particle when situated at $\boldsymbol{r}$. On the lateral boundaries s_p of the unit cell the flux vector satisfies the no-flux condition,

$$\hat{\boldsymbol{n}} \cdot \boldsymbol{J}_0^\infty = 0 \quad \text{on} \quad s_p, \tag{4.4}$$

whereas at the endpoints of the unit cell, P_0^∞ satisfies the jump condition,

$$P_0^\infty(n, s = 0) = P_0^\infty(n, s = l) \qquad \forall n \in [-h, h]\,. \tag{4.5}$$

In addition, P_0^∞ satisfies the normalization condition

$$\int_{\tau_0} P_0^\infty \, d^2\boldsymbol{r} = 1, \tag{4.6}$$

with τ_0 the unit-cell domain and $d^2\boldsymbol{r}$ a two-dimensional "volume" element. Knowledge of the resulting field $P_0^\infty(\boldsymbol{r})$, defined by the above system of equations, enables one to compute $\overline{\boldsymbol{U}}^*$ via the unit-cell quadrature,

$$\overline{\boldsymbol{U}}^* = \int_{\tau_0} \boldsymbol{J}_0^\infty \, d^2\boldsymbol{r}. \tag{4.7}$$

In the presence of an external force field $\boldsymbol{F}$ acting on the particle, the field $\boldsymbol{U}$ required in (4.3) is represented as

$$\boldsymbol{U} = \boldsymbol{U}'(\boldsymbol{r}) + \mathbf{M}(\boldsymbol{r}) \cdot \boldsymbol{F}(\boldsymbol{r}), \tag{4.8}$$

wherein $\boldsymbol{U}'(\boldsymbol{r})$ is the velocity of the particle in the absence of the force (which, in general, may differ from the undisturbed solvent velocity field $\boldsymbol{v}(\boldsymbol{r})$, owing to nonzero particle-size wall effects) and $\mathbf{M}$ is the particle mobility dyadic.

Calculation of $\overline{\mathbf{D}}^*$ requires knowledge of the so-called vector $\boldsymbol{B}(\boldsymbol{r})$-field, arising from deviations of the particle position $\boldsymbol{R}$ at time t from its mean value, $\overline{\boldsymbol{U}}^* t$, at that time. The $\boldsymbol{B}$-field is governed by the steady-state equation

$$\boldsymbol{\nabla} \cdot (P_0^\infty \mathbf{D} \cdot \boldsymbol{\nabla} \boldsymbol{B}) - \boldsymbol{J}_0^\infty \cdot \boldsymbol{\nabla} \boldsymbol{B} = P_0^\infty \overline{\boldsymbol{U}}^*, \tag{4.9}$$

and is subject to the jump condition

$$\boldsymbol{B}(n, s = l) - \boldsymbol{B}(n, s = 0) = -L\hat{\boldsymbol{X}} \qquad \forall n \in [-h, h] \tag{4.10}$$

as well as the no-flux condition

$$\hat{\boldsymbol{n}} \cdot (P_0^\infty \mathbf{D} \cdot \boldsymbol{\nabla} \boldsymbol{B}) = \mathbf{0} \quad \text{on} \quad s_p. \tag{4.11}$$

Equations (4.9)–(4.11) serve to define the $\boldsymbol{B}$-field, albeit only to within an arbitrary additive constant. This constant proves irrelevant, since the evaluation of $\overline{\mathbf{D}}^*$ through the following unit-cell quadrature involves only the gradients of this field. Explicitly,

$$\overline{\mathbf{D}}^* = \int_{\tau_0} P_0^\infty \, (\boldsymbol{\nabla} \boldsymbol{B})^\dagger \cdot \mathbf{D} \cdot (\boldsymbol{\nabla} \boldsymbol{B}) \, d^2\boldsymbol{r}. \tag{4.12}$$

In the present "point-size" particle case, wall effects are neglected. In such circumstances the mobility (as well as the diffusivity) is isotropic and spatially uniform (that is, $\mathbf{M}(\boldsymbol{r}) \equiv M\mathbf{I}$, $\mathbf{D}(\boldsymbol{r}) \equiv D\mathbf{I}$, with $\mathbf{I}$ the dyadic idemfactor), and the solute velocity is accordingly given by

$$\boldsymbol{U}(\boldsymbol{r}) = \boldsymbol{v}(\boldsymbol{r}) + M\boldsymbol{F}(\boldsymbol{r}). \tag{4.13}$$

Consider the transport of a nonconducting colloidal Brownian particle possessing a uniform surface charge density. Owing to this charge, the particle motion is governed by both the bulk flow and the electric field, the latter representing the present counterpart of the generic external force field appearing in (4.13). The size of the particle is assumed small compared with the width $2h$, corresponding to a "point-size" particle. The (dimensional) particle electrophoretic velocity relative to the carrier fluid is given by $M_e\boldsymbol{E}$, where M_e is the electrophoretic mobility. In the limit of a thin Debye layer at the particle surface, this mobility is given by the well-known Smoluchowski relation [14, 23],

$$M_e = \frac{\varepsilon_{el}\zeta_p}{\mu}, \tag{4.14}$$

wherein ζ_p denotes the (uniform) excess potential, relative to ϕ, at the particle surface.[4] The dimensionless counterpart of (4.13) is thus (see (3.14))

$$\boldsymbol{U} = \epsilon^{-1}(1-\alpha)\boldsymbol{\nabla}\phi, \tag{4.15}$$

where $\alpha = \zeta_p/\zeta$. The case $\alpha = 0$ corresponds to an inert particle, one which is unaffected by the presence of the electric field.

To effect a dimensionless formulation, P_0^∞ is normalized with $1/\tau_0$, where $\tau_0 = 2hl$ is the solvent-filled domain of a unit cell. The solute probability flux (normalized by v_0/τ_0) is then

$$\boldsymbol{J}_0^\infty = \epsilon^{-1}(1-\alpha)P_0^\infty\boldsymbol{\nabla}\phi - Pe^{-1}\boldsymbol{\nabla}P_0^\infty, \tag{4.16}$$

wherein

$$Pe = \frac{hv_0}{D} \tag{4.17}$$

is a hybrid Péclet number.

The dimensionless solute probability density field P_0^∞ is here governed by: (i) the solute conservation equation

$$\boldsymbol{\nabla}\cdot\left[\epsilon^{-1}(1-\alpha)P_0^\infty\boldsymbol{\nabla}\phi - Pe^{-1}\boldsymbol{\nabla}P_0^\infty\right] = 0; \tag{4.18}$$

(ii) the no-flux boundary condition

$$\hat{\boldsymbol{n}}\cdot\left[\epsilon^{-1}(1-\alpha)P_0^\infty\boldsymbol{\nabla}\phi - Pe^{-1}\boldsymbol{\nabla}P_0^\infty\right] = 0 \quad \text{at} \quad n = \pm 1; \tag{4.19}$$

and (iii) the unit-cell normalization condition (cf. (4.6))

$$\frac{1}{2}\int_{\tau_0} P_0^\infty \, d^2\boldsymbol{r} = 1, \tag{4.20}$$

wherein the volume element $d^2\boldsymbol{r}$ has been normalized with lh.

It is natural to normalize $\boldsymbol{B}$ with l. The resulting dimensionless $\boldsymbol{B}$-field thus satisfies the equation

$$Pe^{-1}\boldsymbol{\nabla}\cdot(P_0^\infty\boldsymbol{\nabla}\boldsymbol{B}) - \boldsymbol{J}_0^\infty\cdot\boldsymbol{\nabla}\boldsymbol{B} = \epsilon P_0^\infty\overline{\boldsymbol{U}}^*, \tag{4.21}$$

and is subject to the jump condition

$$\boldsymbol{B}(n, S=1;\epsilon) - \boldsymbol{B}(n, S=0;\epsilon) = -\frac{L}{l}\hat{\boldsymbol{X}}, \qquad \forall n \in [-1,1]\,, \tag{4.22}$$

as well as the no-flux condition

$$P_0^\infty\,\hat{\boldsymbol{n}}\cdot\boldsymbol{\nabla}\boldsymbol{B} = \boldsymbol{0} \quad \text{at} \quad n = \pm 1. \tag{4.23}$$

[4] In circumstances wherein the solute particle is comparable in size to Debye layer thickness, Smoluchowski's approximation breaks down; nevertheless, the relation (4.14) may still be used in such situations, with ζ_p representing an effective zeta potential.

4.1. Mean solute velocity. It is easy to verify, using (3.10) and (3.11), that the trial solution

$$P_0^\infty = 1 \tag{4.24}$$

satisfies the boundary-value problem (4.18)–(4.20). This result is nonintuitive, since both the electric and solvent velocity fields are nonuniform in the n- (and s-) directions. However, it is easily verified that $P_0^\infty = \text{const}$ is, in fact, the only solution of the generic transport problem (4.2)–(4.8), at least for point-size particles, provided that (i) the force field is divergence-free; (ii) the force field has a null normal component on the walls; and (iii) the solvent velocity satisfies an impenetrability condition on the walls. In the present case, satisfaction of conditions (i) and (ii) (see (3.10), (3.11)) is associated with the assumption that $\lambda_D/h \to 0$. Satisfaction of condition (ii) is unique to electrophoretic motion in electro-osmotic flows.

In terms of the local (n, S) coordinate system, the probability density flux vector is given by the expression

$$\begin{aligned} \boldsymbol{J}_0^\infty &= \hat{\boldsymbol{n}}\left[\epsilon^{-1}(1-\alpha)\frac{\partial\phi}{\partial n}P_0^\infty - Pe^{-1}\frac{\partial P_0^\infty}{\partial n}\right] \\ &+ \hat{\boldsymbol{s}}\left[\frac{1-\alpha}{1-\epsilon nk}\frac{\partial\phi}{\partial S}P_0^\infty - \frac{\epsilon Pe^{-1}}{1-\epsilon nk}\frac{\partial P_0^\infty}{\partial S}\right]. \end{aligned} \tag{4.25}$$

Substitution of the preceding expressions for ϕ and P_0^∞ gives, with $\boldsymbol{J}_0^\infty = \hat{\boldsymbol{n}}J_n + \hat{\boldsymbol{s}}J_s$, the component fluxes

(4.26a)
$$J_n = \epsilon^2\frac{1-\alpha}{2}\frac{dk}{dS} + \epsilon^3\frac{7(1-\alpha)}{6}k\frac{dk}{dS}n(1-n^2) + \mathcal{O}\left(\epsilon^4\right),$$

(4.26b)
$$J_s = (1-\alpha) + \epsilon(1-\alpha)nk + \epsilon^2(1-\alpha)\left[\left(n^2-\frac{1}{3}\right)k^2 + \frac{1}{3}\int_0^1 k^2(S)\,dS\right] + \mathcal{O}\left(\epsilon^3\right).$$

The mean solute velocity (normalized with v_0) is given by the expression (cf. (4.7))

$$\overline{\boldsymbol{U}}^* = \frac{1}{2}\int_{\tau_0} \boldsymbol{J}_0^\infty\, d^2\boldsymbol{r}. \tag{4.27}$$

From (4.26), together with the relation $d^2\boldsymbol{r} = (1-\epsilon nk)\,dn\,dS$, we obtain $\overline{\boldsymbol{U}}^*$ in the form of an asymptotic series,

$$\overline{\boldsymbol{U}}^* \sim \overline{\boldsymbol{U}}^{(0)} + \epsilon\overline{\boldsymbol{U}}^{(1)} + \epsilon^2\overline{\boldsymbol{U}}^{(2)} + \mathcal{O}\left(\epsilon^3\right), \tag{4.28}$$

in which

$$\overline{\boldsymbol{U}}^{(0)} = (1-\alpha)\int_0^1 \hat{\boldsymbol{s}}(S)\,dS,$$

$$\overline{\boldsymbol{U}}^{(1)} = \boldsymbol{0},$$

$$\overline{\boldsymbol{U}}^{(2)} = \frac{1}{3}(1-\alpha)\left[\int_0^1 \hat{\boldsymbol{n}}(S)\frac{dk}{dS}\,dS + \left(\int_0^1 k^2(S)\,dS\right)\left(\int_0^1 \hat{\boldsymbol{s}}(S)\,dS\right) - \int_0^1 \hat{\boldsymbol{s}}k^2(S)\,dS\right].$$

The first integral appearing in the above brackets may be transformed via use of the identity

$$\int_0^1 \hat{\boldsymbol{n}}(S)\frac{dk}{dS}\,dS = -\int_0^1 k(S)\frac{d\hat{\boldsymbol{n}}}{dS}\,dS,$$

which follows upon integration by parts in conjunction with the "periodicity" of both $\hat{\boldsymbol{n}}$ and k.

With $\boldsymbol{r}_c$ (see (2.1)) normalized via l, the dimensionless parametric equation for the curve Γ is given by

$$\boldsymbol{r} = \epsilon^{-1}\boldsymbol{r}_c(S). \tag{4.29}$$

Thus, the dimensionless counterparts of (2.2)–(2.3) are

$$\frac{d\boldsymbol{r}_c}{dS} = \hat{\boldsymbol{s}}, \quad \frac{d\hat{\boldsymbol{s}}}{dS} = k\hat{\boldsymbol{n}}, \quad \frac{d\hat{\boldsymbol{n}}}{dS} = -k\hat{\boldsymbol{s}}. \tag{4.30}$$

Use of (4.29) and (4.30) eventually yields

$$\overline{\boldsymbol{U}}^* = \overline{U}^*\hat{\boldsymbol{X}}, \tag{4.31}$$

wherein

$$\overline{U}^* = (1-\alpha)\frac{L}{l}\left[1 + \frac{\epsilon^2}{3}\int_0^1 k^2(S)\,dS + \mathcal{O}(\epsilon^3)\right]. \tag{4.32}$$

The $\mathcal{O}(1)$ term accords with intuitive "tortuosity" effects [22], associated with the ratio l/L (> 1) of the curvilinear/rectilinear distances traversed by the solute particle during the same period of time. For typical microfluidic devices this ratio is $\mathcal{O}(1)$. The coefficient $1-\alpha$ reflects the combined effects of electro-osmosis and electrophoresis, such effects being either counteractive or cooperative, according to whether the channel wall and solute particle possess like $(\alpha > 0)$ or unlike $(\alpha < 0)$ charges, respectively. As $\boldsymbol{v}^{(1)}$ is an odd function of n, the explicit functional dependence of the mean velocity upon the curvature appears only at the $\mathcal{O}(\epsilon^2)$ term.

4.2. Dispersivity. Assume the following trial solution for the $\boldsymbol{B}$-field:

$$\boldsymbol{B} = \hat{\boldsymbol{X}}B(n,S;\epsilon). \tag{4.33}$$

Substitution of the above representation into (4.21)–(4.23), in conjunction with use of (4.24), yields the following set of equations governing B, expressed in terms of the local coordinates:

$$\begin{aligned} Pe^{-1}&\left[\frac{\partial^2 B}{\partial n^2} - \frac{\epsilon k}{1-\epsilon nk}\frac{\partial B}{\partial n} + \frac{\epsilon^2}{(1-\epsilon nk)^2}\frac{\partial^2 B}{\partial S^2} + \frac{\epsilon^3 n}{(1-\epsilon nk)^3}\frac{dk}{dS}\frac{\partial B}{\partial S}\right] \\ &- J_n\frac{\partial B}{\partial n} - \frac{\epsilon}{1-\epsilon nk}J_s\frac{\partial B}{\partial S} = \epsilon\overline{U}^*, \end{aligned} \tag{4.34}$$

$$\frac{\partial B}{\partial n} = 0 \quad \text{at} \quad n = \pm 1, \tag{4.35}$$

$$B(n,S=1)-B(n,S=0)=-\frac{L}{l}\qquad\forall n\in[-1,1]\,. \tag{4.36}$$

The ϵ-dependent terms in (4.34) may be expanded into Taylor series, and terms of like order in ϵ collected together, to obtain

$$\begin{aligned}&Pe^{-1}\frac{\partial^2 B}{\partial n^2}-J_n\frac{\partial B}{\partial n}+\epsilon\left(-Pe^{-1}k\frac{\partial B}{\partial n}-J_s\frac{\partial B}{\partial S}-\overline{U}^*\right)\\&+\epsilon^2\left(-Pe^{-1}nk^2\frac{\partial B}{\partial n}+Pe^{-1}\frac{\partial^2 B}{\partial S^2}-nkJ_s\frac{\partial B}{\partial S}\right)\\&+\epsilon^3\left(-Pe^{-1}n^2k^3\frac{\partial B}{\partial n}+2Pe^{-1}nk\frac{\partial^2 B}{\partial S^2}+Pe^{-1}n\frac{dk}{dS}\frac{\partial B}{\partial S}-n^2k^2J_s\frac{\partial B}{\partial S}\right)\\&+\mathcal{O}\left(\epsilon^4\right)=0.\end{aligned} \tag{4.37}$$

Assume the following trial expansion for B:

$$B(n,S;\epsilon)\sim B^{(0)}(n,S)+\epsilon B^{(1)}(n,S)+\epsilon^2 B^{(2)}(n,S)+\cdots\,. \tag{4.38}$$

Substitution into (4.37) of equations (4.26), (4.32), and (4.38) yields a system of equations which can be solved recursively in conjunction with use of the various orders of the boundary condition (4.35), namely,

$$\frac{\partial B^{(i)}}{\partial n}=0\quad\text{at}\quad n=\pm1\qquad(i=0,1,2,\dots), \tag{4.39}$$

together with the decomposed jump condition (4.36),

$$\left.\begin{array}{c}B^{(0)}(n,S=1)-B^{(0)}(n,S=0)=-L/l,\\B^{(i)}(n,S=1)-B^{(i)}(n,S=0)=0\qquad(i=1,2,\dots)\end{array}\right\}\qquad\forall n\in[-1,1]\,. \tag{4.40}$$

As the solution scheme for the present boundary-value problem closely resembles that already given for the Laplace equation governing ϕ (see appendix), we omit details. Calculations up to $\mathcal{O}(\epsilon^2)$ yield[5]

$$B^{(0)}=-\frac{L}{l}S, \tag{4.41a}$$

$$B^{(1)}=0, \tag{4.41b}$$

$$B^{(2)}=\frac{L}{l}(1-\alpha)Pe\frac{3n-n^3}{3}k(S). \tag{4.41c}$$

Calculation of $\overline{\mathbf{D}}^*$ (normalized with D) requires effecting the quadrature (cf. (4.12))

$$\overline{\mathbf{D}}^*=\frac{1}{2\epsilon^2}\int_{\tau_0}P_0^\infty\,(\boldsymbol{\nabla B})^\dagger\cdot(\boldsymbol{\nabla B})\,d^2\boldsymbol{r}. \tag{4.42}$$

Introduction of the preceding expression obtained for $\boldsymbol{B}$ eventually yields

$$\overline{\mathbf{D}}^*=\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}\,\overline{D}^*, \tag{4.43}$$

[5]Since $\boldsymbol{B}$ is defined only to within an additive constant, we conveniently set $B^{(i)}(n,S=0)=0$.

wherein $\overline{D}^*$, represented as the sum of respective "molecular" and "convective" contributions, is

$$\overline{D}^* = \overline{D}^M + \overline{D}^C, \tag{4.44}$$

in which

$$\overline{D}^M = \left(\frac{L}{l}\right)^2 \left[1 + \frac{\epsilon^2}{3}\int_0^1 k^2(S)\,dS + \mathcal{O}\left(\epsilon^3\right)\right], \tag{4.45a}$$

$$\overline{D}^C = \frac{8\epsilon^2}{15}\left(\frac{L}{l}\right)^2 (1-\alpha)^2 Pe^2 \int_0^1 k^2(S)\,dS + \mathcal{O}\left(\epsilon^3\right). \tag{4.45b}$$

Whereas the (dimensional) molecular dispersion coefficient is directly proportional to the solute particle's molecular diffusivity D, the comparable convective coefficient is inversely proportional to D. The $\mathcal{O}(1)$ term in $\overline{D}^M$ represents the combined "tortuosity" effects of the extended curvilinear channel length and the decreased effective width of the serpentine channel [22]. The $\overline{D}^C$ term of $\mathcal{O}\left(\epsilon^2\right)$ constitutes the natural extension to the constant-radius results of [10].

The convective coefficient (4.45b), which vanishes for curvature-free (i.e., rectilinear) electro-osmotic plug flow, constitutes a nonnegative contribution originating from second-order curvature effects. These arise from comparable first-order curvature corrections to the otherwise uniform flow and electric fields, both of which are laterally antisymmetric (see, e.g., (3.23b)). Whereas, to terms of dominant order, this results in a null contribution to the average solute velocity (see (4.32)), the corresponding variance in the Brownian particle position, which constitutes a quadratic term, does not vanish.

As a simple example of the use of the present results, consider the case where each period of the serpentine channel consists of two semicircular arcs, each possessing a (dimensional) radius $L/4$. (This geometry can also approximate spiral channel configurations; see [6].) The total channel length is therefore given by $l = \pi L/2$, and the curvature is equal to $\pm 4/L$. As would be expected, the discontinuity in k does not affect the value of the integrals in (4.45). The dimensionless magnitude of $|k|$ is equal almost everywhere to $4(l/L)$; this readily yields $\overline{D}^C = 128\epsilon^2(1-\alpha)^2 Pe^2/15$.

5. Circular cross-section. Sections 2–4 furnished the flow and subsequent macrotransport analyses for laterally unbounded serpentine devices, eventually resulting in the macrotransport coefficients $\overline{\boldsymbol{U}}^*$ and $\overline{\mathbf{D}}^*$. While the two-dimensional configuration analyzed thus far is predominant in microfluidic devices [7], the present asymptotic scheme may be easily generalized for three-dimensional flows, emphasizing the generic nature of the analysis. In this section we demonstrate the extension of previous analysis to serpentine devices possessing circular cross sections, which constitute another simple laterally bounded channel geometry.

5.1. Geometry. Consider a serpentine channel of uniform circular cross section, as in Figure 4. As in the comparable two-dimensional case, it is convenient to focus on the planar center-line curve Γ defining the axis of the channel. The boundary s_p of this channel consists of the locus of circles of radius h centered about Γ and lying normal to it. This construction may be generalized to define a "radial" coordinate, say ρ, whereby the surface $\rho = \text{const}$ constitutes the collection of circles of radius ρ centered about Γ and lying normal to it. Consequently, it is natural to adopt a

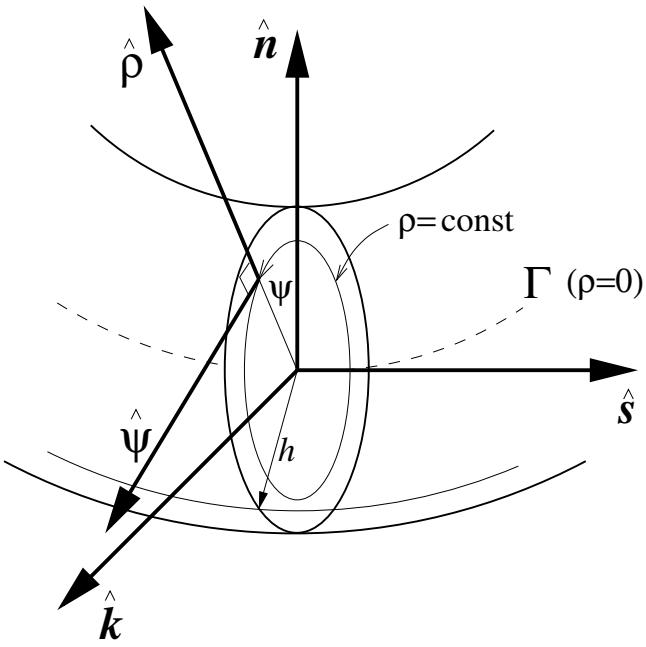


FIG. 4. *Curvilinear cylindrical-like coordinate system.*

system of local cylindrical coordinates, (ρ, ψ, s), appropriate to the configuration of the present serpentine channel. The surface $\psi = \text{const}$ corresponds to the collection of rays originating from Γ and lying normal to it, forming an angle ψ relative to the osculating plane. (The positive sense of ψ is taken to lie in the counter-clockwise direction from $\hat{\boldsymbol{n}}$ to $\hat{\boldsymbol{k}}$.)

Define the respective radial and azimuthal unit vectors associated with the coordinates ρ and ψ,

$$\hat{\boldsymbol{\rho}} = \hat{\boldsymbol{n}}(s)\cos\psi + \hat{\boldsymbol{k}}\sin\psi, \tag{5.1a}$$

$$\hat{\boldsymbol{\psi}} = -\hat{\boldsymbol{n}}(s)\sin\psi + \hat{\boldsymbol{k}}\cos\psi. \tag{5.1b}$$

The triad $(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{s}})$, in that order, form a right-handed orthonormal system. The position vector at the point (ρ, ψ, s) is given by

$$\boldsymbol{r}\,(\rho, \psi, s) = \boldsymbol{r}_c(s) + \rho\hat{\boldsymbol{\rho}}(\psi, s). \tag{5.2}$$

Use of (2.3b) and (5.1) yields the differential vector displacement $d\boldsymbol{r}$ between the neighboring points (ρ, ψ, s) and $(\rho + d\rho, \psi + d\psi, s + ds)$:

$$d\boldsymbol{r} = \hat{\boldsymbol{\rho}}\, d\rho + \hat{\boldsymbol{\psi}}\rho\, d\psi + \hat{\boldsymbol{s}}\,(1 - \rho k\cos\psi)\; ds.$$

The metrical coefficients associated with the coordinates (ρ, ψ, s) are thus given by $(1, \rho, 1 - \rho k\cos\psi)$, respectively, whence the gradient operator appropriate to the serpentine cylindrical coordinate system (cf. (2.6)) is

$$\boldsymbol{\nabla} = \hat{\boldsymbol{\rho}}\frac{\partial}{\partial\rho} + \hat{\boldsymbol{\psi}}\frac{1}{\rho}\frac{\partial}{\partial\psi} + \hat{\boldsymbol{s}}\frac{1}{1 - \rho k\cos\psi}\frac{\partial}{\partial s}. \tag{5.3}$$

Use of (5.3) yields [17] the following expression for the Laplacian of a generic scalar field $f(\rho,\psi,s)$:

$$\nabla^2 f = \frac{\partial^2 f}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial f}{\partial\rho} - \frac{k\cos\psi}{1-\rho k\cos\psi}\frac{\partial f}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2 f}{\partial\psi^2} + \frac{k\sin\psi}{\rho\,(1-\rho k\cos\psi)}\frac{\partial f}{\partial\psi} + \frac{1}{(1-\rho k\cos\psi)^2}\frac{\partial^2 f}{\partial s^2} + \frac{\rho\cos\psi}{(1-\rho k\cos\psi)^3}\frac{dk}{ds}\frac{\partial f}{\partial s}. \tag{5.4}$$

In the case of a uniform (i.e., s-independent) curvature k, these expressions degenerate to their respective counterparts in toroidal coordinates (cf. [2]).

5.2. Electro-osmotic flow. The mathematical formulation of the present electrokinetic problem possesses a structure similar to that encountered in the two-dimensional case, (3.1)–(3.7), as all length variables are normalized here with ρ_0, rather than a. Thus, $\epsilon = \rho_0/l$ (cf. (3.13)). In the nondimensionalization, all length variables are normalized with h, the cylinder radius. With $S = \epsilon s$, the dimensionless reformulation of the potential problem in terms of the local coordinates (ρ,ψ,S) consists of (i) Laplace's equation in the fluid domain:

$$\frac{\partial^2\phi}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial\phi}{\partial\rho} - \frac{\epsilon k\cos\psi}{1-\epsilon\rho k\cos\psi}\frac{\partial\phi}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2\phi}{\partial\psi^2} + \frac{\epsilon k\sin\psi}{\rho\,(1-\epsilon\rho k\cos\psi)}\frac{\partial\phi}{\partial\psi} + \frac{\epsilon^2}{(1-\epsilon\rho k\cos\psi)^2}\frac{\partial^2\phi}{\partial S^2} + \frac{\epsilon^3\rho\cos\psi}{(1-\epsilon\rho k\cos\psi)^3}\frac{dk}{dS}\frac{\partial\phi}{\partial S} = 0; \tag{5.5}$$

(ii) the channel wall boundary conditions (at $\rho = 1$),

$$\frac{\partial\phi}{\partial n} = 0; \tag{5.6}$$

and (iii) the serpentine periodicity conditions

$$\phi(\rho,\psi,S=1;\epsilon) - \phi(\rho,\psi,S=0;\epsilon) = 1 \qquad \forall(\rho,\psi)\in[0,1]\times[0,2\pi]\,. \tag{5.7}$$

In the limit $\epsilon \ll 1$ the above problem may be solved using perturbation expansions in powers of ϵ. The analysis is similar to that of section 3.2, whence we omit details. The electric potential is eventually found to be given by the expression

$$\phi = S + \frac{\epsilon^2}{4}\left[S\int_0^1 k^2(x)\,dx - \int_0^S k^2(x)\,dx\right] + \epsilon^3\left[\frac{3\rho-\rho^3}{8}\frac{dk}{dS}\cos\psi + E(S)\right] + \epsilon^4\left[\frac{1}{192}k\frac{dk}{dS}\left(42\rho^2 - 21\rho^4 + 52\rho^2\cos 2\psi - 26\rho^4\cos 2\psi\right) + F(S)\right] + \mathcal{O}\left(\epsilon^5\right), \tag{5.8}$$

wherein $E(S)$ and $F(S)$ are arbitrary periodic functions of S (cf. (3.22)). Substitution into (3.14) yields the velocity components

$$u = \frac{3\epsilon^2}{8}\frac{dk}{dS}(1-\rho^2)\cos\psi + \frac{\epsilon^3}{48}k\frac{dk}{dS}\left(\rho-\rho^3\right)(21+26\cos 2\psi) + \mathcal{O}\left(\epsilon^4\right), \tag{5.9a}$$

$$v = \frac{\epsilon^2}{8}\frac{dk}{dS}(\rho^2-3)\sin\psi + \frac{13\epsilon^3}{48}k\frac{dk}{dS}\left(\rho^3-2\rho\right)\sin 2\psi + \mathcal{O}\left(\epsilon^4\right), \tag{5.9b}$$

$$w = 1 + \epsilon\rho k(S)\cos\psi + \frac{\epsilon^2}{4}\left[(2\rho^2-1)k^2(S) + 2\rho^2 k^2(S)\cos 2\psi + \int_0^1 k^2(x)\,dx\right] + \mathcal{O}\left(\epsilon^3\right). \tag{5.9c}$$

5.3. Macrotransport analysis. The (dimensional) equations governing the transport of a colloidal Brownian particle are posed by equations (4.2)–(4.15), with minor modifications reflecting the three-dimensional nature of the present problem. To effect the dimensionless formulation, normalize P_0^∞ with $1/\tau_0$, where $\tau_0 = \pi h^2 l$ is the volume of a unit cell. The solute probability flux (normalized with v_0/τ_0) is then given by (4.16), with Pe given by (4.17). Accordingly, P_0^∞ is governed by the solute conservation equation

$$\boldsymbol{\nabla}\cdot\left[\epsilon^{-1}(1-\alpha)P_0^\infty\boldsymbol{\nabla}\phi - Pe^{-1}\boldsymbol{\nabla}P_0^\infty\right] = 0, \tag{5.10}$$

the no-flux boundary condition

$$\hat{\boldsymbol{\rho}}\cdot\left[\epsilon^{-1}(1-\alpha)P_0^\infty\boldsymbol{\nabla}\phi - Pe^{-1}\boldsymbol{\nabla}P_0^\infty\right] = 0 \quad \text{at} \quad \rho = 1, \tag{5.11}$$

and the unit-cell normalization condition (cf. (4.20))

$$\frac{1}{\pi}\int_{\tau_0} P_0^\infty\, d^3\boldsymbol{r} = 1, \tag{5.12}$$

wherein the volume element $d^3\boldsymbol{r}$ has been normalized with $h^2 l$.

The dimensionless $\boldsymbol{B}$-field (normalized with l) is governed by the differential equation

$$Pe^{-1}\boldsymbol{\nabla}\cdot(P_0^\infty\boldsymbol{\nabla}\boldsymbol{B}) - \boldsymbol{J}_0^\infty\cdot\boldsymbol{\nabla}\boldsymbol{B} = \epsilon P_0^\infty\overline{\boldsymbol{U}}^*, \tag{5.13}$$

the boundary condition

$$P_0^\infty\,\hat{\boldsymbol{n}}\cdot\boldsymbol{\nabla}\boldsymbol{B} = \boldsymbol{0} \quad \text{at} \quad \rho = 1, \tag{5.14}$$

and the jump condition

$$\boldsymbol{B}(\rho,\psi,S=1;\epsilon) - \boldsymbol{B}(\rho,\psi,S=0;\epsilon) = -\frac{L}{l}\hat{\boldsymbol{X}} \qquad \forall(\rho,\psi)\in[0,1]\times[0,2\pi]. \tag{5.15}$$

As the arguments leading to the identity (4.24) are independent of the unit-cell specific geometry, the identity continues to prevail in the present problem. The probability flux in terms of the local (ρ,ψ,s) coordinate system is therefore given by

$$\begin{aligned}\boldsymbol{J}_0^\infty &\triangleq \hat{\boldsymbol{\rho}}J_\rho + \hat{\boldsymbol{\psi}}J_\psi + \hat{\boldsymbol{s}}J_s\\ &= \hat{\boldsymbol{\rho}}\epsilon^{-1}(1-\alpha)\frac{\partial\phi}{\partial\rho} + \hat{\boldsymbol{\psi}}\epsilon^{-1}(1-\alpha)\frac{\partial\phi}{\rho\partial\psi} + \hat{\boldsymbol{s}}\frac{1-\alpha}{1-\epsilon\rho k\cos\psi}\frac{\partial\phi}{\partial S}.\end{aligned} \tag{5.16}$$

Substitution of (5.8) yields

(5.17)

$$\begin{aligned}J_\rho &= \frac{3\epsilon^2}{8}(1-\alpha)\frac{dk}{dS}(1-\rho^2)\cos\psi + \frac{\epsilon^3}{48}(1-\alpha)k\frac{dk}{dS}(\rho-\rho^3)(21+26\cos 2\psi) + \mathcal{O}\left(\epsilon^4\right),\\ J_\psi &= \frac{\epsilon^2}{8}(1-\alpha)\frac{dk}{dS}(\rho^2-3)\sin\psi + \frac{13\epsilon^3}{48}(1-\alpha)k\frac{dk}{dS}(\rho^3-2\rho)\sin 2\psi + \mathcal{O}\left(\epsilon^4\right),\\ J_s &= (1-\alpha) + \epsilon(1-\alpha)\rho k\cos\psi + \frac{\epsilon^2}{4}(1-\alpha)\left[2\rho^2k^2 + 2\rho^2k^2\cos 2\psi - k^2 + \int_0^1 k^2(x)\,dx\right]\\ &\quad + \mathcal{O}\left(\epsilon^3\right).\end{aligned}$$

The mean solute velocity is given by the expression (cf. (4.27))

$$\bar{\boldsymbol{U}}^* = \frac{1}{\pi}\int_{\tau_0} \boldsymbol{J}_0^\infty \, d^3\boldsymbol{r}. \tag{5.18}$$

In conjunction with the expression

$$d^3\boldsymbol{r} = (1-\epsilon\rho k\cos\psi)\, d\rho\, d\psi\, dS,$$

for the volume element, substitution of (5.17) into (5.18) eventually yields

$$\bar{\boldsymbol{U}}^* = \hat{\boldsymbol{X}}\bar{U}^*,$$

wherein (cf. (4.31))

$$\bar{U}^* = (1-\alpha)\frac{L}{l}\left[1+\frac{\epsilon^2}{4}\int_0^1 k^2(S)\, dS + \mathcal{O}(\epsilon^3)\right]. \tag{5.19}$$

Similarly to (4.33), we assume the following trial solution for the $\boldsymbol{B}$-field:

$$\boldsymbol{B} = \hat{\boldsymbol{X}} B(\rho,\psi,S;\epsilon).$$

Substitution of the preceding into (5.13)–(5.14) furnishes the differential equation

$$Pe^{-1}\nabla^2 B - J_\rho\frac{\partial B}{\partial\rho} - J_\psi\frac{\partial B}{\rho\partial\psi} - \frac{\epsilon}{1-\epsilon\rho k\cos\psi}J_s\frac{\partial B}{\partial S} = \epsilon\bar{U}^*, \tag{5.20}$$

the boundary condition

$$\frac{\partial B}{\partial\rho} = 0 \quad \text{at} \quad \rho = 1, \tag{5.21}$$

and the jump condition

$$B(\rho,\psi,S=1) - B(\rho,\psi,S=0) = -\frac{L}{l} \qquad \forall(\rho,\psi)\in[0,1]\times[0,2\pi], \tag{5.22}$$

governing the scalar B-field. Expansion of B into an asymptotic series, similar to (4.38), eventually yields the solution

$$B \sim -\frac{L}{l}S + \epsilon^2\frac{L}{l}(1-\alpha)Pe\frac{3\rho-\rho^3}{4} + \mathcal{O}(\epsilon^3). \tag{5.23}$$

Performing the quadrature in the expression for the dispersivity (cf. (4.42)), namely,

$$\overline{\mathbf{D}}^* = \frac{1}{\pi\epsilon^2}\int_{\tau_0} P_0^\infty\, (\boldsymbol{\nabla}\boldsymbol{B})^\dagger\cdot(\boldsymbol{\nabla}\boldsymbol{B})\; d^2\boldsymbol{r}, \tag{5.24}$$

eventually yields the anticipated form, (4.43)–(4.44), wherein (cf. (4.45))

$$\bar{D}^M = \left(\frac{L}{l}\right)^2\left[1+\frac{\epsilon^2}{4}\int_0^1 k^2(S)\, dS + \mathcal{O}\left(\epsilon^3\right)\right], \tag{5.25a}$$

$$\bar{D}^C = \frac{5\epsilon^2}{48}\left(\frac{L}{l}\right)^2(1-\alpha)^2Pe^2\int_0^1 k^2(S)\, dS + \mathcal{O}\left(\epsilon^3\right). \tag{5.25b}$$

6. Comparison with existing literature. Here, we compare our results with those obtained using alternative models of turn-induced dispersion phenomena. Initial investigations (see, e.g., [13]) employed intuitive kinematical models to obtain the "skew" produced by a constant-radius turn, neglecting transverse diffusion. This skew, quantifying the extent of distortion accompanying the introduction of an initially uniform band of solute into a pre-existing flow, was attributed to two additive sources: transverse variation in electric field strength, and overall migration distance within the turn (the so-called "racetrack effect"). Culbertson, Jacobson, and Ramsey [5] accounted for molecular diffusion using empirical expressions, for which the coefficients appearing therein were obtained from a fit with experimental data. Those authors proposed the use of "complementary" turns to reduce dispersion in systems operating at large Péclet numbers.

Griffiths and Nilson [10] provided an analytical model for diffusion effects in a constant-radius turn. Their curvature-induced shear-flow expression is equivalent to the present $v^{(1)}$ (see (3.23b)). The authors have calculated solute variance following a single turn. The variance for small Péclet numbers was evaluated using a method-of-moments formula, whereas that in the opposite extreme of large Péclet numbers was calculated using a kinematical skewing model. A composite expression for the variance, presumably valid for all Péclet numbers, was suggested, having the form of a fractional expression, which degenerates to the desired forms in the respective limits of small and large Péclet numbers. This solution was compared with both numerical Monte Carlo simulations as well as with the experimental data of [5].

In a related paper, Molho et al. [16] presented a two-dimensional single-turn dispersion model. The added variance, resulting from a single turn, was obtained using a method-of-moments approach. These authors postulated an initially symmetric solute band, thus enabling the solution of the advective-diffusion equation. The resulting variance was found to be a function of both Péclet number and turn length. It was further purposed that superposition of this dispersivity with the purely molecular dispersion occurring in the straight segments of the channel would provide a valid model for channel geometries consisting of more than one turn.

Obviously, all of the above-mentioned studies have employed the assumption of small curvature, either explicitly or implicitly (via use of an equivalent expression for the curvature-induced shear $v^{(1)}$). However, owing to the absence of a rigorous specification of the curvilinear geometry, the preceding analyses take account of curvature effects in a rather intuitive manner, thereby embodying only leading-order effects for small curvature.[6] This contrasts with the present asymptotic scheme, which allows for systematic expansions in ϵ.

Considerable effort has been expended in the above-mentioned studies to correlate and otherwise reconcile single-turn models with both simulation and experimental results. The underlying idea in those studies is that the total variance occurring in a multiturn system can be obtained by simply superposing the respective expressions appropriate to the several geometrical elements of which the serpentine system is composed (e.g., single-radius turns and straight segments). This procedure is appropriate for small Péclet numbers, $Pe \ll 1$, where the analyte plug can achieve a stationary state before encountering the next turn.

[6] For example, both the convective and diffusive terms appearing in the solute transport equation of [16] are incomplete owing to the failure to recognize that neither the gradient operator nor the concomitant Laplacian operator possesses "standard" Cartesian forms when expressed in curvilinear coordinates. The missing terms in their expressions (cf. (2.6)–(2.7)) vanish only for zero curvature.

The present analysis does not require that the Péclet number be strictly small. Nevertheless, the applicability of these results to the large Péclet number case is limited by the time-scale disparity underlying macrotransport analyses. The disparity assumption (see the discussion following (4.1)) may be written in the dimensionless form $Pe \ll N\epsilon$. Thus, for practical devices operating at large Pe, many turns would prove necessary for N-independent behavior to be established.

It is therefore obvious that the single- and multi-turn approaches provide similar results in the case of small Péclet number transport in a single turn. There, the diffusive unit-cell time-scale, l^2/D, is small compared with the transit time around the turn, $\mathcal{O}(l/v_0)$. Solute distribution thus achieves a stationary state almost instantaneously, which is exactly the assumption upon which macrotransport theory is based (which, for $Pe \sim \mathcal{O}(1)$, requires $N \gg 1$). In that case, and for constant-radius turns (spatially uniform curvature), the present dispersion coefficient (4.45b) can be expressed as arising from a superposition of "single-turn contributions," each having a form similar to the intermediate expression (15) of [10].

Several of the cited papers dealing with curved geometries have also proposed schemes for minimizing solute dispersion. Obviously, one way would be to increase the radius of curvature, for example, by using spiral geometries [6]. Another interesting method for reducing shear-induced dispersion invokes the use of a sinusoidal wavy channel wall at the inner track of the channel turn [9]. The resulting extension of the inner track length leads to a reduced average velocity on that side of the channel, thus reducing the curvature-induced shear. The wavy geometry lends itself to analytic analysis, the latter having been employed to suggest optimal channel shapes. Finally, a simple and effective way to reduce plug dispersion entails the use of narrow curved segments [20]. Obviously, constant-radius models are inadequate to describe the latter geometry, which is therefore analyzed alternatively using molecular dynamics simulations. This point demonstrates the benefit of extending our robust scheme to more complex geometries, for example, to the case of nonconstant channel width.

7. Concluding remarks. Electrokinetics, in conjunction with macrotransport theory, has been employed within the asymptotic framework of a regular perturbation scheme to obtain the mean velocity and dispersivity of a charged colloidal point-size Brownian particle entrained in an electro-osmotic solvent Stokes flow taking place within a thin serpentine channel. The results obtained apply to arbitrarily shaped (albeit periodic) channel configurations. The present electro-osmotic analysis extends the leading-order tortuosity results previously obtained by [22] (for nonelectro-osmotic, pressure-driven solvent flows). Explicitly, the combined contributions of both velocity and electric field deviations to the serpentine-scale macrotransport coefficients $\overline{U}^*$ and $\overline{D}^*$ have been evaluated. Unsurprisingly, our main contribution to the existing literature resides in the expression obtained for (the convective part of) $\overline{D}^*$, as no velocity variance—and hence no Taylor dispersion—is present in the leading-order electro-osmotic flow occurring in straight channel geometries (at least for thin Debye double-layers and point-size solute particles).

While only leading-order curvature corrections were obtained in the present contributions, the generic asymptotic scheme may be utilized to obtain higher-order terms, thus extending the range of applicability of the macrotransport description. Moreover, since the asymptotic scheme results in *regular* perturbation expansions in ϵ, it is expected that addition of such terms would extend the scheme's range of validity even to $\epsilon \lesssim 1$, thus providing dispersion models which are valid for "tight"

turns. Obviously, the validity of such extensions depends upon the neglect of inertial flow effects. It is also important to note that whereas the common thin-channel assumption is essential for the solution of the electrokinetic problem, this assumption is by no means necessary for the macrotransport analysis. Thus, the present scheme can, in principle, be employed for any periodic cell geometry, possibly requiring numerical solution of Laplace's equation.

The present analysis, while limited to idealized geometries, outlines a generic scheme for the subsequent evaluation of device-scale macrotransport coefficients appropriate to realistic chip geometries, incorporating both electro-osmotic and electrophoretic effects. The present flow-electrostatics solution scheme may also be used in conjunction with more realistic models, involving finite-size solute particles (relative to the channel width), for which steric volume-exclusion effects, as well as hydrodynamic [4, 15] (and perhaps colloidal [14]) wall effects, would serve to modify the preceding macrotransport analysis.

Appendix. Solution of the two-dimensional electrostatic problem.

The $\mathcal{O}(1)$ and $\mathcal{O}(\epsilon)$ problem. The leading-order problem is posed by the requirements that

$$\frac{\partial^2 \phi^{(0)}}{\partial n^2} = 0,$$
$$\frac{\partial \phi^{(0)}}{\partial n} = 0 \quad \text{at} \quad n = \pm 1.$$

This necessitates that

$$\frac{\partial \phi^{(0)}}{\partial n} = 0 \qquad \forall n \in [-1, 1],$$

or, alternatively,

$$\phi^{(0)} = \phi^{(0)}(S).$$

As a consequence of the latter condition, the $\mathcal{O}(\epsilon)$ problem is identical in form to the preceding $\mathcal{O}(1)$ problem, eventually yielding

$$\phi^{(1)} = \phi^{(1)}(S).$$

The $\mathcal{O}(\epsilon^2)$ problem. The solutions of the pertinent equations

$$\frac{\partial^2 \phi^{(2)}}{\partial n^2} = -\frac{d^2 \phi^{(0)}}{dS^2},$$
$$\frac{\partial \phi^{(2)}}{\partial n} = 0 \quad \text{at} \quad n = \pm 1,$$

supplemented by the periodicity condition

$$\phi^{(0)}(n, S = 1) - \phi^{(0)}(n, S = 0) = 1,$$

are

$$\phi^{(2)} = \phi^{(2)}(S)$$

and[7]

$$\phi^{(0)}(S) = S. \tag{A.1}$$

The $\mathcal{O}(\epsilon^3)$ problem. Here, the relevant equations

$$\frac{\partial^2 \phi^{(3)}}{\partial n^2} = -n\frac{dk}{dS} - \frac{d^2\phi^{(1)}}{dS^2},$$
$$\frac{\partial \phi^{(3)}}{\partial n} = 0 \quad \text{at} \quad n = \pm 1,$$

supplemented by the periodicity condition

$$\phi^{(1)}(n, S = 1) - \phi^{(1)}(n, S = 0) = 0,$$

furnish the following solutions:

$$\phi^{(3)}(n, S) = \frac{3n - n^3}{6}\frac{dk}{dS} + C(S)$$

and

$$\phi^{(1)} \equiv 0. \tag{A.2}$$

Here, $C(S)$ is an arbitrary (albeit "periodic"[8]) function of S. Its S-dependence may, if desired, be established from analysis of the $\mathcal{O}(\epsilon^5)$ problem.

The $\mathcal{O}(\epsilon^4)$ problem. Using similar arguments, the governing equations

$$\frac{\partial^2 \phi^{(4)}}{\partial n^2} = \frac{1 - 7n^2}{2}k\frac{dk}{dS} - \frac{d^2\phi^{(2)}}{dS^2},$$
$$\frac{\partial \phi^{(4)}}{\partial n} = 0 \quad \text{at} \quad n = \pm 1,$$

supplemented by the periodicity condition

$$\phi^{(2)}(n, S = 1) - \phi^{(2)}(n, S = 0) = 0,$$

furnish the following solutions:

$$\phi^{(4)}(n, S) = \frac{7}{24}(2n^2 - n^4)k\frac{dk}{dS} + D(S)$$

and

$$\phi^{(2)}(S) = \frac{1}{3}\left[S\int_0^1 k^2(x)\,dx - \int_0^S k^2(x)\,dx\right]. \tag{A.3}$$

Here, $D(S)$ is an arbitrary (albeit "periodic") function of S. If desired, it may be evaluated from analysis of the $\mathcal{O}(\epsilon^6)$ problem.

Acknowledgment. The authors are grateful to Eli Lilly and Company for their financial support and encouragement.

[7]Since ϕ is defined only to within an additive constant, we conveniently set $\phi^{(i)}(n, S = 0) = 0$.
[8]By "periodic" it is meant that $C(0) = C(1)$.

REFERENCES

[1] R. Aris, *Vectors, Tensors, and the Basic Equations of Fluid Mechanics*, Dover, New York, 1962.

[2] S. A. Berger, L. Talbot, and L. S. Yao, *Flow in curved pipes*, Ann. Rev. Fluid Mech., 15 (1983), pp. 461–512.

[3] H. Brenner and D. A. Edwards, *Macrotransport Processes*, Butterworth–Heinemann, Boston, 1993.

[4] H. Brenner and L. J. Gaydos, *The constrained Brownian movement of spherical particles in cylindrical pores of comparable radius: Models of the diffusive and convective transport of solute particles in membranes and porous media*, J. Colloid Interface Sci., 58 (1977), pp. 312–356.

[5] C. T. Culbertson, S. C. Jacobson, and J. M. Ramsey, *Dispersion sources for compact geometries on microchips*, Anal. Chem., 70 (1998), pp. 3781–3789.

[6] C. T. Culbertson, S. C. Jacobson, and J. M. Ramsey, *Microchip devices for high-efficiency separations*, Anal. Chem., 72 (2000), pp. 5814–5819.

[7] S. K. Cummings, R. H. Griffiths, R. H. Nilson, and P. H. Paul, *Conditions for similitude between the fluid velocity and electric field in electroosmotic flows*, Anal. Chem., 72 (2000), pp. 2526–2532.

[8] P. H. Daskopoulos and A. M. Lenhoff, *Dispersion coefficient for laminar flow in curved tubes*, AIChE J., 34 (1988), pp. 2052–2058.

[9] D. Dutta and D. T. Leighton, *A low dispersion geometry for microchip separation devices*, Anal. Chem., 74 (2002), pp. 1007–1016.

[10] S. K. Griffiths and R. H. Nilson, *Band spreading in two-dimensional microchannel turns for electrokinetic species transport*, Anal. Chem., 72 (2000), pp. 5473–5482.

[11] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics*, Prentice–Hall, Englewood Cliffs, NJ, 1965.

[12] S. C. Jacobson, R. Hergenroder, L. B. Koutny, R. J. Warmack, and J. M. Ramsey, *Effects of injection schemes and column geometry on the performance of microchip electrophoresis devices*, Anal. Chem., 66 (1994), pp. 1107–1113.

[13] V. Kasicka, Z. Prusik, B. Gas, and M. Stedry, *Contribution of capillary coiling to zone dispersion in capillary zone electrophoresis*, Electrophoresis, 16 (1995), pp. 2034–2038.

[14] H. J. Keh and J. L. Anderson, *Boundary effects on electrophoretic motion of colloidal spheres*, J. Fluid Mech., 153 (1985), pp. 417–439.

[15] G. M. Mavrovouniotis and H. Brenner, *Hindered sedimentation, diffusion, and dispersion coefficients for Brownian spheres in circular cylindrical pores*, J. Colloid Interface Sci., 124 (1988), pp. 269–283.

[16] J. I. Molho, A. E. Herr, B. P. Mosier, J. G. Santiago, T. W. Kenny, R. A. Brennen, G. B. Gordon, and B. Mohammadi, *Optimization of turn geometries for microchip electrophoresis*, Anal. Chem., 73 (2001), pp. 1350–1360.

[17] P. Moon and D. E. Spencer, *Field Theory Handbook*, Springer-Verlag, New York, 1988.

[18] F. A. Morrison, *Electrophoresis of a particle of arbitrary shape*, J. Colloid Interface Sci., 34 (1970), pp. 210–214.

[19] J. T. G. Overbeek, *Electrokinetic phenomena*, in Colloid Science, H. R. Kruyt, ed., Elsevier, New York, 1952.

[20] B. M. Paegel, L. D. Hutt, P. C. Simpson, and R. A. Mathies, *Turn geometry for minimizing band broadening in microfabricated capillary electrophoresis channels*, Anal. Chem., 72 (2000), pp. 3030–3037.

[21] R. F. Probstein, *Physicochemical Hydrodynamics*, Butterworth, New York, 1989.

[22] B. Rush, K. D. Dorfman, H. Brenner, and S. Kim, *Dispersion by pressure-driven flow in serpentine microfluidic channels*, Ind. Eng. Chem. Res., 41 (2002), pp. 4652–4662.

[23] W. B. Russel, D. A. Saville, and W. R. Schowalter, *Colloidal Dispersions*, Cambridge University Press, Cambridge, UK, 1989.

[24] H. A. Stone and S. Kim, *Microfluidics: Basic issues, applications, and challenges*, AIChE J., 47 (2001), pp. 1250–1254.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1125–1132

GEOMETRICAL OPTICS APPROXIMATION FOR NONLINEAR EQUATIONS*

F. BASS†, V. FREILIKHER†, A. A. MARADUDIN‡, AND V. PROSENTSOV†

Abstract. A wide class of nonlinear equations is studied in the geometrical optics approximation. It is shown that a nonlinear equation with coefficients dependent on the amplitude of the function sought can be reduced to a system of quasi-linear equations of the gas-dynamics type. As an illustration, the Hamilton–Jacobi equation with a specific form of the nonlinear operator has been solved, and the propagation of monochromatic waves and of point source radiation in nonlinear media has been studied.

Key words. nonlinear equations, quasi-linear equations, geometrical optics approximation, Hamilton–Jacobi equation

AMS subject classifications. 47J25, 78A05, 70H20

DOI. 10.1137/S0036139903426617

The geometrical optics approximation is an efficient and popular method in different areas of modern physics (acoustic and electromagnetic wave propagation, quantum mechanics, etc.). Due to its simplicity and reasonable accuracy, this method is frequently applied to the solution of various specific nonlinear problems [1], [2], [3], [4]. In this paper we propose a modification of the geometrical optics approximation applicable to an arbitrary form of a nonlinear equation (differential, integral, or finite difference) when the nonlinearity depends on the modulus of the unknown function.

1. The method. The basic equation of our problem is

$$\widehat{H}\left\{\rho, \mathbf{R}, \frac{\partial}{\partial \mathbf{R}}\right\}\Psi(\mathbf{R}) = 0, \tag{1.1}$$

where $\widehat{H}$ is a nonlinear operator, $\Psi(\mathbf{R})$ is the function to be found, $\mathbf{R} = \{\mathbf{r}(R_1, R_2, R_3), R_4\}$ is the four-dimensional space-time vector, $\mathbf{r}$ is the three-dimensional spatial radius vector, R_4 is the time component, and $\rho = |\Psi|^2$. Note that the operator $\widehat{H}$ is nonlinear with respect to the unknown function. However, it could depend (linearly) on derivatives of any order (including infinite series), which enables not only differential equations, but nonlinear integral and finite difference equations as well, to be studied.

In what follows it is expedient to deal with the integral form of (1.1). To do this we represent $\Psi(\mathbf{R})$ in the form of a Fourier integral

$$\Psi(\mathbf{R}) = \int \widetilde{\Psi}(\mathbf{K}) e^{i\mathbf{KR}} d\mathbf{K}, \tag{1.2}$$

substitute (1.2) in (1.1), and obtain

*Received by the editors April 24, 2003; accepted for publication (in revised form) October 9, 2003; published electronically April 21, 2004.
http://www.siam.org/journals/siap/64-4/42661.html
†The Jack and Pearl Resnick Institute of Advanced Technology, Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel.
‡Department of Physics and Astronomy, University of California, Irvine, CA 92697 (aamaradu@uci.edu).

$$\int K\left\{\rho, \mathbf{R}, \mathbf{R}-\mathbf{R}'\right\} \Psi(\mathbf{R}') d\mathbf{R}' = 0. \tag{1.3}$$

Here the kernel K is given by

$$K(\rho, \mathbf{R}, \mathbf{R}-\mathbf{R}') = \frac{1}{(2\pi)^4} \int H\left\{\rho, \mathbf{R}, i\mathbf{K}\right\} e^{i\mathbf{K}(\mathbf{R}-\mathbf{R}')} d\mathbf{K}, \tag{1.4}$$

where the function $H\left\{\rho, \mathbf{R}, i\mathbf{K}\right\}$ is obtained from the operator $\widehat{H}\left\{\rho, \mathbf{R}, \frac{\partial}{\partial \mathbf{R}}\right\}$by replacing the differential operator $\frac{\partial}{\partial \mathbf{R}}$ by the vector $i\mathbf{K}$. Obviously, the four-dimensional vector $\mathbf{K}$ has the form $\mathbf{K} = \{\mathbf{k}(K_1,K_2,K_3), -\omega)$. Solutions of (1.3) can be represented in the following form:

$$\Psi(\mathbf{R}) = \rho^{1/2}(\mathbf{R}) e^{iS(\mathbf{R})}. \tag{1.5}$$

When the kernel in (1.3) is a sharp function of $(\mathbf{R}-\mathbf{R}')$ in comparison with $\Psi(\mathbf{R}')$, the functions $\rho(\mathbf{R}')$ and $S(\mathbf{R}')$ can be expanded in power series in the vicinity of the point $\mathbf{R}$. Simple (though rather lengthy) calculations yield the following set of equations:

$$\frac{\partial}{\partial \mathbf{R}} \rho \upsilon = 0, \tag{1.6}$$

$$H\left\{\rho, \mathbf{R}, i\mathbf{P}\right\} = 0, \tag{1.7}$$

where $\frac{\partial}{\partial \mathbf{R}}$ is a vector with components $\{\frac{\partial}{\partial R_1}, \ldots, \frac{\partial}{\partial R_4}\}$, $\boldsymbol{\upsilon} = \frac{\partial H}{\partial \mathbf{P}}$, and $\mathbf{P} = \frac{\partial S}{\partial \mathbf{R}}$.

These equations are valid if the following inequalities hold:

$$\begin{aligned} &\frac{R_k}{S\rho}\left|\frac{\partial \rho}{\partial R_k}\right| \ll 1, \\ &\frac{R_k^n}{S^n}\left|\frac{\partial^n S}{\partial R_k^n}\right| \ll 1, \qquad n = 1, 2. \end{aligned} \tag{1.8}$$

Relations (1.8) are the usual conditions for the applicability of the quasi-classical method.

Equations (1.6) and (1.7) are the continuity equation and (nonlinear) Hamilton–Jacobi equation, respectively. Equation (1.7) can by transformed (by differentiation with respect to **R)** into the form of an equation of motion that has the form

$$\boldsymbol{\upsilon}\, \frac{\partial \mathbf{P}}{\partial \mathbf{R}} + \xi \frac{\partial \rho}{\partial \mathbf{R}} = \mathbf{F}, \tag{1.9}$$

where $\xi = \frac{\partial H}{\partial \rho}$ and $\mathbf{F} = -\frac{\partial H}{\partial \mathbf{R}}$. Note that the classical-mechanical interpretation of the parameters $\boldsymbol{\upsilon}$, $\mathbf{P}$, $\mathbf{F}$, and H is that they are the four-dimensional velocity, momentum, force, and Hamilton function, respectively.

The set of geometrical optics equations related to $\mathbf{P}$ is quasi-linear, and it may be transformed to a linear set with the help of a hodographic transformation, when H is independent of $\mathbf{R}$ and $\mathbf{F} = 0$ [5], [6]. The system of equations (1.6) and (1.7)

may be rewritten in a form analogous to gas-dynamics equations [7], [8]. To do this we represent all solutions of (1.7) in the form

$$P_4 = -h(\mathbf{r}, t, \rho, \mathbf{p}), \tag{1.10}$$

($t \equiv R_4$ is time, $\mathbf{r} = \{R_1, R_2, R_3\}, \mathbf{p} = \{p_1, p_2, p_3\}$ is the three-dimensional momentum vector) and then differentiate the equality obtained, (1.10), with respect to $\mathbf{r}$. This yields the following equation of motion:

$$\frac{\partial \mathbf{p}}{\partial t} + \mathbf{u}\frac{\partial \mathbf{p}}{\partial \mathbf{r}} + \xi\frac{\partial \rho}{\partial \mathbf{r}} = \mathbf{f}, \tag{1.11}$$

where

$$\mathbf{u} = \frac{\partial h}{\partial \mathbf{p}}, \quad \xi = \frac{\partial h}{\partial \rho}, \quad \mathbf{f} = -\frac{\partial h}{\partial \mathbf{r}}.$$

The functions $\mathbf{u}$ and $\mathbf{f}$ are the three-dimensional velocity and force, respectively. The continuity equation in these variables reads as

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho \mathbf{u})}{\partial \mathbf{r}} = 0. \tag{1.12}$$

Equations (1.11) and (1.12) look (at least formally) like the well-known system of gas-dynamical equations for the density ρ and $\mathbf{p}$. The important difference is that in our case the Hamiltonian depends (due to the nonlinearity) on ρ and, therefore, the coefficient ξ in the equation of motion (1.11) is defined by the Hamilton function, while in gas dynamics it should be found from some additional conditions—for example, from the equation of state [5]. Note that, while in gas dynamics ξ is equal to the square of the velocity of sound, here it arises from the nonlinearity in the Hamiltonian and has no general physical meaning (ξ can, for example, be negative for some $h(\rho)$). It should also be noted that the vector coefficient $\mathbf{f}$ on the right-hand side of (1.11) (external force in gas dynamics) arises in the case under consideration from the spatial inhomogeneity of the Hamiltonian.

The geometrical optics method developed above can be generalized for more complex equations, such as

$$\widehat{H}\left\{\mathbf{R}, \left(\widehat{H}_i\Psi\right)^n \times \left(\widehat{H}_k\Psi^*\right)^n, \frac{\partial}{\partial \mathbf{R}}\right\}\Psi = 0, \tag{1.13}$$

where $\widehat{H}_i$ and $\widehat{H}_k$ are nonlinear operators introduced by (1.1). It is easy to see that if $\widehat{H}_i = \widehat{H}_k = \text{const}$ and $n = 1$, (1.13) takes the form (1.1).

2. Examples.

2.1. Nonlinear Hamilton–Jacobi equation. If a Hamilton–Jacobi equation can be represented as

$$\widehat{H}\Psi = \widehat{H}_0\left\{\rho, \frac{\partial}{\partial \mathbf{R}}\right\}\Psi(\mathbf{R}) + \widehat{H}'\left\{\rho, \mathbf{R}, \frac{\partial}{\partial \mathbf{R}}\right\}\Psi(\mathbf{R}) = 0, \tag{2.1}$$

where $\left|\widehat{H}_0\Psi\right| \gg \left|\widehat{H}'\Psi\right|$, it is expedient to seek solutions of (1.7) and (2.1) in the form

$$\Psi(\mathbf{R}) = (\rho_0 + \rho'(\mathbf{R}))^{1/2} e^{i(\mathbf{P}_0\mathbf{R} + S'(\mathbf{R}))}, \tag{2.2}$$

where $\mathbf{P}_0$ and ρ are constants, and $|\frac{\partial S'}{\partial \mathbf{R}}| \ll |\mathbf{P}_0|$, $|\rho'| \ll |\rho_0|$. H_0 is independent of $\mathbf{R}$, and hence $\mathbf{F} = 0$. Such a problem arises, for example, in studies of wave propagation in a nonlinear homogeneous on average ($\widehat{H}_0$ is independent of $\mathbf{R}$) medium with weak fluctuations of the dielectric constant.

After substitution of the expression (2.2) into (1.6) and (1.7) and linearization of the equations obtained with respect to S' and ρ', we find the following equations for S' and ρ':

$$\rho'(\mathbf{R}) = -\frac{H'(\rho_0, \mathbf{R}, i\mathbf{P}_0)}{\xi_0} - \frac{\upsilon_{0n}}{\xi_0}\frac{\partial S'(\mathbf{R})}{\partial R_n}, \tag{2.3}$$

$$\beta_m \frac{\partial \rho'(\mathbf{R})}{\partial R_m} = -\rho_0 \left(\frac{\partial \upsilon_{0m}}{\partial P_{0n}} \frac{\partial^2 S'(\mathbf{R})}{\partial R_m \partial R_n} + \frac{\partial \upsilon'_m}{\partial R_m} \right), \tag{2.4}$$

where

$$\beta_m = \frac{\partial(\rho_0 \upsilon_{0m})}{\partial \rho_0}, \qquad \upsilon_{0m} = \frac{\partial H_0(\rho_0, i\mathbf{P}_0)}{\partial P_{0m}},$$
$$\xi_0 = \frac{\partial H_0(\rho_0, i\mathbf{P}_0)}{\partial \rho_0}, \qquad \upsilon'_m = \frac{\partial H'(\rho_0, \mathbf{R}, i\mathbf{P}_0)}{\partial P_{0m}}.$$

Substituting ρ' from (2.3) into (2.4) we obtain the equation for S',

$$\eta_{mn}\frac{\partial^2 S'(\mathbf{R})}{\partial R_m \partial R_n} = \beta_m \frac{\partial H'(\rho_0, \mathbf{R}, i\mathbf{P}_0)}{\partial R_m} - \rho_0 \xi_0 \frac{\partial \upsilon'_m}{\partial R_m} = \Phi(\mathbf{R}), \tag{2.5}$$

where $\eta_{mn} = \xi_0 \rho_0 \frac{\partial \upsilon_{0m}}{\partial P_{0n}} - \upsilon_{0n}\beta_m$. Equation (2.5) is a partial differential equation of the second order with constant coefficients. It can be solved with the help of a Fourier transformation. Assuming the solution in the form

$$S'(\mathbf{R}) = \int \widetilde{S}'(\mathbf{K}) e^{i\mathbf{K}\mathbf{R}} d\mathbf{K} \tag{2.6}$$

and substituting it into (2.5) we obtain

$$\widetilde{S}'(\mathbf{K}) = -\frac{\widetilde{\Phi}(\mathbf{K})}{\eta_{mn} K_m K_n}, \tag{2.7}$$

where $\widetilde{S}'(\mathbf{K})$ and $\widetilde{\Phi}(\mathbf{K})$ are the Fourier transforms of $S'(\mathbf{R})$ and $\Phi(\mathbf{R})$, respectively. Thus, $S'(\mathbf{R})$ can be represented as

$$S'(\mathbf{R}) = \int G(\mathbf{R} - \mathbf{R}')\Phi(\mathbf{R}')d\mathbf{R}', \tag{2.8}$$

where the Green function $G(\mathbf{R} - \mathbf{R}')$ is given by [9]

$$G(\mathbf{R} - \mathbf{R}') = \frac{1}{(2\pi)^4} \int \frac{e^{i\mathbf{K}(\mathbf{R}-\mathbf{R}')}}{\eta_{mn} K_m K_n} d\mathbf{K}, \tag{2.9}$$

and ρ' is found by substituting (2.9) into (2.3). Here the nonlinearity shows itself in the dependence of the tensor η_{mn} on ρ_0 and P_0.

In the case that $\mathbf{R}$ is a one- or three-dimensional vector, results can be obtained immediately:

$$S'(R_1) = \frac{1}{\eta_{11}} \int_0^{R_1} [\beta_1 H'(\rho_0, R_1, iP_{01}) - \rho_0 \xi_0 v_1'] \, dR_1,$$

$$\rho'(R_1) = -\frac{1}{\xi_0} \left[1 + \frac{v_{01}\beta_1}{\eta_{11}}\right] H'(\rho_0, R_1, iP_{01}) + \frac{v_{01}\rho_0 v_1'}{\eta_{11}} \tag{2.10}$$

and

$$S'(\mathbf{r}) = \int G(\mathbf{r} - \mathbf{r}') \left[\beta_m \frac{\partial H'(\rho_0, \mathbf{r}', i\mathbf{p}_0)}{\partial R_m'} - \rho_0 \xi_0 \frac{\partial v_m'}{\partial R_m'}\right] d\mathbf{r}'$$

$$\rho'(\mathbf{r}) = -\frac{H'(\rho_0, \mathbf{r}, i\mathbf{p}_0)}{\xi_0} - \frac{v_{0m}}{\xi_0} \int \frac{\partial G(\mathbf{r} - \mathbf{r}')}{\partial R_m} \left[\beta_m \frac{\partial H'(\rho_0, \mathbf{r}', i\mathbf{p}_0)}{\partial R_m'} - \rho_0 \xi_0 \frac{\partial v_m'}{\partial R_m'}\right] d\mathbf{r}',$$

where the Green function is [7]

$$\mathbf{G}(\mathbf{r} - \mathbf{r}') = \frac{1}{4\pi \sqrt{|\eta| \, \eta_{mn}^{-1} (R_m - R_m')(R_n - R_n')}}. \tag{2.11}$$

In the linear limit, when $\xi_0 = 0$ and $\boldsymbol{v}_0$ is independent of ρ_0, so that $\beta = \boldsymbol{v}_0$, we obtain from (2.5)

$$\boldsymbol{v}_0 \frac{\partial S'}{\partial \mathbf{R}} = -H'(\mathbf{R}). \tag{2.12}$$

If one of the axes (call it R_v) is directed along the vector $\boldsymbol{v}_0$, (2.12) takes the form

$$\frac{\partial S'(R_v, \mathbf{R}_\perp)}{\partial R_v} = -\frac{1}{v_0} H'(R_v, \mathbf{R}_\perp), \tag{2.13}$$

and its solution is

$$S' = -\frac{1}{v_0} \int_0^{R_v} H'(R_v, \mathbf{R}_\perp) dR_v, \tag{2.14}$$

where $R_\perp$ is the three-dimensional vector perpendicular to $\boldsymbol{v}_0$. Hence it follows that the linear case can be reduced to a one-dimensional problem.

2.2. Plane waves in nonlinear media. The fundamental property of a plane wave is that its phase and amplitude are functions of a scalar product $\zeta = \mathbf{R} \bullet \mathbf{q}$, where $\mathbf{q}$ is a (four-dimensional) wave vector. In this case the Hamilton–Jacobi equation has the form

$$H\left\{\zeta, \rho, \frac{\partial S}{\partial \mathbf{R}}\right\} = 0. \tag{2.15}$$

We seek the solution of (2.15) in the form

$$\rho = \rho(\zeta), \; S(\mathbf{R}) = \mathbf{P}_0 \mathbf{R} + \varphi(\zeta). \tag{2.16}$$

Substitution of (2.16) into (1.6) and (1.7) gives

$$H\{\zeta, \rho, p_\zeta, \mathbf{P}_0\} = 0,$$
$$\rho v(\zeta, \rho, p_\zeta, \mathbf{P}_0) = \mathbf{j}, \tag{2.17}$$

where $p_\zeta = \frac{\partial \varphi}{\partial \zeta}$. The second equation in (2.17) is the flux conservation law. The constant $\mathbf{j}$ (flux) should be found from the initial conditions. If at $t = 0$, for example, $\rho = \rho_0$ and $\boldsymbol{v} = \boldsymbol{v}_0$, the flux $\mathbf{j} = \rho_0 \boldsymbol{v}_0$. Equations (2.17) are a system of ordinary differential equations with respect to ρ and p_ζ that in general may have several solutions, some of which could be unstable. This question will be considered in greater detail below.

Now we employ the general theory presented above to study the propagation of a monochromatic wave in a nonlinear medium. Consider a medium whose parameters depend only on ζ, and wave propagation is described by the Helmholtz equation

$$\widehat{H}\Psi(\zeta) = \left[\frac{d^2}{d\zeta^2} + w(\zeta, \rho)\right]\Psi(\zeta) = 0, \tag{2.18}$$

where $w(\zeta, \rho)$ (the permittivity of the medium) is an arbitrary function of its arguments. In this case (1.6) and (1.7) lead to the following equations:

$$p_\zeta^2 = w(\zeta, \rho), \qquad \rho p_\zeta = j. \tag{2.19}$$

This is a generalization of the WKB method to the nonlinear case [10]. Substituting p_ζ from the second equation of the set (2.19) into the first one, we obtain the following equation for ρ:

$$\rho^2 w(\zeta, \rho) = j^2. \tag{2.20}$$

To proceed farther the explicit form of $w(\zeta, p)$ should be specified. Assuming that w has the Kerr form,

$$w(\zeta, \rho) = w_1(\zeta) + w_2(\zeta)\rho, \tag{2.21}$$

and substituting w from (2.21) into (2.20), we obtain a cubic equation for determining ρ:

$$\rho^2[w_1(\zeta) + w_2(\zeta)\rho] = j^2. \tag{2.22}$$

This equation has three solutions. If these solutions are real and positive, phase hysteresis takes place. This situation was investigated in [11]. Note that in our case the step will move; i.e., a shock wave will propagate. When $\rho \gg |\frac{w_1}{w_2}|$, the first term in (2.22) can be neglected, and ρ is given by

$$\rho = \frac{j^{2/3}}{w_2^{1/3}(\zeta)}. \tag{2.23}$$

When ρ is known, S can be found from the second equation of the set (2.19) as

$$S(\zeta) \equiv \int_0^\zeta p_\zeta d\zeta = j^{1/3} \int_0^\zeta \sqrt[3]{w_2(\zeta)}\, d\zeta. \tag{2.24}$$

Combining (2.23) and (2.24), we obtain for ψ

$$\psi(\zeta) = \frac{j^{1/6}}{w_2^{1/6}} \exp\left[ij^{1/3}\int_0^\zeta w_2^{1/3}(\zeta)d\zeta\right]. \tag{2.25}$$

Another limiting case, $\rho \ll |\frac{w_1}{w_2}|$, corresponds to the result obtained in the WKB approximation in the linear theory [10]. One can see that, in contrast to the results of linear geometrical optics, in the case under consideration the solution $\psi(\zeta)$ is defined by the nonlinearity, $w_2(\zeta)$. The functional dependencies of the amplitude and phase on w are also different: in the linear case $\rho \sim w^{1/2}(\zeta)$ (not $\sim w_1^{-1/3}(\zeta)$ as it is in (2.23)), and the integrand in the phase integral is equal to $w_1^{1/2}(\zeta)$ (in contrast to $w_2^{1/3}(\zeta)$ in (2.25)).

2.3. Point-like antenna in a nonlinear medium. We next consider the radiation of a point-like source in a spherically symmetric nonlinear medium. When the antenna is located at the origin ($r = 0$), the monochromatic spherically symmetric field is described by the Helmholtz equation

$$\Delta\psi(r) + w(r,\rho)\psi(r) = \delta(r). \tag{2.26}$$

Assuming that $w(r,\rho)$ is finite for all r and ρ, we will seek the spherically symmetric solution that at $r \neq 0$ satisfies the equation

$$\frac{1}{r^2}\frac{d}{dr}\frac{d\psi(r)}{dr} + w(r,\rho)\psi(r) = 0, \tag{2.27}$$

which after the substitution $\psi(r) = \widetilde{\psi}(r)/r$ takes the form

$$\frac{d^2\widetilde{\psi}(r)}{dr^2} + w(r,\omega)\widetilde{\psi}(r) = 0,\ (\widetilde{\rho} = \left|\widetilde{\psi}\right|^2). \tag{2.28}$$

Equation (2.28) is similar to (2.18), which was investigated previously. Therefore, if we represent the spherically symmetric solution of (2.26) as

$$\psi(r) = \frac{\widetilde{\rho}^{1/2}}{r}e^{iS}, \tag{2.29}$$

$\widetilde{\rho}$ and S can be found from (2.19) with ζ replaced by r. It only remains to calculate the constant j. To this end we note that in the vicinity of the source $w \approx w(0,\widetilde{\rho}(0)) \equiv w_0$, where $\widetilde{\rho}(0)$ is the (known) intensity of the source, and the solution of (2.26) has the form

$$\psi(r) = \frac{\sqrt{\widetilde{\rho}(0)}}{r}e^{i\sqrt{w_0}r}. \tag{2.30}$$

Employing (2.19) we obtain

$$j = \widetilde{\rho}(0)\sqrt{w_0}. \tag{2.31}$$

3. Conclusions. A geometrical optics approximation has been developed for nonlinear equations of a rather general form. By the use of this method, solutions of a nonlinear Hamilton–Jacobi equation have been found, and the propagation of plane and spherical waves in nonlinear media has been studied.

REFERENCES

[1] A. B. Shvartsburg, *The self-action of the electromagnetic field in a medium with considerable nonlinearity*, Opt. and Quantum El., 8 (1976), pp. 393-398.
[2] A. B. Shvartsburg, *the nonstationary evolution of localized wave fields in nonlinear dispersive media*, in Nonlinear Electromagnetics, P. L. E. Uslenghi, ed., Academic Press, New York, 1980, pp. 133–187.
[3] I. Dajani, E. C. Morse, and R. W. Ziolkowski, *Weakly nonlinear geometrical optics in plasmas*, Phys. D, 64 (1993), pp. 237-250.
[4] S. C. Yap, B. C. Quek, and K. S. Low, *General eikonal approximation. 2. Propagation of stationary electromagnetic waves in linear and nonlinear media*, J. Opt. Soc. Amer. A, 15, (1998), pp. 2725–2729.
[5] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Butterworth, Oxford, 1995.
[6] B. L. Rozdestvenskii and N. N. Ivanenko, *Systems of Quasilinear Equations and Their Applications to Gas Dynamics*, AMS, Providence, RI, 1983.
[7] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 8, *Electrodynamics of Continuous Media*, Pergamon Press, Oxford, 1984.
[8] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley Interscience, New York, 1974.
[9] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 2, *The Classical Theory of Fields*, Pergamon Press, Oxford, 1975.
[10] N. Froman and P. O. Froman, *JWKB Approximation, Contributions to the Theory*, North–Holland, Amsterdam, 1965.
[11] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 1, *Mechanics*, Pergamon Press, Oxford, 1976.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1133–1155

STATISTICAL STABILITY IN TIME REVERSAL*

GEORGE PAPANICOLAOU†, LEONID RYZHIK‡, AND KNUT SØLNA§

Abstract. When a signal is emitted from a source, recorded by an array of transducers, time-reversed, and re-emitted into the medium, it will refocus approximately on the source location. We analyze the refocusing resolution in a high frequency remote-sensing regime and show that, because of multiple scattering in an inhomogeneous or random medium, it can improve beyond the diffraction limit. We also show that the back-propagated signal from a spatially localized narrow-band source is self-averaging, or statistically stable, and relate this to the self-averaging properties of functionals of the Wigner distribution in phase space. Time reversal from spatially distributed sources is self-averaging only for broad-band signals. The array of transducers operates in a remote-sensing regime, so we analyze time reversal with the parabolic or paraxial wave equation.

Key words. wave propagation, random medium, Liouville–Ito equation, stochastic flow, time reversal

AMS subject classifications. 35L05, 60H15, 35Q60

DOI. 10.1137/S0036139902411107

1. Introduction.

In time reversal experiments, a signal emitted by a localized source is recorded by an array and then re-emitted into the medium time-reversed, that is, the tail of the recorded signal is sent back first. In the absence of absorption, the re-emitted signal propagates back toward the source and focuses approximately on it. This phenomenon has numerous applications in medicine, underwater acoustics, and elsewhere and has been extensively studied in the literature, both from the experimental and theoretical points of view [12, 13, 14, 15, 16, 20, 24, 25, 31]. Recently time reversal has been also the subject of active mathematical research in the context of wave propagation and imaging in random media [2, 3, 4, 5, 7, 8, 9, 32]. A schematic description of a time reversal experiment is presented in Figure 1.1.

For a point source in a homogeneous medium, the size of the refocused spot is approximately $\lambda L/a$, where λ is the central wavelength of the emitted signal, L is the distance between the source and the transducer array, and a is the aperture of the array. We assume here that the array is operating in the remote-sensing regime $a \ll L$. Multiple scattering in a randomly inhomogeneous medium creates *multipathing*, which means that the transducer array can capture waves that were initially moving away from it but got scattered onto it by the inhomogeneities. As a result, the array captures a wider aperture of rays emanating from the original source and appears to be larger than its physical size. Therefore, somewhat contrary to intuition, the inhomogeneities of the medium do not destroy the refocusing but enhance its resolution. The refocused spot is now $\lambda L/a_e$, where $a_e > a$ is the *effective* size of the

*Received by the editors July 10, 2002; accepted for publication (in revised form) July 3, 2003; published electronically April 21, 2004.
http://www.siam.org/journals/siap/64-4/41110.html

†Department of Mathematics, Stanford University, Stanford, CA 94305 (papanico@math.stanford.edu). The research of this author was supported in part by AFOSR grant F49620-01-1-0465, NSF grant DMS-9971972, and ONR grant N00014-02-1-0088.

‡Department of Mathematics, University of Chicago, Chicago, IL 60637 (ryzhik@math.uchicago.edu). The research of this author was supported in part by NSF grant DMS-9971742, an Alfred P. Sloan Fellowship, and ONR grant N00014-02-1-0089.

§Department of Mathematics, University of California, Irvine, CA 92697 (ksolna@math.uci.edu). The research of this author was supported in part by NSF grant DMS-0093992 and ONR grant N00014-02-1-0089.

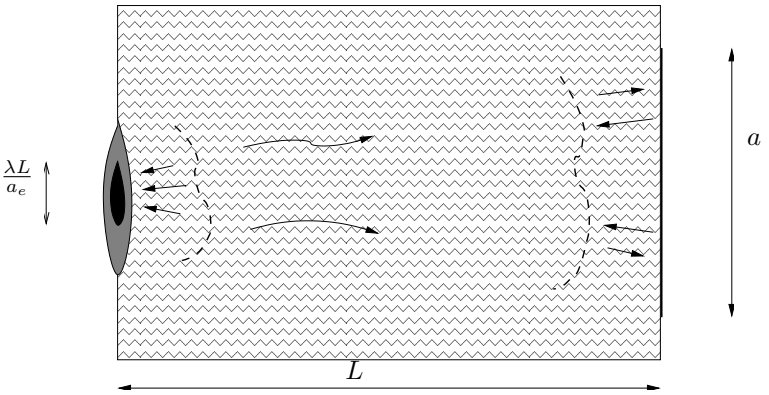


FIG. 1.1. *A pulse propagates toward a time reversal array of size a. The propagation distance L is large compared to a. The ambient medium has a randomly varying index of refraction with a typical correlation length that is small compared to L. The signal is time-reversed at the array and sent back into the medium. The back-propagated signal refocuses with spot size $\lambda L/a_e$, where a_e is the effective aperture of the array (see section* 3.3*) and $a_e > a$.*

array in the randomly scattering medium, and depends on L. The enhancement of refocusing resolution by multipathing is called *superresolution* [7]. The time-reversed pulse is also *self-averaging*, and refocusing near the source is therefore *statistically stable*, which means that it does not depend on the particular realization of the random medium. There is some loss of energy in the refocused signal because of scattering away from the array, but this can be overcome by amplification, up to a point.

The purpose of this paper is to explore in detail the mathematical basis of pulse stabilization, beyond what was done in [7]. We want to explore in particular in what regime of parameters statistical stability is observed in time reversal. We show here that for high frequency waves in a remote-sensing regime, spatially localized sources lead to statistically stable superresolution in time reversal even for narrow-band signals. We also show that, when the source is spatially distributed, only for broad-band signals do we have statistical stability in time reversal. The regime where our analysis holds is a high frequency one, more appropriate to optical or infrared time reversal than to ultrasound, sonar, or microwave radar. In this regime we can make precise what "spatially localized" or "distributed" means (see section 3.1). The numerical simulations in [7] and [8], which are set in an ultrasound or underwater sound regime, indicate that time reversal is not statistically stable for narrow-band signals even for localized sources. Only for broad-band signals is time reversal statistically stable in the regime of ultrasound experiments or sonar.

If the aperture of the transducer array is small with $a/L \ll 1$, the Fresnel number $L/(ka^2)$ is of order one, and the random inhomogeneities are weak, which is often the case, we may analyze wave propagation in the paraxial or parabolic approximation [29]. The wave field is then given approximately by

$$u(t, \mathbf{x}, z) = \frac{1}{2\pi} \int e^{i\omega(z/c_0 - t)} \psi(z, \mathbf{x}; \omega/c_0) d\omega, \tag{1.1}$$

where the complex amplitude ψ satisfies the parabolic or Schrödinger equation

$$2ik\psi_z + \Delta_{\mathbf{x}}\psi + k^2(n^2 - 1)\psi = 0. \tag{1.2}$$

Here $\mathbf{x} = (x, y)$ are the coordinates transverse to the direction of propagation z, the wave number $k = \omega/c_0$, and $n(\mathbf{x}, z) = c_0/c(\mathbf{x}, z)$ is the random index of refraction relative to a reference speed c_0. The fluctuations of the refraction index,

$$\sigma\mu\left(\frac{\mathbf{x}}{l}, \frac{z}{l}\right) = n^2(\mathbf{x}, z) - 1, \tag{1.3}$$

are assumed to be a stationary random field with mean zero, variance σ^2, correlation length l, and normalized covariance with dimensionless arguments

$$R(\mathbf{x}, z) = E\{\mu(\mathbf{x} + \mathbf{x}', z + z')\mu(\mathbf{x}', z')\}. \tag{1.4}$$

A convenient tool for the analysis of wave propagation in a random medium is the Wigner distribution [19, 28] defined by

$$W(z, \mathbf{x}, \mathbf{p}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\mathbf{p}\cdot\mathbf{y}} \psi\left(\mathbf{x} - \frac{\mathbf{y}}{2}, z\right) \overline{\psi\left(\mathbf{x} + \frac{\mathbf{y}}{2}, z\right)} d\mathbf{y}, \tag{1.5}$$

where $d = 1$ or 2 is the transverse dimension and the bar denotes complex conjugate. The Wigner distribution may be interpreted as phase space wave energy, and it is particularly well suited for high frequency asymptotics and random media [28]. The quantity of principal interest in time reversal, the time-reversed and back-propagated wave field, can also be expressed in terms of the Wigner distribution (see section 3.1). The self-averaging properties of the back-propagated field are related to the self-averaging properties of functionals of the Wigner distribution in the form of integrals of W over the wave numbers $\mathbf{p}$.

In the next section we introduce a precise scaling that corresponds to (a) high frequency, (b) long propagation distance, (c) narrow beam propagation, and (d) weak random fluctuations. In the asymptotic limit where the small parameters go to zero, the Wigner distribution satisfies a stochastic partial differential equation (SPDE), a Liouville–Ito equation, that has the form

$$dW(z, \mathbf{x}, \mathbf{p}; k) = \left(-\frac{\mathbf{p}}{k} \cdot \nabla_{\mathbf{x}} W + \frac{k^2 D}{2} \Delta_{\mathbf{p}} W\right) dz - \frac{k}{2} \nabla_{\mathbf{p}} W \cdot d\mathbf{B}(\mathbf{x}, z), \tag{1.6}$$

where $\mathbf{B}(\mathbf{x}, z)$ is a vector-valued Brownian field with covariance

$$E\{B_i(\mathbf{x}_1, z_1) B_j(\mathbf{x}_2, z_2)\} = -\left(\frac{\partial^2 R_0(\mathbf{x}_1 - \mathbf{x}_2)}{\partial x_i \partial x_j}\right) z_1 \wedge z_2, \tag{1.7}$$

where $z_1 \wedge z_2 = \min\{z_1, z_2\}$, and in the isotropic case

$$D = -\frac{R_0''(0)}{4}, \quad R_0(\mathbf{x}) = \int_{-\infty}^{\infty} R(\mathbf{x}, s) ds. \tag{1.8}$$

In section 2.5 we analyze this SPDE in the asymptotic limit of small correlation length for $\mathbf{B}(\mathbf{x}, z)$ in the transverse variables $\mathbf{x}$ and show that $W(z, \mathbf{x}, \mathbf{p}; k)$'s with different wave vectors $\mathbf{p}$ are uncorrelated. From this decorrelation property, we deduce that for localized sources the time-reversed back-propagated field is self-averaging, even for narrow-band signals. For distributed sources, it is self-averaging only for broad-band signals. We show in detail in section 3 how the asymptotic theory is used in time reversal. In Appendix A we introduce other scalings which lead to the same averaged SPDE, but we do not analyze them in detail.

Throughout the paper we define the Fourier transform by

$$\hat{f}(\mathbf{k}) = \int d\mathbf{x} e^{-i\mathbf{k}\cdot\mathbf{x}} f(\mathbf{x})$$

so that

$$f(\mathbf{x}) = \int \frac{d\mathbf{k}}{(2\pi)^d} e^{i\mathbf{k}\cdot\mathbf{x}} \hat{f}(\mathbf{k}).$$

2. Scaling and asymptotics.

2.1. The rescaled problem. To carry out the asymptotic analysis we begin by rewriting the Schrödinger equation (1.2) in dimensionless form. Let L_z and $L_{\mathbf{x}}$ be characteristic length scales in the propagation direction, as, for example, the distance L between the source and the transducer array for L_z and a multiple of the array size a for L_x. We introduce a dimensionless wave number $k' = k/k_0$ with $k_0 = \omega_0/c_0$ and ω_0 a central frequency. We rescale $\mathbf{x}$ and z by $\mathbf{x} = L_{\mathbf{x}}\mathbf{x}'$, $z = L_z z'$ and rewrite (1.2) in the new coordinates, dropping primes:

$$2ik\frac{\partial\psi}{\partial z} + \frac{L_z}{k_0 L_{\mathbf{x}}^2}\Delta\psi + k^2 k_0 L_z \sigma\mu\left(\frac{\mathbf{x}L_{\mathbf{x}}}{l}, \frac{zL_z}{l}\right)\psi = 0. \tag{2.1}$$

The physical parameters that characterize the propagation problem are (a) the central wave number k_0, (b) the strength of the fluctuations σ, and (c) the correlation length l. We now introduce three dimensionless variables

$$\delta = \frac{l}{L_{\mathbf{x}}}, \quad \varepsilon = \frac{l}{L_z}, \quad \gamma = \frac{1}{k_0 l}, \tag{2.2}$$

which are the reciprocals of the *transverse scale* relative to correlation length, the reciprocal of the *propagation distance* relative to correlation length, and the central *wave length* relative to the correlation length. We will assume that the dimensionless parameters γ, σ, ε, and δ are small:

$$\gamma \ll 1; \quad \sigma \ll 1; \quad \delta \ll 1; \quad \varepsilon \ll 1. \tag{2.3}$$

This is a regime of parameters where superresolution phenomena can be observed.

To make the scaling more precise we introduce the Fresnel number

$$\theta = \frac{L_z}{k_0 L_{\mathbf{x}}^2} = \gamma\frac{\delta^2}{\varepsilon}. \tag{2.4}$$

We can then rewrite the Schrödinger equation (2.1) in the form

$$2ik\theta\psi_z + \theta^2\Delta_{\mathbf{x}}\psi + \frac{k^2\delta}{\varepsilon^{1/2}}\mu\left(\frac{\mathbf{x}}{\delta}, \frac{z}{\varepsilon}\right)\psi = 0, \tag{2.5}$$

provided that we relate ε to σ and δ by

$$\varepsilon = \sigma^{2/3}\delta^{2/3} \tag{2.6}$$

so that $\varepsilon \to 0$ is a white noise limit. One way that the asymptotic regime (2.3) can be realized is with the ordering

$$\theta \ll \varepsilon \ll \delta \ll 1\ , \tag{2.7}$$

and $\gamma \ll \sigma^{4/3}\delta^{-2/3}$, corresponding to the high frequency limit. We see from the scaled Schrödinger equation (2.5) that this regime can be given the following interpretation. We have first a *high frequency* limit $\theta \to 0$, then a *white noise* limit $\varepsilon \to 0$, and then a *broad beam* limit $\delta \to 0$. We will analyze in detail and interpret these limits in the following sections. Another scaling in which (2.3) is realized is $\varepsilon \ll \theta \ll \delta \ll 1$. This is a regime in which the white noise limit is carried out first, then the high frequency limit, and then the broad beam limit. We do not analyze this case here. Additional comments on scaling are provided in Appendix A.

It is instructive to express the constraints (2.6) and (2.7) in terms of the dimensional parameters of the problem. First, both the size of the transverse scale $L_{\mathbf{x}}$ and the propagation distance L_z should be much larger than the correlation length l of the medium. Moreover, (2.6) implies that the longitudinal and transverse scales should be related by

$$\frac{L_z}{L_{\mathbf{x}}} = \frac{\delta}{\varepsilon} = \left(\frac{\delta}{\sigma^2}\right)^{1/3} \gg 1$$

so that we are indeed in the beam approximation. The first inequality in (2.7) implies that

$$\frac{L_z}{L_{\mathbf{x}}} \ll \sqrt{k_0 l} = \frac{1}{\sqrt{\gamma}},$$

and, with the above choice of L_z, this implies that

$$\frac{\gamma^{3/2}}{\sigma^2} \ll \frac{L_{\mathbf{x}}}{l} \ll \frac{1}{\sigma^2}.$$

2.2. The high frequency limit. A convenient tool for the study of the high frequency limit, especially in random media, is the Wigner distribution. It is often used in the context of energy propagation [19, 28], but it is also useful in analyzing time reversal phenomena [2, 3, 5, 7]. Let $\phi_\theta(\mathbf{x})$ be a family of functions oscillating on a small scale θ. The Wigner distribution is a function of the physical space coordinate $\mathbf{x}$ and wave vector $\mathbf{p}$ defined as

$$W_\theta(\mathbf{x},\mathbf{p}) = \int_{\mathbb{R}^d} \frac{d\mathbf{y}}{(2\pi)^d} e^{i\mathbf{p}\cdot\mathbf{y}} \phi_\theta\left(\mathbf{x} - \frac{\theta\mathbf{y}}{2}\right) \overline{\phi_\theta\left(\mathbf{x} + \frac{\theta\mathbf{y}}{2}\right)}. \tag{2.8}$$

The family W_θ is bounded in the space of Schwartz distributions $\mathcal{S}'(\mathbb{R}^d \times \mathbb{R}^d)$ if the functions ϕ_θ are uniformly bounded in $L^2(\mathbb{R}^d)$. Therefore, there exists a subsequence $\theta_k \to 0$ such that W_{θ_k} converges weakly as $k \to \infty$ to a limit measure $W(\mathbf{x},\mathbf{p})$. This limit $W(\mathbf{x},\mathbf{p})$ is nonnegative and is customarily interpreted as the limit phase space energy density because

$$|\phi_{\theta_k}(\mathbf{x})|^2 \to \int_{\mathbb{R}^d} W(\mathbf{x},\mathbf{p}) d\mathbf{p} \quad \text{as } k \to 0 \tag{2.9}$$

in the weak sense. This allows us to think of $W(\mathbf{x},\mathbf{p})$ as a local energy density.

Let $W_\theta(z,\mathbf{x},\mathbf{p})$ be the Wigner distribution of the solution ψ of the Schrödinger equation (2.5), in the transversal space-variable $\mathbf{x}$. A straightforward calculation

shows that $W_\theta(z,\mathbf{x},\mathbf{p})$ satisfies in a weak sense the linear evolution equation

$$\begin{aligned}\frac{\partial W_\theta}{\partial z}+\frac{\mathbf{p}}{k}\cdot\nabla_{\mathbf{x}}W_\theta \\ =\frac{ik\delta}{2\sqrt{\varepsilon}}\int e^{i\mathbf{q}\cdot\mathbf{x}/\delta}\hat{\mu}\left(q,\frac{z}{\varepsilon}\right)\frac{W_\theta\left(\mathbf{p}-\frac{\theta\mathbf{q}}{2\delta}\right)-W_\theta\left(\mathbf{p}+\frac{\theta\mathbf{q}}{2\delta}\right)}{\theta}\frac{d\mathbf{q}}{(2\pi)^d}.\end{aligned} \tag{2.10}$$

In the limit $\theta\to 0$, the solution converges weakly in $\mathcal{S}'$, for each realization, to the (weak) solution of the random Liouville equation

$$\frac{\partial W}{\partial z}+\frac{\mathbf{p}}{k}\cdot\nabla_{\mathbf{x}}W+\frac{k}{2\sqrt{\varepsilon}}\nabla_{\mathbf{x}}\mu\left(\frac{\mathbf{x}}{\delta},\frac{z}{\varepsilon}\right)\cdot\nabla_{\mathbf{p}}W=0. \tag{2.11}$$

The initial condition at $z=0$ is $W(0,\mathbf{x},\mathbf{p})=W_I(\mathbf{x},\mathbf{p})$, the limit Wigner distribution of the initial wave function.

2.3. The white noise limit. In this section we take the white noise limit $\varepsilon\to 0$ in the random Liouville equation (2.11) whose solution we now denote by W_ε. We can do this using the asymptotic theory of stochastic differential equations and flows [22, 6, 21, 26] as follows. Using the method of characteristics, the solution of the Liouville equation (2.11) may be written in the form

$$W_\varepsilon(z,\mathbf{x},\mathbf{p})=W_I(\mathbf{X}_\varepsilon(z;\mathbf{x},\mathbf{p}),\mathbf{P}_\varepsilon(z;\mathbf{x},\mathbf{p})),$$

where the processes $\mathbf{X}_\varepsilon(z;\mathbf{x},\mathbf{p})$ and $\mathbf{P}_\varepsilon(z;\mathbf{x},\mathbf{p})$ are solutions of the characteristic equations

$$\frac{d\mathbf{X}_\varepsilon}{dz}=-\frac{1}{k}\mathbf{P}_\varepsilon,\quad \frac{d\mathbf{P}_\varepsilon}{dz}=-\frac{k}{2\sqrt{\varepsilon}}\nabla_{\mathbf{x}}\mu\left(\frac{\mathbf{X}_\varepsilon}{\delta},\frac{z}{\varepsilon}\right)$$

with the initial conditions $\mathbf{X}_\varepsilon(0)=\mathbf{x}$ and $\mathbf{P}_\varepsilon(0)=\mathbf{p}$. We assume here that the fluctuation process $\mu(\mathbf{x},z)$ is twice differentiable. The asymptotic theory of random differential equations with rapidly oscillating coefficients implies that, under suitable conditions on μ, in the limit $\varepsilon\to 0$, the processes $\mathbf{X}_\varepsilon$, $\mathbf{P}_\varepsilon$ converge weakly (in the probabilistic sense) and uniformly on compact sets in $\mathbf{x},\mathbf{p}$ to the limit processes $\mathbf{X}(z)$, $\mathbf{P}(z)$ that satisfy a system of stochastic differential equations

$$d\mathbf{P}=-\frac{k}{2}d\mathbf{B}(z),\quad d\mathbf{X}=-\frac{1}{k}\mathbf{P}dz,\quad \mathbf{X}(0)=\mathbf{x},\quad \mathbf{P}(0)=\mathbf{p}.$$

The random process $\mathbf{B}(z)$ is a Brownian motion with the covariance function

$$\begin{aligned}E\left\{B_i(z_1)B_j(z_2)\right\}&=-\frac{\partial^2R_0(0)}{\partial x_i\partial x_j}dsz_1\wedge z_2\\ &=\delta_{ij}\left(-R_0''(0)\right)z_1\wedge z_2,\end{aligned} \tag{2.12}$$

in the isotropic case, where

$$R_0(\mathbf{x})=\int_{-\infty}^{\infty}R(\mathbf{x},s)ds \tag{2.13}$$

is a function of $|\mathbf{x}|$. This implies that the average Wigner distribution $W_\varepsilon^{(1)}(z,\mathbf{x},\mathbf{p}) = E\{W_\varepsilon(z,\mathbf{x},\mathbf{p})\}$ converges as $\varepsilon \to 0$ uniformly on compact sets to the solution of the advection-diffusion equation in phase space

$$\frac{\partial W^{(1)}}{\partial z} + \frac{\mathbf{p}}{k}\cdot\nabla_{\mathbf{x}} W^{(1)} = \frac{k^2 D}{2}\Delta_{\mathbf{p}} W^{(1)} \tag{2.14}$$

with the initial data $W^{(1)}(0,\mathbf{x},\mathbf{p}) = W_I(\mathbf{x},\mathbf{p})$. Here the diffusion coefficient D is given by

$$D = -\frac{R_0''(0)}{4}. \tag{2.15}$$

The one-point moments $E\left\{[W_\varepsilon(z,\mathbf{x},\mathbf{p})]^N\right\}$ converge as $\varepsilon \to 0$ to the functions $W^{(N)}(z,\mathbf{x},\mathbf{p})$ that satisfy the same equation (2.14) but with the initial data $W^{(N)}(0,\mathbf{x},\mathbf{p}) = [W_0(\mathbf{x},\mathbf{p})]^N$. This is similar to the spot dancing phenomenon [11], where all one-point moments are governed by the same Brownian motion. In particular we have that

$$W^{(2)}(z,\mathbf{x},\mathbf{p}) \neq \left[W^{(1)}(z,\mathbf{x},\mathbf{p})\right]^2$$

so that the process W_ε does not converge to a deterministic one, in the strong sense pointwise.

2.4. Multipoint moment equations. As in the previous section, we may also study the white noise limit $\varepsilon \to 0$ of the higher moments of $W_\varepsilon(z,\mathbf{x},\mathbf{p})$ at different points

$$W_\varepsilon^{(N)}(z,\mathbf{x}^1,\dots,\mathbf{x}^N,\mathbf{p}^1,\dots,\mathbf{p}^N) = E\left\{[W_\varepsilon(z,\mathbf{x}^1,\mathbf{p}^1)]^{r_1}\cdot\ldots\cdot[W_\varepsilon(z,\mathbf{x}^N,\mathbf{p}^N)]^{r_N}\right\}.$$

Here the points $(\mathbf{x}^m,\mathbf{p}^m)$ are all distinct, and $(\mathbf{x}^n,\mathbf{p}^n) \neq (\mathbf{x}^m,\mathbf{p}^m)$. We may account for moments that have different powers of W_ε at different points by taking different powers r_j of $W_\varepsilon(\mathbf{x}^j,\mathbf{p}^j)$.

We now consider the joint process $(\mathbf{X}_\varepsilon(z;\mathbf{x}^m,\mathbf{p}^m),\mathbf{P}_\varepsilon(z;\mathbf{x}^m,\mathbf{p}^m))$, $m = 1,\dots,N$. As $\varepsilon \to 0$, it converges to the solution of the system of stochastic differential equations

$$d\mathbf{P}_i^m = -\frac{k}{2}\sum_{n=1}^{N}\sum_{j=1}^{d}\sigma_{ij}\left(\frac{\mathbf{X}^m-\mathbf{X}^n}{\delta}\right)dB_j^n(z), \quad d\mathbf{X}^m = -\frac{1}{k}\mathbf{P}^m dz, \tag{2.16}$$

with the initial conditions

$$\mathbf{X}^m(0) = \mathbf{x}^m, \quad \mathbf{P}^m(0) = \mathbf{p}^m.$$

The d-dimensional Brownian motions $\mathbf{B}^m$, $m = 1,\dots,N$, have the standard covariance tensor

$$E\left\{B_i^m(z_1)B_j^n(z_2)\right\} = \delta_{mn}\delta_{ij}z_1\wedge z_2, \quad i,j = 1,\dots,d,\ m,n = 1,\dots,N.$$

The symmetric tensor $\sigma_{ij}(\mathbf{x})$ is determined from

$$\sum_{k=1}^{d}\sigma_{ik}(\mathbf{x})\sigma_{jk}(\mathbf{x}) = -\left(\frac{\partial^2 R_0(\mathbf{x})}{\partial x_i \partial x_j}\right). \tag{2.17}$$

We assume that (2.17) has a solution that is differentiable in $\mathbf{x}$, which is compatible with the fact that the matrix on the right is, by Bochner's theorem, nonnegative definite.

The moments $W_\varepsilon^{(N)}$ converge as $\varepsilon \to 0$ to the solution of the advection-diffusion equation

$$\frac{\partial W^{(N)}}{\partial z} + \sum_{m=1}^{N} \frac{\mathbf{p}^m}{k} \cdot \nabla_{\mathbf{x}^m} W^{(N)} = \frac{k^2 D}{2} \sum_{m=1}^{N} \Delta_{\mathbf{p}^m} W^{(N)} - \frac{k^2}{4} \sum_{\substack{n,m=1 \\ n>m}}^{N} \sum_{i,j=1}^{d} \frac{\partial^2 R_0((\mathbf{x}^n - \mathbf{x}^m)/\delta)}{\partial x_i \partial x_j} \frac{\partial^2 W^{(N)}}{\partial p_i^n \partial p_j^m} \tag{2.18}$$

with the initial data

$$W^{(N)}(0, \mathbf{x}_1, \ldots, \mathbf{x}^N, \mathbf{p}^1, \ldots, \mathbf{p}^N) = [W_I(\mathbf{x}^1, \mathbf{p}^1)]^{r_1} \cdot \ldots \cdot [W_I(\mathbf{x}^N, \mathbf{p}^N)]^{r_N}.$$

From (2.18) we can calculate moments of functionals of W_ε of the form

$$W_{\varepsilon,\phi}(z) = \int W_\varepsilon(z, \mathbf{x}, \mathbf{p}) \phi(\mathbf{x}, \mathbf{p}) d\mathbf{x} d\mathbf{p}.$$

For example, as $\varepsilon \to 0$, we have that

$$E\left\{[W_{\varepsilon,\phi}(z)]^2\right\} \to \int W^{(2)}(z, \mathbf{x}_1, \mathbf{p}_1, \mathbf{x}_2, \mathbf{p}_2) \phi(\mathbf{x}_1, \mathbf{p}_1) \phi(\mathbf{x}_2, \mathbf{p}_2) d\mathbf{x}_1 d\mathbf{p}_1 d\mathbf{x}_2 d\mathbf{p}_2.$$

A convenient way to deal not only with the limit of N-point moments but also with the full limit process $W(z, \mathbf{x}, \mathbf{p})$, at all points $\mathbf{x}, \mathbf{p}$ simultaneously, is provided by the theory of stochastic flows [23]. For this we need to show that $W_\varepsilon(z, \mathbf{x}, \mathbf{p})$ converges weakly (in the probabilistic sense) as $\varepsilon \to 0$ to the process $W(z, \mathbf{x}, \mathbf{p})$ that satisfies the SPDE

$$dW_\delta = \left[-\frac{\mathbf{p}}{k} \cdot \nabla_{\mathbf{x}} W_\delta + \frac{k^2 D}{2} \Delta_{\mathbf{p}} W_\delta \right] dz - \frac{k}{2} \nabla_{\mathbf{p}} W_\delta \cdot d\mathbf{B}\left(\frac{\mathbf{x}}{\delta}, z\right). \tag{2.19}$$

Here the Gaussian random field $\mathbf{B}(\mathbf{x}, z)$ has covariance

$$E\{B_i(\mathbf{x}_1, z_1) B_j(\mathbf{x}_2, z_2)\} = -\left(\frac{\partial^2 R_0(\mathbf{x}_1 - \mathbf{x}_2)}{\partial x_i \partial x_j} \right) z_1 \wedge z_2.$$

We call (2.19) the Liouville–Ito equation. It allows us to treat all equations of the form (2.18) simultaneously, and it is a convenient tool for simulation and analysis. The dimensionless wave number k can be scaled out of (2.19) by writing $W(z, \mathbf{x}, \mathbf{p}; k) = W(z, \mathbf{x}, \frac{\mathbf{p}}{k}; 1)$ so that we need only consider (2.19) with $k = 1$. We will use this scaling in section 3.1.

Note that unlike the single Brownian motion (2.12) that governs the evolution of one-point moments, the Brownian field that enters the SPDE (2.19) depends explicitly on the dimensionless correlation length δ in the transverse direction. Therefore, the limit process also depends on δ, and we denote it by W_δ.

2.5. Statistical stability in the broad beam limit. We will now consider the limit $\delta \to 0$ of the process $W_\delta(z, \mathbf{x}, \mathbf{p})$ when the transverse dimension $d \geq 2$. We are particularly interested in the behavior of functionals of W_δ as $\delta \to 0$. The analysis of one-point moments in section 2.3 showed that they do not depend on δ and are governed by a standard Brownian motion. Therefore the process W_δ does not have a pointwise deterministic limit. However, we will show that functionals of W_δ become deterministic in the limit $\delta \to 0$. We refer to this phenomenon as *statistical stabilization* and give conditions for it to happen. Stabilization plays an important role in time reversal, imaging, and other applications, as discussed in the introduction.

THEOREM 2.1. *Assume that $\phi(\mathbf{p})$ is a smooth test function of rapid decay, the transverse correlation function $R_0(\mathbf{x})$ has compact support, the initial Wigner distribution $W_I(\mathbf{x}, \mathbf{p})$ is uniformly bounded and Lipschitz continuous, and the transverse dimension $d \geq 2$. Define*

$$I_{\delta,\phi}(z, \mathbf{x}) = \int W_\delta(z, \mathbf{x}, \mathbf{p}) \phi(\mathbf{p}) d\mathbf{p}. \tag{2.20}$$

Then

$$\lim_{\delta \to 0} E\left\{I^2_{\delta,\phi}(z, \mathbf{x})\right\} = E^2\left\{I_{\delta,\phi}(z, \mathbf{x})\right\}, \tag{2.21}$$

where $E\{I_{\delta,\phi}(z, \mathbf{x})\}$ is independent of δ.

The independence of δ for the expectation of $I_{\delta,\phi}(z, \mathbf{x})$ follows immediately from taking expectations in the stochastic differential equation (2.19). The assumption of compact support for $R_0(\mathbf{x})$ is not essential but simplifies the proof. We have already noted that the Wigner distribution W_δ itself does not stabilize. However, (2.21) implies that

$$\lim_{\delta \to 0} \mathrm{Var}\left\{I_{\delta,\phi}\right\} = \lim_{\delta \to 0} E\left\{I^2_{\delta,\phi}(z)\right\} - E^2\left\{I_{\delta,\phi}\right\} = 0. \tag{2.22}$$

Therefore, any smooth functional of the form (2.20) stabilizes in the limit $\delta \to 0$; that is,

$$I_{\delta,\phi} \approx E\{I_{\delta,\phi}\} \tag{2.23}$$

in mean square, and the expectation of $I_{\delta,\phi}$ does not depend on δ. We prove Theorem 2.1 in Appendix B.

In the applications of the asymptotic theory to time reversal, we need functionals $I_{\delta,\phi}$ not only of the form (2.20), but also of the form

$$J_\delta(z, \mathbf{x}) = \int W_\delta(z, \mathbf{x}, \mathbf{p}) d\mathbf{p}. \tag{2.24}$$

We need to show that such functionals are well defined with probability one and to analyze their behavior as $\delta \to 0$. This is done in the following theorem.

THEOREM 2.2. *Under the same hypotheses of Theorem* 2.1, *and with a nonnegative initial Wigner distribution $W_I \geq 0$, the functional J_δ is bounded, continuous, and nonnegative with probability one. In the limit $\delta \to 0$, we have*

$$\lim_{\delta \to 0} E\left\{J^2_\delta(z, \mathbf{x})\right\} = E^2\left\{J_\delta(z, \mathbf{x})\right\}, \tag{2.25}$$

where $E\{J_\delta(z, \mathbf{x})\}$ does not depend on δ.

The proof of this theorem is given in Appendix B.

What is important in both Theorems 2.1 and 2.2 is that we do integrate over the wave numbers $\mathbf{p}$ because there is no pointwise stabilization. In time reversal applications, as in section 3.1, we actually need Theorem 2.2 when the integration is only over a line segment in $\mathbf{p}$ space, and the dimension of the latter is $d \geq 2$. Its proof follows from the one of Theorem 2.2.

3. Application to time reversal in a random medium. We will now apply these results to the time reversal problem [7] described in the introduction. A wave emitted from the plane $z = 0$ propagates through the random medium and is recorded on the time reversal mirror at L. It is then *reversed* in time and re-emitted into the medium. The back-propagated signal refocuses approximately at the source, as shown in Figure 1.1. There are two striking features of this refocusing in random media. One is that it is statistically stable; that is, it does not depend on the particular realization. The other is superresolution; that is, the refocused spot is tighter than in the deterministic case. We discuss these two issues in this section.

3.1. The time-reversed and back-propagated field. We assume that the wave source at $z = 0$ is distributed on a scale σ_s around a point $\mathbf{x}_0$; that is,

$$\psi_\theta(z=0,\mathbf{x};k) = e^{i\mathbf{p}_0\cdot(\mathbf{x}-\mathbf{x}_0)/\theta}\psi_0\left(\frac{\mathbf{x}-\mathbf{x}_0}{\sigma_s};k\right),$$

where ψ_0 is a rapidly decaying and smooth function of $\mathbf{x}$ and k. The width of the source σ_s could be large or small compared to the Fresnel number θ, and this affects the statistical stability of the time-reversed back-propagated field, as we explain in this section. The Green's function, $G_\theta(z,\mathbf{x},\xi;k)$, solves the parabolic wave equation (2.5) with a point source at $(\mathbf{x},z) = (\xi,0)$. Using its symmetry properties and the fact that time reversal $t \to -t$ is equivalent to $\omega \to -\omega$ or $k \to -k$, the back-propagated time-reversed field on the plane of the source has the form

$$\begin{aligned}(3.1)\quad &\psi_\theta^B(L,\mathbf{x}_0,\xi;k)\\ &= \int\!\!\int G_\theta(L,\mathbf{x},\mathbf{x}_0+\theta\xi;k)\overline{G_\theta(L,\mathbf{x}_0+\eta,\mathbf{x};k)}e^{-i\mathbf{p}_0\cdot\eta/\theta}\psi_0\left(\frac{\eta}{\sigma_s};-k\right)\chi_A(\mathbf{x})d\mathbf{x}d\eta.\end{aligned}$$

The complex field amplitude ψ_θ^B is evaluated at $\mathbf{x}_0+\theta\xi$, in the plane $z = 0$. We scale the observation point off $\mathbf{x}_0$ by θ because we expect that the spot size of the refocused signal will be comparable to the lateral spread of the initial wave function. We denote by χ_A the aperture function of the time reversal mirror. It could be its characteristic function, occupying the region A in the plane $z = L$,

$$\chi_A(\mathbf{x}) = \begin{cases} 1, & \mathbf{x}\in A,\\ 0, & \mathbf{x}\notin A,\end{cases}$$

or a more general aperture function like a Gaussian. The time reversal mirror is located in the plane $z = L$.

After changing variables, the back-propagated field is given by

$$\begin{aligned}&\psi_\theta^B(L,\mathbf{x}_0,\xi;k)\\ &= \theta^d\int G_\theta(L,\mathbf{x},\mathbf{x}_0+\theta\xi;k)\overline{G_\theta(L,\mathbf{x},\mathbf{x}_0+\theta\eta;k)}e^{-i\mathbf{p}_0\cdot\eta}\psi_0\left(\frac{\theta\eta}{\sigma_s};-k\right)\chi_A(\mathbf{x})d\mathbf{x}d\eta\\ &= \theta^d\int G_\theta(L,\mathbf{x}_0+\theta\xi,\mathbf{x};k)\overline{G_\theta(L,\mathbf{x}_0+\theta\eta,\mathbf{x};k)}e^{-i\mathbf{p}_0\cdot\eta}\psi_0\left(\frac{\theta\eta}{\sigma_s};-k\right)\chi_A(\mathbf{x})d\mathbf{x}d\eta.\end{aligned}$$

It is now convenient to introduce the Wigner distribution

(3.2)
$$W_\theta(z,\mathbf{x}_0,\mathbf{p};k)=\int \frac{\theta^d e^{i\mathbf{p}\cdot\mathbf{y}}}{(2\pi)^d} G_\theta(z,\mathbf{x}_0-\mathbf{y}\theta/2,\mathbf{x};k)\overline{G_\theta(z,\mathbf{x}_0+\mathbf{y}\theta/2,\mathbf{x};k)}\chi_A(\mathbf{x})d\mathbf{x}d\mathbf{y},$$

and express the back-propagated field as

(3.3)
$$\begin{aligned}&\psi_\theta^B(L,\mathbf{x}_0,\xi;k)\\ &\quad=\int e^{i\mathbf{p}\cdot(\xi-\eta)}W_\theta\left(L,\mathbf{x}_0+\frac{\theta(\xi+\eta)}{2},\mathbf{p};k\right)e^{-i\mathbf{p}_0\cdot\eta}\psi_0\left(\frac{\theta\eta}{\sigma_s};-k\right)d\mathbf{p}d\eta.\end{aligned}$$

The Wigner distribution is scaled differently here from in (2.8) because of the way we have scaled the source function.

In the high frequency limit $\theta\to 0$, $W_\theta(z,\mathbf{x},\mathbf{p};k)$ tends to $W(z,\mathbf{x},\mathbf{p};k)$, which solves the random Liouville equation (2.11). Then, in the white noise limit, it solves the Liouville–Ito equation (2.19). The mean of W solves (2.14), in the high frequency and white noise limit, with initial data

(3.4)
$$W(0,\mathbf{x},\mathbf{p};k)=\frac{\chi_A(\mathbf{x})}{(2\pi)^d}.$$

Let

(3.5)
$$\beta=\frac{\sigma_s}{\theta}$$

be the ratio of the width of the source to the Fresnel number and assume that it remains fixed as $\theta\to 0$. In this limit, the time-reversed and back-propagated field is given by

(3.6)
$$\begin{aligned}\psi^B(L,\mathbf{x}_0,\xi;k)&=\int e^{i\mathbf{p}\cdot(\xi-\eta)}W\left(L,\mathbf{x}_0,\frac{\mathbf{p}}{k}\right)e^{-i\mathbf{p}_0\cdot\eta}\psi_0(\eta/\beta;-k)d\mathbf{p}d\eta\\ &=\int e^{i\mathbf{p}\cdot\xi}W\left(L,\mathbf{x}_0,\frac{\mathbf{p}}{k}\right)\beta^d\hat{\psi}_0(\beta(\mathbf{p}+\mathbf{p}_0);-k)d\mathbf{p}.\end{aligned}$$

Here we have used the scaling $W(z,\mathbf{x},\mathbf{p};k)=W(z,\mathbf{x},\frac{\mathbf{p}}{k};1)$ in (2.19), and we have dropped the last argument $k=1$.

3.2. Statistical stability. From the form (3.6) of the back-propagated and time-reversed field we see that when $\beta=O(1)$ (or small), which means that σ_s is comparable to the Fresnel number θ (or smaller), we can apply the results of section 2.5 and conclude that it is statistically stable or self-averaging in the broad beam limit $\delta\to 0$. Theorems 2.1 and 2.2 are exactly what is needed for this. The fact that the initial function (3.4) may be discontinuous at the boundary of the set A is not a problem. This is because we may approximate the function χ_A from above and below by two smooth positive functions, to which we may apply Theorems 2.1 and 2.2, and then use the maximum principle to deduce the decorrelation property when the initial data is χ_A. We have, therefore,

$$\psi^B(L,\mathbf{x}_0,\xi;k)\approx\langle\psi^B(L,\mathbf{x}_0,\xi;k)\rangle$$

in the sense of convergence in probability or in mean square, in the broad beam limit $\delta\to 0$, for each fixed frequency $\omega=kc_0$. Statistical stability of time reversal does

not depend on having a broad-band signal if the source is localized in space. This is true in the regime of parameters reflected by the scaling $\theta \ll \varepsilon \ll \delta$ considered here, which is a high frequency regime encountered in optical or infrared applications like ladar. The numerical experiments in [7] and [8] are closer to the regime of ultrasound experiments [16] and in underwater sound propagation, which is different from the high frequency regime analyzed here.

For distributed sources, the parameter β is large, and we cannot apply Theorems 2.1 and 2.2 to (3.6). It is necessary for statistical stability in this case to have broad-band signals. For β large, the time-reversed and back-propagated signal in the time domain has the form

$$\begin{aligned}(3.7)\quad &\psi^B(L,\mathbf{x}_0,\xi,t)\\ &= (2\pi)^d e^{-i(\mathbf{p}_0\cdot\xi+k_0c_0t)}\psi_0(\xi/\beta)\int W\left(L,\mathbf{x}_0,\frac{\mathbf{p}_0}{k_0+k}\right)e^{-ikc_0t}\hat{g}(-c_0k)\frac{c_0dk}{2\pi}\\ &= (2\pi)^d e^{-i(\mathbf{p}_0\cdot\xi+\omega_0t)}\psi_0(\xi/\beta)\int W\left(L,\mathbf{x}_0,\frac{c_0\mathbf{p}_0}{\omega_0+\omega}\right)e^{-i\omega t}\hat{g}(-\omega)\frac{d\omega}{2\pi}\end{aligned}$$

with $\hat{g}(c_0k)$ the Fourier transform of the initial pulse relative to the central frequency $\omega_0 = c_0k_0$. This means that we have replaced the actual wave number k by k_0+k, or ω by $\omega_0+\omega$, with the new ω, the baseband frequency, bounded by Ω, the bandwidth, $|\omega| \le \Omega < \omega_0$. The integration is over the bandwidth $[-\Omega,\Omega]$. This integral is well defined with probability one and is self-averaging in the broad beam limit $\delta \to 0$ by Theorem 2.2 and the remark following it. We will compute its average in section 3.4.

3.3. The effective aperture of the array. From the explicit expression for the Green's function of (2.14), with $k=1$,

$$\begin{aligned}W^{(1)}(z,\mathbf{x},\mathbf{p};\mathbf{x}^0,\mathbf{p}^0) = &\int \frac{d\mathbf{w}d\mathbf{r}}{(2\pi)^{2d}}\exp\left(i\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}^0)+i\mathbf{r}\cdot(\mathbf{p}-\mathbf{p}^0)-iz\mathbf{w}\cdot\mathbf{p}^0\right)\\ &\times\exp\left(-\frac{Dz}{2}\left[r^2+z\mathbf{r}\cdot\mathbf{w}+\frac{w^2z^2}{3}\right]\right),\end{aligned}$$

and with the time reversal mirror a distance L from the source and $\mathbf{x}_0=0$, it follows from (3.6) that

$$\begin{aligned}(3.8)\quad &\langle\psi^B(L,\xi;k)\rangle\\ &= \int\frac{d\mathbf{p}d\mathbf{y}d\mathbf{w}}{(2\pi)^{2d}}e^{i\mathbf{p}\cdot\xi}\beta^d\hat{\psi}_0(\beta(\mathbf{p}-\mathbf{p}_0);-k)\chi_A(\mathbf{y})\exp\left[-i\mathbf{w}\cdot\mathbf{y}-iL\mathbf{w}\cdot\frac{\mathbf{p}}{k}-\frac{DL^3w^2}{6}\right].\end{aligned}$$

The high frequency white noise limit of the *self-averaging* time-reversed and back-propagated field is therefore given by a convolution

$$(3.9)\qquad \langle\psi^B(L,\xi;k)\rangle = \psi_0^\beta(\cdot,-k)*\mathcal{W}(\cdot)(\xi)$$

with

$$(3.10)\qquad \mathcal{W}(\eta) = \mathcal{W}(\eta;L,k) = \frac{k^d}{(2\pi L)^d}\hat{\chi}_A(\eta k/L)\ e^{-\eta^2/(2\sigma_M^2)},$$

the *point spread function*, and

$$(3.11)\qquad \psi_0^\beta(\eta,-k) = e^{-i\mathbf{p}_0\cdot\eta}\psi_0(\eta/\beta)\hat{g}(-kc_0)$$

with $\psi_0(\eta/\beta)$ the spatial source distribution function and $\hat{g}$ the Fourier transform of the pulse shape function $g(t)$. This notation is consistent with (3.7), with the time factor $e^{-ik_0c_0t}$ omitted, along with the horizontal phase e^{ikz}, which cancels in time reversal. We have also introduced the refocused *spot size* with multipathing

$$\sigma_M^2 = \frac{3}{DLk^2} = \frac{L^2}{k^2a_e^2} \tag{3.12}$$

and the *effective aperture* $a_e = a_e(L)$,

$$a_e = \sqrt{\frac{DL^3}{3}}, \tag{3.13}$$

which we now interpret.

If the time reversal mirror is the whole plane $z = L$, then $\chi_A \equiv 1$ and

$$\left\langle \psi^B(L,\xi;k) \right\rangle = \psi_0^\beta(\xi,-k).$$

In this case, the back-propagated field is the source field reversed in time, both in the random and in the deterministic case. The point spread function $\mathcal{W}$ determines the resolution of the refocused signal for a time reversal mirror of finite aperture. Multipathing in a random medium gives rise to the Gaussian factor (3.12) whose variance is σ_M^2. We can give an interpretation of this variance, or spot size, as follows. For a square time reversal mirror of size a, the Fourier transform of χ_A is the sinc function so that

$$\mathcal{W}(\eta_1,\eta_2;L,k) = \frac{1}{\pi\eta_1}\sin\left(\frac{\eta_1 ka}{2L}\right)\frac{1}{\pi\eta_2}\sin\left(\frac{\eta_2 ka}{2L}\right)e^{-(\eta_1^2+\eta_2^2)/(2\sigma_M^2)}.$$

For a deterministic medium ($D = 0$), the Rayleigh resolution is the distance η_F to the first zero of the sine, the first Fresnel zone in either direction,

$$\eta_F = \frac{2\pi L}{ka} = \frac{\lambda L}{a}.$$

In general, if χ_A is supported by a region of size a, we may define the Fresnel resolution, or the Fresnel *spot size*, by

$$\sigma_F = \frac{L}{ka}.$$

For *weak multipathing*, we have $\sigma_M \gg \sigma_F$ and

$$\mathcal{W}(\eta;L,k) \sim \left(\frac{k}{2\pi L}\right)^d \hat{\chi}_A(\eta k/L)\ ,$$

which is the diffractive point spread function whose integral over $\eta \in R^d$ is one. If, however, we have *strong multipathing*, $\sigma_M \ll \sigma_F$, then we may approximate $\hat{\chi}_A(\eta k/L)$ by $\hat{\chi}_A(0) = a^d$ in (3.10), and the point spread function becomes

$$\mathcal{W}(\eta;L,k) \sim \left(\frac{ka}{2\pi L}\right)^d e^{-|\eta|^2/(2\sigma_M^2)}.$$

By writing the variance (spot size) σ_M^2 in the form (3.12) we can interpret a_e as an effective aperture of the time reversal mirror. We can rewrite the point spread function in terms of a normalized Gaussian as

$$\mathcal{W}(\eta; L, k) \sim \left(\frac{\sigma_M}{\sqrt{2\pi}\sigma_F}\right)^d \frac{e^{-|\eta|^2/(2\sigma_M^2)}}{(2\pi\sigma_M^2)^{d/2}}$$

with the factor in front of the normalized Gaussian also equal to

$$\left(\frac{a}{\sqrt{2\pi}a_e}\right)^d.$$

This means that, when there is strong multipathing, the integral of the point spread function over R^d is equal not to one but to this ratio, which can be much smaller than one if $a_e \gg a$. Multipathing produces a tighter point spread function, but there is also loss of energy, as of course we should expect.

A more direct interpretation for the effective aperture can be given if the time reversal mirror has a Gaussian aperture function

$$\chi_A(\eta) = e^{-|\eta|^2/(2a^2)}.$$

The point spread function $\mathcal{W}$ now has the form

$$\mathcal{W}(\eta; L, k) = \left(\frac{ka}{\sqrt{2\pi}L}\right)^d e^{-|\eta|^2/(2\sigma_g^2)},$$

with

$$\sigma_g = \frac{L}{ka_g},$$

and the effective aperture a_g given by

$$a_g = \sqrt{a^2 + \frac{DL^3}{3}} = \sqrt{a^2 + a_e^2}.$$

Clearly, $a_g \approx a_e$ when there is strong multipathing and $a_e \gg a$. Written with a normalized Gaussian, the point spread function for a Gaussian aperture has the form

$$\mathcal{W}(\eta) = \left(\frac{a}{a_g}\right)^d \frac{e^{-|\eta|^2/(2\sigma_g^2)}}{(2\pi\sigma_g^2)^{d/2}}.$$

3.4. Broad-band time reversal for distributed sources. For a distributed source, its support σ_s is large compared to the Fresnel number θ, so the ratio $\beta = \sigma_s/\theta$ is large. In this case we can compute the average of (3.7) the same way as we did in (3.8), and we find that

$$\begin{aligned}(3.14)\quad &\langle \Psi^B(L, \mathbf{x}_0, \xi, t)\rangle \\ &= (2\pi)^d e^{-i(\mathbf{p}_0\cdot\xi + k_0 c_0 t)}\psi_0(\xi/\beta)\int \left\langle W\left(L, \mathbf{x}_0, \frac{\mathbf{p}_0}{k_0+k}\right)\right\rangle e^{-ikc_0 t}\hat{g}(-c_0 k)\frac{c_0 dk}{2\pi} \\ &= e^{-i(\mathbf{p}_0\cdot\xi + k_0 c_0 t)}\psi_0(\xi/\beta)\int \frac{d\mathbf{y}d\mathbf{w}c_0 dk}{(2\pi)^{d+1}}\chi_A(\mathbf{y})e^{i(\frac{L\mathbf{w}\mathbf{p}_0}{k_0+k} - \mathbf{w}\cdot\mathbf{y} - kc_0 t)}e^{-\frac{DL^3 w^2}{6}}\hat{g}(-c_0 k).\end{aligned}$$

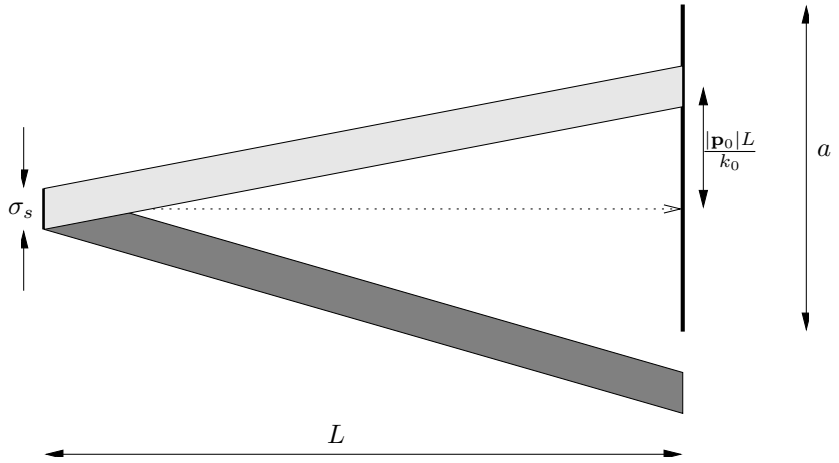


FIG. 3.1. *A directed field propagates from a distributed source of size σ_s toward the time reversal mirror of size a. The time-reversed back-propagated field depends on the location of the mirror relative to the direction of the propagating beam.*

The $\mathbf{y}$ integral on the right gives the Fourier transform of the aperture function $\chi_A(\mathbf{y})$, so with $\omega_0 = c_0 k_0$ and a change of variable from k to $\omega = c_0 k$, we have

$$\begin{aligned}(3.15)\qquad &\langle \Psi^B(L, \mathbf{x}_0, \xi, t)\rangle \\ &= e^{-i(\mathbf{p}_0\cdot\xi+\omega_0 t)}\psi_0(\xi/\beta)\int \frac{d\omega}{2\pi} e^{-i\omega t}\hat{g}(-\omega)\ \chi_A * \left(\frac{e^{-x^2/(2a_e^2)}}{(2\pi a_e^2)^{d/2}}\right)\left(\frac{Lc_0\mathbf{p}_0}{\omega_0+\omega}\right).\end{aligned}$$

Here the star denotes convolution with respect to the spatial variables $\mathbf{x}$, and a_e is the effective aperture defined by (3.13).

When multipathing is weak, we can ignore the Gaussian factor in the convolution and we have

$$\begin{aligned}(3.16)\qquad &\langle \Psi^B(L, \mathbf{x}_0, \xi, t)\rangle \\ &= e^{-i(\mathbf{p}_0\cdot\xi+\omega_0 t)}\psi_0(\xi/\beta)\int \frac{d\omega}{2\pi} e^{-i\omega t}\hat{g}(-\omega)\ \chi_A\left(\frac{Lc_0\mathbf{p}_0}{\omega_0+\omega}\right).\end{aligned}$$

In the opposite case, when there is strong multipathing and the effective aperture is much larger than the physical one, $a_e \gg a$, we have

$$\begin{aligned}(3.17)\qquad &\langle \Psi^B(L, \mathbf{x}_0, \xi, t)\rangle \\ &= e^{-i(\mathbf{p}_0\cdot\xi+\omega_0 t)}\psi_0(\xi/\beta)\left(\frac{a}{\sqrt{2\pi}a_e}\right)^d \int \frac{d\omega}{2\pi} e^{-i\omega t}\hat{g}(-\omega) e^{-\frac{1}{2}(\frac{Lc_0\mathbf{p}_0}{a_e(\omega_0+\omega)})^2}.\end{aligned}$$

To interpret these results, we note first that a distributed source function of the form (3.11) can be considered as a phased array emitting an inhomogeneous plane wave, a beam, in the direction $(k, \mathbf{p}_0)$, within the paraxial or parabolic approximation. The ratio $|\mathbf{p}_0|/k$ is the tangent of the angle the direction vector makes with the z axis, and $L|\mathbf{p}_0|/k$ is the transverse distance of the beam center to the center of the phased array (see Figure 3.1). If for each ω the beam displacement vector $Lc_0\mathbf{p}_0/(\omega_0+\omega)$

is inside the set A occupied by the time reversal array, then we recover at the source the full pulse in (3.16), time-reversed,

$$\langle \Psi^B(L, \mathbf{x}_0, \xi, t)\rangle = e^{-i(\mathbf{p}_0\cdot\xi+\omega_0 t)}\psi_0(\xi/\beta)g(-t).$$

If, however, for some frequencies the transverse displacement vector is outside the time reversal array, these frequencies will be nulled in the integration and a distorted time pulse will be received at the source. Depending on the position of the time reversal mirror relative to the beam, high or low frequencies may be nulled.

In a strongly multipathing medium, the situation is quite different because the expression (3.17), or more generally (3.15), now holds. Even if the beam from the phased array does not intercept the time reversal mirror at all, we will still get a time-reversed signal at the source but with a much diminished amplitude. If the beam falls entirely within the time reversal mirror, then the time-reversed pulse will be a distorted form of $g(-t)$, with its amplitude reduced by the factor $(a/a_e)^d$. An interesting and important application of the time reversal of a beam in a random medium is the possibility of *estimating* the effective aperture a_e by pointing the beam in different directions toward the time reversal mirror, measuring the time-reversed signal that back-propagates to the source, that is, to the phased array, and inferring a_e by fitting the measurements to (3.15).

4. Summary and conclusions. We have analyzed and explained two important phenomena associated with time reversal in a random medium:

- superresolution of the back-propagated signal due to multipathing,
- self-averaging that gives a statistically stable refocusing.

Our analysis is based on a specific asymptotic limit (see section 2.1), where the longitudinal distance of propagation is much larger than the size of the time reversal mirror, which in turn is much larger than the correlation length of the medium, fluctuations in the index of refraction are weak, and the wave length is short compared to the correlation length. This asymptotic regime is more relevant to optical or infrared time reversal than it is to sonar or ultrasound. We have related the self-averaging properties of the back-propagated signal to those of functionals of the Wigner distribution. Self-averaging of these functionals implies the statistical stability of the time-reversed and back-propagated signal in the frequency domain, provided that the source function is not too broad compared to the Fresnel number (2.4). Time reversal refocusing of waves emitted from a distributed source is self-averaging only in the time domain.

We apply our theoretical results about stochastic Wigner distributions to time reversal and discuss in detail superresolution and statistical stability in section 3.

Appendix A. The white noise limit and the parabolic approximation. We collect here some comments on the scaling analysis of section 2.1 and refer to [1, 27, 33] for additional comments and results on scaling and asymptotics in the high frequency and white noise regime.

The dimensionless parameters $\delta, \varepsilon, \gamma$ introduced by (2.2) in section 2.1, along with the Fresnel number θ defined by (2.4), lead to the scaled parabolic wave equation (2.5). If we do not make the parabolic approximation and keep the ψ_{zz} term, we have the scaled Helmholtz equation, with the phase e^{ikz} removed,

$$\frac{\varepsilon^2\theta^2}{\delta^2}\psi_{zz} + 2ik\theta\psi_z + \theta^2\Delta_{\mathbf{x}}\psi + \frac{k^2\delta}{\varepsilon^{1/2}}\mu\left(\frac{\mathbf{x}}{\delta}, \frac{z}{\varepsilon}\right)\psi = 0. \tag{A.1}$$

Here, as in (2.5), we relate the strength of the fluctuations σ to ε and δ by (2.6). Is the parabolic approximation valid in the ordering (2.7), $\theta \ll \varepsilon \ll \delta \ll 1$, that we have analyzed? The answer is yes, but not before both θ *and* ε limits have been taken, in which case the scaled Wigner distribution (2.8) converges to the Liouville–Ito process that is defined by the SPDE (2.19).

It is in the white noise limit $\varepsilon \to 0$, with Fresnel number θ and δ fixed, that the parabolic approximation is valid for (A.1), as was pointed out in [1]. This is easily seen if the random fluctuations μ are differentiable in z. The parabolic approximation is clearly not valid in the high frequency limit $\theta \to 0$, before the white noise limit $\varepsilon \to 0$ is also taken. In the white noise limit, the wave function $\psi(z, \mathbf{x})$ satisfies an Ito–Schrödinger equation

$$2ik\theta d_z\psi + \theta^2 \Delta_{\mathbf{x}} \psi dz + \frac{ik^3\delta^2}{4\theta} R_0(0)\psi dz + k^2\delta\psi d_z B\left(\frac{\mathbf{x}}{\delta}, z\right) = 0. \tag{A.2}$$

Here R_0 is the integrated covariance of the fluctuations μ given by (2.15) and (2.13), and the Brownian field $B(\mathbf{x}, z)$ has covariance

$$\langle B(\mathbf{x}, z_1) B(\mathbf{y}, z_2)\rangle = R_0(\mathbf{x} - \mathbf{y}) z_1 \wedge z_2.$$

This Ito–Schrödinger equation is the result of the central limit theorem applied to (A.1). Let

$$B^\varepsilon(\mathbf{x}, z) = \frac{1}{\sqrt{\varepsilon}} \int_0^z \mu\left(\mathbf{x}, \frac{s}{\varepsilon}\right) ds.$$

Then, as $\varepsilon \to 0$, this process converges weakly, under suitable hypotheses, to the Brownian field $B(\mathbf{x}, z)$ with the above covariance. The extra term in (A.2) is the Stratonovich correction.

The white noise limit for SPDEs is analyzed in [10] and a rigorous theory of the Ito–Schrödinger equation is given in [11]. The ergodic theory of the Ito–Schroedinger equation is explored in [17]. Wave propagation in the parabolic approximation with white noise fluctuations is considered in detail in [18, 30].

The scaled Wigner distribution for the process ψ, defined by (2.8), satisfies the stochastic transport equation

$$\begin{aligned} &d_z W_\theta(z, \mathbf{x}, \mathbf{p}) + \frac{\mathbf{p}}{k} \cdot \nabla_x W_\theta(z, \mathbf{x}, \mathbf{p}) dz \\ &= \frac{k^2\delta^2}{4\theta^2} \int \frac{d\mathbf{q}}{(2\pi)^d} \hat{R}_0(\mathbf{q}) \left(W_\theta\left(z, \mathbf{x}, \mathbf{p} + \frac{\theta\mathbf{q}}{\delta}\right) - W_\theta(z, \mathbf{x}, \mathbf{p}) \right) dz \\ &+ \frac{ik\delta}{2\theta} \int \frac{d\mathbf{q}}{(2\pi)^d} e^{i\mathbf{q}\cdot\mathbf{x}/\delta} \left(W_\theta\left(z, \mathbf{x}, \mathbf{p} - \frac{\theta\mathbf{q}}{2\delta}\right) - W_\theta\left(z, \mathbf{x}, \mathbf{p} + \frac{\theta\mathbf{q}}{2\delta}\right) \right) d_z \hat{B}(\mathbf{q}, z), \end{aligned} \tag{A.3}$$

which is derived from (A.2) using the Ito calculus. The Wigner process W_θ converges in the limit $\theta \to 0$ to the Liouville–Ito process defined by the SPDE (2.19).

Appendix B. Decorrelation of the Wigner process.

B.1. Proof of Theorem 2.1. We give here the proof of Theorems 2.1 and 2.2. We consider Theorem 2.1 first. It will follow from the Lebesgue dominated convergence theorem if we show that for $\mathbf{p}_1 \neq \mathbf{p}_2$

$$E\{W_\delta(z, \mathbf{x}, \mathbf{p}_1) W_\delta(z, \mathbf{x}, \mathbf{p}_2)\} - E\{W_\delta(z, \mathbf{x}, \mathbf{p}_1)\} E\{W_\delta(z, \mathbf{x}, \mathbf{p}_2)\} \to 0 \tag{B.1}$$

as $\delta \to 0$ because the function W_δ is uniformly bounded and $E\{W_\delta(z,\mathbf{x},\mathbf{p}_1)\}$ does not depend on δ. Furthermore, the correlation function at the same spatial point, but for two different values of the wave vector, $U_\delta^{(2)}(z,\mathbf{x},\mathbf{p}_1,\mathbf{p}_2) = E\{W_\delta(z,\mathbf{x},\mathbf{p}_1)W_\delta(z,\mathbf{x},\mathbf{p}_2)\}$, is the solution of (2.18) with $N=2$ and the initial data

$$W_\delta^{(2)}(0,\mathbf{x}_1,\mathbf{p}_1,\mathbf{x}_2,\mathbf{p}_2) = W_I(\mathbf{x}_1,\mathbf{p}_1)W_I(\mathbf{x}_2,\mathbf{p}_2),$$

evaluated at $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}$. Therefore, $U_\delta^{(2)}$ may be represented as

$$U_\delta^{(2)}(z,\mathbf{x}_1,\mathbf{p}_1,\mathbf{x}_2,\mathbf{p}_2) = E\left\{W_I(\mathbf{X}_\delta^1(z),\mathbf{P}_\delta^1(z))W_I(\mathbf{X}_\delta^2(z),\mathbf{P}_\delta^2(z))\right\}.$$

The processes $\mathbf{X}_\delta^{1,2}$ and $\mathbf{P}_\delta^{1,2}$ satisfy the system of stochastic differential equations (2.16), which may be more explicitly written as

$$\begin{aligned}
d\mathbf{P}_\delta^1 &= -\left[\sigma(0)d\mathbf{B}^1(z) + \frac{1}{2}\sigma\left(\frac{\mathbf{X}_\delta^1 - \mathbf{X}_\delta^2}{\delta}\right)d\mathbf{B}^2(z)\right], \\
d\mathbf{P}_\delta^2 &= -\left[\sigma(0)d\mathbf{B}^2(z) + \frac{1}{2}\sigma\left(\frac{\mathbf{X}_\delta^2 - \mathbf{X}_\delta^1}{\delta}\right)d\mathbf{B}^1(z)\right], \\
d\mathbf{X}_\delta^1 &= -\mathbf{P}_\delta^1 dz, \quad d\mathbf{X}_\delta^2 = -\mathbf{P}_\delta^2 dz
\end{aligned} \tag{B.2}$$

with the initial conditions $\mathbf{X}_\delta^{1,2}(0) = \mathbf{x}$, $\mathbf{P}_\delta^m(0) = \mathbf{p}_m$, $m = 1,2$. Here $\sigma^2(0) = D$, the diffusion coefficient (2.15), and the coupling matrix $\sigma(\mathbf{x})$ is given by (2.17). Recall that $W_\delta(z,\mathbf{x},\mathbf{p},k) = W_\delta(z,\mathbf{x},\mathbf{p}/k;1)$, and we need only consider the case $k=1$.

It is convenient to introduce the processes $\mathbf{X}^{1,2}$ and $\mathbf{P}^{1,2}$ that are solutions of (2.16) with no coupling,

$$\begin{aligned}
&d\mathbf{P}^m = -\sigma(0)d\mathbf{B}^m(z), \quad d\mathbf{X}^m = -\mathbf{P}^m dz, \\
&\mathbf{X}^{1,2}(0) = \mathbf{x}, \quad \mathbf{P}^m(0) = \mathbf{p}_m, \quad m = 1,2,
\end{aligned} \tag{B.3}$$

and define the deviations of the solutions of the coupled system of stochastic differential equations (B.2) from those of (B.3): $\mathbf{Z}_\delta^m = \mathbf{X}_\delta^m - \mathbf{X}^m$, $\mathbf{S}_\delta^m = \mathbf{P}_\delta^m - \mathbf{P}^m$. Then we have

$$\begin{aligned}
&d\mathbf{S}_\delta^1 = -\frac{1}{2}\sigma\left(\frac{\mathbf{X}_\delta^1 - \mathbf{X}_\delta^2}{\delta}\right)d\mathbf{B}^2(z), \quad d\mathbf{S}_\delta^2 = -\frac{1}{2}\sigma\left(\frac{\mathbf{X}_\delta^2 - \mathbf{X}_\delta^1}{\delta}\right)d\mathbf{B}^1(z), \\
&d\mathbf{Z}_\delta^1 = -\mathbf{S}_\delta^1 dz, \quad d\mathbf{Z}_\delta^2 = -\mathbf{S}_\delta^2 dz
\end{aligned} \tag{B.4}$$

with the initial data $\mathbf{S}_\delta^m(0) = \mathbf{Z}^m(0) = 0$. Define

$$\begin{aligned}
&\mathcal{V}(\mathbf{X}^1,\mathbf{X}^2,\mathbf{P}^1,\mathbf{P}^2,\mathbf{Z}_\delta^1,\mathbf{Z}_\delta^2,\mathbf{S}_\delta^1,\mathbf{S}_\delta^2) \\
&= W_I(\mathbf{X}^1 + \mathbf{Z}_\delta^1, \mathbf{P}^1 + \mathbf{S}_\delta^1)W_I(\mathbf{X}^2 + \mathbf{Z}_\delta^2, \mathbf{P}^2 + \mathbf{S}_\delta^2) - W_I(\mathbf{X}^1,\mathbf{P}^1)W_I(\mathbf{X}^2,\mathbf{P}^2).
\end{aligned} \tag{B.5}$$

Then we have, with the above notation,

$$\begin{aligned}
&E\{W_\delta(z,\mathbf{x},\mathbf{p}_1)W_\delta(z,\mathbf{x},\mathbf{p}_2)\} - E\{W_\delta(z,\mathbf{x},\mathbf{p}_1)\}E\{W_\delta(z,\mathbf{x},\mathbf{p}_2)\} \\
&= E\left\{\mathcal{V}(\mathbf{X}^1(z),\mathbf{X}^2(z),\mathbf{P}^1(z),\mathbf{P}^2(z),\mathbf{Z}_\delta^1(z),\mathbf{Z}_\delta^2(z),\mathbf{S}_\delta^1(z),\mathbf{S}_\delta^2(z))\right\} \\
&\le CE\left\{|\mathbf{Z}_\delta^1(z)| + |\mathbf{Z}_\delta^2(z)| + |\mathbf{S}_\delta^1(z)| + |\mathbf{S}_\delta^2(z)|\right\}
\end{aligned} \tag{B.6}$$

since W_I is a Lipschitz function.

Let us assume for simplicity that the correlation function $R(\mathbf{x})$ has compact support inside the set $|\mathbf{x}| \le M$. Then the coupling term in (B.2) is nonzero only when $|\mathbf{X}_\delta^1 - \mathbf{X}_\delta^2| \le M\delta$. We introduce the processes $\mathbf{Q}_\delta = \mathbf{P}_\delta^1 - \mathbf{P}_\delta^2$ and $\mathbf{Y}_\delta = \mathbf{X}_\delta^1 - \mathbf{X}_\delta^2$ that govern (B.4). They satisfy the stochastic differential equations

$$d\mathbf{Q}_\delta = -\left[\sigma(0) - \frac{1}{2}\sigma\left(\frac{\mathbf{Y}_\delta}{\delta}\right)\right] d\tilde{\mathbf{B}}, \quad d\mathbf{Y}_\delta = -\mathbf{Q}_\delta dz, \tag{B.7}$$
$$\mathbf{Q}_\delta(0) = \mathbf{p}_1 - \mathbf{p}_2, \ \mathbf{Y}_\delta(0) = 0$$

with $\tilde{\mathbf{B}} = \mathbf{B}^1 - \mathbf{B}^2$ being a Brownian motion.

In order to prove the theorem, we show that the coupling term $\sigma(\cdot)$ in (B.2) introduces only lower order correction terms; that is, $\mathbf{S}_\delta^m$ and $\mathbf{Z}_\delta^m$ are small. We first show that after a small "time" τ, the points $\mathbf{X}_\delta^m$ are driven apart since $\mathbf{Q}_\delta(0) = \mathbf{P}_\delta^1(0) - \mathbf{P}_\delta^2(0) \neq 0$. Then we show that after the points have separated the probability that they come close so that the coupling term $\sigma(\cdot)$ becomes nonzero is small. This "nonrecurrence" condition requires that the spatial dimension $d \ge 2$. It follows that to leading order, the points $\mathbf{X}_\delta^m$ are uncorrelated when $d \ge 2$ and that the coupling term introduces only lower order corrections. A similar argument for $d = 1$ would require an estimate of the time that points that are originally separated in the spatial variable spend near each other, where the coupling term in (B.2) is not zero.

We need the following two lemmas. The first one shows that particles that start at the same point $\mathbf{x}$, with different initial directions $\mathbf{p}_1$ and $\mathbf{p}_2$, get separated with a large probability.

LEMMA B.1. *Let $\mathbf{Y}_\delta$, $\mathbf{Q}_\delta$ solve* (B.7) *with $\mathbf{Y}_\delta(0) = 0$, $\mathbf{Q}_\delta(0) = \mathbf{q} \neq 0$. Then for any $\varepsilon > 0$ there exists $\tau_0(\varepsilon) > 0$ that depends only on $\mathbf{q} = \mathbf{p}_1 - \mathbf{p}_2$ but not on δ so that we have $P(|\mathbf{Y}_\delta(\tau)| \ge \frac{|\mathbf{q}|\tau}{2}) \ge 1 - \varepsilon$ for all $\tau \le \tau_0(\varepsilon)$.*

The second lemma shows that, after the particles are separated, the probability that they come close to each other is small.

LEMMA B.2. *Given any fixed $r > 0$ and $z > 0$, if $\mathbf{Y}_\delta$, $\mathbf{Q}_\delta$ solve* (B.7) *with $|\mathbf{Y}_\delta(0)| \ge r$, $\mathbf{Q}_\delta(0) = \mathbf{q} \neq 0$, then $P\left(\inf_{0\le s\le z} |\mathbf{Y}_\delta(s)| \le M\delta\right) \to 0$ as $\delta \to 0$.*

We prove Theorem 2.1 before proving Lemmas B.1 and B.2.

Proof. Let z and $\mathbf{q} = \mathbf{p}_1 - \mathbf{p}_2$ be fixed and defined as above. Given $\varepsilon > 0$, then for any $\tau < \tau_0(\varepsilon)$ (with τ_0 as defined in Lemma B.1), Lemma B.2 and the Markov property of the Brownian motion imply that

$$P\left(\mathbf{S}_\delta^m(z) = \mathbf{S}_\delta^m(\tau) \Big| |\mathbf{Y}_\delta(\tau)| \ge \frac{\tau|\mathbf{q}|}{2}\right) \ge 1 - \varepsilon$$

and

$$P\left(\mathbf{Z}_\delta^m(z) = \mathbf{Z}_\delta^m(\tau) + (z-\tau)\mathbf{S}_\delta^m(\tau) \Big| |\mathbf{Y}_\delta(\tau)| \ge \frac{\tau|\mathbf{q}|}{2}\right) \ge 1 - \varepsilon$$

for $\delta < \delta_0(\tau, \varepsilon)$. Furthermore,

$$E\left\{|\mathbf{Z}_\delta^1(\tau)| + |\mathbf{Z}_\delta^2(\tau)| + |\mathbf{S}_\delta^1(\tau)| + |\mathbf{S}_\delta^2(\tau)| \Big| |\mathbf{Y}_\delta(\tau)| \ge \frac{\tau|\mathbf{q}|}{2}\right\} \tag{B.8}$$
$$\le E\left\{|\mathbf{Z}_\delta^1(\tau)| + |\mathbf{Z}_\delta^2(\tau)| + |\mathbf{S}_\delta^1(\tau)| + |\mathbf{S}_\delta^2(\tau)|\right\}/(1-\varepsilon) \le C\tau$$

because the function σ is uniformly bounded. Therefore, we have

$$E\left\{\mathcal{V}(\mathbf{X}^1, \mathbf{X}^2, \mathbf{P}^1, \mathbf{P}^2, \mathbf{Z}_\delta^1, \mathbf{Z}_\delta^2, \mathbf{S}_\delta^1, \mathbf{S}_\delta^2)\right\}$$

$$= E\left\{\mathcal{V}(\mathbf{X}^1,\mathbf{X}^2,\mathbf{P}^1,\mathbf{P}^2,\mathbf{Z}^1_\delta,\mathbf{Z}^2_\delta,\mathbf{S}^1_\delta,\mathbf{S}^2_\delta)\Big||\mathbf{Y}_\delta(\tau)|\geq\frac{\tau|\mathbf{q}|}{2}\right\}P\left(|\mathbf{Y}_\delta(\tau)|\geq\frac{\tau|\mathbf{q}|}{2}\right) \tag{B.9}$$

$$+E\left\{\mathcal{V}(\mathbf{X}^1,\mathbf{X}^2,\mathbf{P}^1,\mathbf{P}^2,\mathbf{Z}^1_\delta,\mathbf{Z}^2_\delta,\mathbf{S}^1_\delta,\mathbf{S}^2_\delta)\Big||\mathbf{Y}_\delta(\tau)|\leq\frac{\tau|\mathbf{q}|}{2}\right\}P\left(|\mathbf{Y}_\delta(\tau)|\leq\frac{\tau|\mathbf{q}|}{2}\right)=I+II.$$

The second term above is small because the probability for $\mathbf{Y}_\delta(\tau)$ to be very small is bounded by Lemma B.1. More precisely, given $\varepsilon>0$ and $\tau<\tau_0(\varepsilon)$, Lemma B.1 implies that

$$II\leq C\varepsilon. \tag{B.10}$$

The first term in (B.9) corresponds to the more likely scenario that $\mathbf{Y}_\delta$ at time τ has left the ball of radius $\tau|\mathbf{q}|/2$. We estimate it as follows. The probability that $\mathbf{Y}_\delta$ re-enters the ball of radius $M\delta$ is small according to Lemma B.2. Moreover, if $\mathbf{Y}_\delta$ stays outside this ball, the difference variables $\mathbf{Z}^m$ and $\mathbf{S}^m$ are bounded in terms of their values at time τ. The latter are small if τ is small. More precisely, using (B.8), we choose τ so small that

$$E\left\{|\mathbf{Z}^1_\delta(\tau)|+|\mathbf{Z}^2_\delta(\tau)|+|\mathbf{S}^1_\delta(\tau)|+|\mathbf{S}^2_\delta(\tau)|\Big||\mathbf{Y}_\delta(\tau)|\geq\frac{\tau|\mathbf{q}|}{2}\right\}\leq\varepsilon.$$

Then we obtain

$$I\leq E\left\{\mathcal{V}(\mathbf{X}^1,\mathbf{X}^2,\mathbf{P}^1,\mathbf{P}^2,\mathbf{Z}^1_\delta,\mathbf{Z}^2_\delta,\mathbf{S}^1_\delta,\mathbf{S}^2_\delta)\Big||\mathbf{Y}_\delta(\tau)|\geq\frac{\tau|\mathbf{q}|}{2}\right\}$$

$$\leq E\left\{\mathcal{V}(\mathbf{X}^1,\mathbf{X}^2,\mathbf{P}^1,\mathbf{P}^2,\mathbf{Z}^1_\delta,\mathbf{Z}^2_\delta,\mathbf{S}^1_\delta,\mathbf{S}^2_\delta)\Big||\mathbf{Y}_\delta(\tau)|\geq\frac{\tau|\mathbf{q}|}{2}\text{ and }\inf_{\tau\leq s\leq z}|\mathbf{Y}_\delta(s)|\leq M\delta\right\}$$

$$\times P\left(\inf_{\tau\leq s\leq z}|\mathbf{Y}_\delta(s)|\leq M\delta\Big||\mathbf{Y}_\delta(\tau)|\geq\frac{\tau|\mathbf{q}|}{2}\right)$$

$$+E\left\{|\mathbf{Z}^1_\delta(z)|+|\mathbf{Z}^2_\delta(z)|+|\mathbf{S}^1_\delta(z)|+|\mathbf{S}^2_\delta(z)|\Big||\mathbf{Y}_\delta(\tau)|\geq\frac{\tau|\mathbf{q}|}{2}\text{ and }\inf_{\tau\leq s\leq z}|\mathbf{Y}_\delta(s)|\geq M\delta\right\}$$

$$\times P\left(\inf_{\tau\leq s\leq z}|\mathbf{Y}_\delta(s)|\geq M\delta\Big||\mathbf{Y}_\delta(\tau)|\geq\frac{\tau|\mathbf{q}|}{2}\right)=I_1+I_2.$$

The term I_1 goes to zero as $\delta\to0$ by Lemma B.2. However, if the conditions in I_2 hold, then

$$\mathbf{S}^m_\delta(z)=\mathbf{S}^m_\delta(\tau),\quad\mathbf{Z}^m_\delta(z)=\mathbf{Z}^m_\delta(\tau)-\frac{1}{k}(z-\tau)\mathbf{S}^m_\delta(\tau).$$

Therefore, the term I_2 may be bounded with the help of (B.8) by

$$I_2\leq E\left\{|\mathbf{Z}^1_\delta(z)|+|\mathbf{Z}^2_\delta(z)|+|\mathbf{S}^1_\delta(z)|+|\mathbf{S}^2_\delta(z)|\Big||\mathbf{Y}_\delta(\tau)|\geq\frac{\tau|\mathbf{q}|}{2}\text{ and }\inf_{\tau\leq s\leq z}|\mathbf{Y}_\delta(s)|\geq M\delta\right\}$$

$$\leq CE\left\{|\mathbf{Z}^1_\delta(\tau)|+|\mathbf{Z}^2_\delta(\tau)|+|\mathbf{S}^1_\delta(\tau)|+|\mathbf{S}^2_\delta(\tau)|\Big||\mathbf{Y}_\delta(\tau)|\geq\frac{\tau|\mathbf{q}|}{2}\right\}\leq C\tau.$$

Putting together (B.9), (B.10), and the above bounds on I_1 and I_2, we obtain

$$E\left\{|\mathbf{Z}^1_\delta(z)|+|\mathbf{Z}^2_\delta(z)|+|\mathbf{S}^1_\delta(z)|+|\mathbf{S}^2_\delta(z)|\right\}\leq C\varepsilon$$

for $\delta<\bar{\delta}$, and Theorem 2.1 follows from (B.6). □

B.2. Proofs of Lemmas B.1 and B.2. We first prove Lemma B.1.

Proof. We write

$$Q_\delta(z) = \mathbf{q} - \int_0^z \left(\sigma(0) - \frac{1}{2}\sigma(\mathbf{Y}(s)/\delta)\right)\, d\tilde{\mathbf{B}}(s) \;\equiv\; \mathbf{q} + \tilde{\mathbf{Q}}_\delta(z)$$

so that

$$\mathbf{Y}_\delta(t) = -\mathbf{q}t - \int_0^t \tilde{\mathbf{Q}}_\delta(s)ds.$$

Then we have

$$P\left(\sup_{0\le s\le\tau} |\tilde{\mathbf{Q}}_\delta(s)| \;>\; r\right) \le C\tau/r^2 \tag{B.11}$$

and hence

$$P\left(|\mathbf{Y}_\delta(\tau) + \tau\mathbf{q}| > r\tau\right) \le P\left(\sup_{0\le s\le\tau} |\tilde{\mathbf{Q}}_\delta(s)| > r\right) \le C\tau/r^2.$$

We let $r = |\mathbf{q}|/2$ in the above formula and obtain

$$P\left(|\mathbf{Y}_\delta(\tau)| < \frac{\tau|\mathbf{q}|}{2}\right) \le \frac{C}{|\mathbf{q}|^2}\tau,$$

and the conclusion of Lemma B.1 follows. ∎

Finally, we prove Lemma B.2.

Proof. Let τ_δ be the first time $\mathbf{Y}_\delta(z)$ enters the ball of radius $M\delta$,

$$\tau_\delta = \inf\left\{z :\; |\mathbf{Y}_\delta(z)| \le M\delta\right\},$$

with $\mathbf{Y}_\delta(0) = \mathbf{Y}^0 \neq 0$. For $0 < \alpha < 1$, let $\Delta z = \delta^{1-\alpha}$, $n = \lceil z/\Delta z\rceil$, $J_i = (i\Delta z, (i+1)\Delta z)$, and $p < 1$. Note that, until the time τ_δ, the process $(\mathbf{Y}_\delta, \mathbf{Q}_\delta)$ coincides with the process $(\mathbf{Y}, \mathbf{Q})$ governed by (B.7) without the coupling term $\sigma(\mathbf{Y}_\delta/\delta)$. We find

$$P(\tau_\delta < z) \le \sum_{i=0}^{n-1} \left\{ P\left(|\mathbf{Y}(i\Delta z)| < M\delta^p\right) + P\left(\inf_{s\in J_i} |\mathbf{Y}(s)| < M\delta \;\middle|\; |\mathbf{Y}(i\Delta z)| \ge M\delta^p\right)\right\}.$$

The process $\mathbf{Y}(s)$ is Gaussian with mean $\mathbf{Y}^0$ and variance $\mathcal{O}(s^2)$. Therefore, there is a $\bar\delta > 0$ such that for $\delta < \bar\delta$

$$P(|\mathbf{Y}(i\Delta z)| < M\delta^p) \le C\delta^{dp}.$$

If we assume

$$p < 1 - \alpha, \tag{B.12}$$

then also

$$\begin{aligned} P(\tau_\delta < z) &\le nC\left(\delta^{dp} + P\left(\sup_{0<s<\Delta z} |\mathbf{Y}(s) - \mathbf{Y}^0| \ge M[\delta^p - \delta]\right)\right) \\ &\le C\left(\delta^{dp+\alpha-1} + \delta^{\alpha-1}\frac{E\left\{\mathbf{B}(\delta z)^{2r}\right\}\Delta z^{2r}}{(\delta^p - \delta)^{2r}}\right) \\ &\le C\left[\delta^{dp+\alpha-1} + \delta^{\alpha-1-rp+3r(1-\alpha)/2}\right]. \end{aligned}$$

Note that, with $p < 1 - \alpha$ and r large enough, there is a $q > 0$ so that

$$P(\tau_\delta < z) \le C\delta^q$$

if $d \ge 2$, and Lemma B.2 follows. ∎

B.3. Proof of Theorem 2.2. We need to first show that

$$J_\delta(z,\mathbf{x}) = \int W_\delta(z,\mathbf{x},\mathbf{p})d\mathbf{p} \tag{B.13}$$

is finite with probability one. The stochastic flow $(\mathbf{X}_\delta(z,\mathbf{x},\mathbf{p}),\mathbf{P}_\delta(z,\mathbf{x},\mathbf{p})$ is continuous in $(z,\mathbf{x},\mathbf{p})$ with probability one, so $W_\delta(z,\mathbf{x},\mathbf{p}) = W_I(\mathbf{X}_\delta(z,\mathbf{x},\mathbf{p}),\mathbf{P}_\delta(z,\mathbf{x},\mathbf{p}))$ is bounded and continuous. It is, moreover, nonnegative if $W_I \geq 0$. We know that

$$\int E\{W_\delta(z,\mathbf{x},\mathbf{p})\}d\mathbf{p}$$

is finite and independent of δ, and the order of integration and expectation can be interchanged by Tonelli's theorem. This theorem implies in addition that $J_\delta(z,\mathbf{x})$ is finite with probability one.

We can now consider

$$E\{J_\delta^2(z,\mathbf{x})\} = \int E\{W_\delta(z,\mathbf{x},\mathbf{p}_1)W_\delta(z,\mathbf{x},\mathbf{p}_2)\}d\mathbf{p}_1 d\mathbf{p}_2.$$

The integrand is bounded by an integrable function uniformly in δ because

$$E\{W_\delta(z,\mathbf{x},\mathbf{p}_1)W_\delta(z,\mathbf{x},\mathbf{p}_2)\} \leq E^{1/2}\{W_\delta^2(z,\mathbf{x},\mathbf{p}_1)\}E^{1/2}\{W_\delta(z,\mathbf{x},\mathbf{p}_2)\};$$

the right side does not depend on δ and is integrable. Therefore, by the Lebesgue dominated convergence theorem and the results of the previous section, we have that

$$\lim_{\delta\to 0} E\{J_\delta^2(z,\mathbf{x})\} = E^2\{J_\delta(z,\mathbf{x})\}$$

and the right side does not depend on δ. This completes the proof of Theorem 2.2.

REFERENCES

[1] F. Bailly, J.F. Clouet, and J.P. Fouque, *Parabolic and Gaussian white noise approximation for wave propagation in random media*, SIAM J. Appl. Math., 56 (1996), pp. 1445–1470.

[2] G. Bal and L. Ryzhik, *Time reversal for classical waves in random media*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 1041–1046.

[3] G. Bal and L. Ryzhik, *Time reversal and refocusing in random media*, SIAM J. Appl. Math., 63 (2003), pp. 1475–1498.

[4] G. Bal, G. Papanicolaou, and L. Ryzhik, *Self-averaging in time reversal for the parabolic wave equation*, Stoch. Dyn., 2 (2002), pp. 507–531.

[5] G. Bal, G. Papanicolaou, and L. Ryzhik, *Radiative transport limit for the random Schrödinger equation*, Nonlinearity, 15 (2002), pp. 513–529.

[6] G. Blankenship and G. C. Papanicolaou, *Stability and control of stochastic systems with wide-band noise disturbances*, SIAM J. Appl. Math., 34 (1978), pp. 437–476.

[7] P. Blomgren, G. Papanicolaou, and H. Zhao, *Super-resolution in time-reversal acoustics*, J. Acoust. Soc. Am., 111 (2002), pp. 230–248.

[8] L. Borcea, C. Tsogka, G. Papanicolaou, and J. Berryman, *Imaging and time reversal in random media*, Inverse Problems, 18 (2002), pp. 1247–1279.

[9] J. Berryman, L. Borcea, G. Papanicolaou, and C. Tsogka, *Statistically stable ultrasonic imaging in random media*, J. Acoust. Soc. Am., 111 (2002), pp. 230–248.

[10] R. Bouc and E. Pardoux, *Asymptotic analysis of PDEs with wide-band noise disturbances and expansion of the moments*, Stochastic Anal. Appl., 2 (1984), pp. 369–422.

[11] D. Dawson and G. Papanicolaou, *A random wave process*, Appl. Math. Optim., 12 (1984), pp. 97–114.

[12] D. Dowling and D. Jackson, *Phase conjugation in underwater acoustics*, J. Acoust. Soc. Am., 89 (1990), pp. 171–181.

[13] D. Dowling and D. Jackson, *Narrow-band performance of phase-conjugate arrays in dynamic random media*, J. Acoust. Soc. Am., 91 (1992), pp. 3257–3277.
[14] M. Fink and J. de Rosny, *Time-reversed acoustics in random media and in chaotic cavities*, Nonlinearity, 15 (2002), pp. R1–R18.
[15] M. Fink, D. Cassereau, A. Derode, C. Prada, P. Roux, M. Tanter, J.L. Thomas, and F. Wu, *Time-reversed acoustics*, Rep. Progr. Phys., 63 (2000), pp. 1933–1995.
[16] M. Fink and C. Prada, *Acoustic time-reversal mirrors*, Inverse Problems, 17 (2001), pp. R1–R38.
[17] J.P. Fouque, G.C. Papanicolaou, and Y. Samuelides, *Forward and Markov approximation: The strong intensity fluctuations regime revisited*, Waves Random Media, 8 (1998), pp. 303–314.
[18] K. Furutsu, *Random Media and Boundaries: Unified Theory, Two-Scale Method, and Applications*, Springer-Verlag, Berlin, 1993.
[19] P. Gérard, P. Markovich, N. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 323–380.
[20] W. Hodgkiss, H. Song, W. Kuperman, T. Akal, C. Ferla, and D. Jackson, *A long-range and variable focus phase-conjugation experiment in a shallow water*, J. Acoust. Soc. Am., 105 (1999), pp. 1597–1604.
[21] H. Kesten and G. Papanicolaou, *A limit theorem for turbulent diffusion*, Comm. Math. Phys., 65 (1979), pp. 97–128.
[22] G. Papanicolaou and W. Kohler, *Asymptotic analysis of deterministic and stochastic equations with rapidly varying components*, Comm. Math. Phys., 45 (1975), pp. 217–232.
[23] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge Stud. Adv. Math. 24, Cambridge University Press, Cambridge, UK, 1997.
[24] W. Kuperman, W. Hodgkiss, H. Song, T. Akal, C. Ferla, and D. Jackson, *Phase-conjugation in the ocean*, J. Acoust. Soc. Am., 102 (1997), pp. 1–16.
[25] W. Kuperman, W. Hodgkiss, H. Song, T. Akal, C. Ferla, and D. Jackson, *Phase conjugation in the ocean: Experimental demonstration of an acoustic time reversal mirror*, J. Acoust. Soc. Am., 103 (1998), pp. 25–40.
[26] H. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press Series in Signal Processing, Optimization, and Control, MIT Press, Cambridge, MA, 1984.
[27] B. Nair and B. S. White, *High-frequency wave propagation in random media—a unified approach*, SIAM J. Appl. Math., 51 (1991), pp. 374–411.
[28] L. Ryzhik, G. Papanicolaou, and J. B. Keller, *Transport equations for elastic and other waves in random media*, Wave Motion, 24 (1996), pp. 327–370.
[29] F. Tappert, *The parabolic approximation method*, in Wave Propagation and Underwater Acoustics, Lecture Notes in Phys. 70, Springer-Verlag, Berlin, 1977, pp. 224–287.
[30] V. I. Tatarskii, A. Ishimaru, and V. U. Zavorotny, eds., *Wave Propagation in Random Media (Scintillation)*, SPIE, Bellingham, WA, IOP, Bristol, UK, 1993.
[31] J. L. Thomas and M. Fink, *Ultrasonic beam focusing through tissue inhomogeneities with a time reversal mirror: Application to transskull therapy*, IEEE Trans. Ultrasonics, Ferroelectrics and Frequency Control, 43 (1996), pp. 1122–1129.
[32] C. Tsogka and G. Papanicolaou, *Time reversal through a solid-liquid interface and super-resolution*, Inverse Problems, 18 (2002), pp. 1639–1657.
[33] B. S. White, *The stochastic caustic*, SIAM J. Appl. Math., 44 (1984), pp. 127–149.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1156–1174

STRUCTURE FROM MOTION: A NEW LOOK FROM THE POINT OF VIEW OF INVARIANT THEORY*

PIERRE-LOUIS BAZIN† AND MIREILLE BOUTIN‡

Abstract. We present a novel simple formulation of the problem of 3D object reconstruction from images. In this formulation, the object is seen as lying at the intersection of the projection of orbits of custom built Lie group actions. The group parameters correspond to unknown irrelevant quantities such as the camera orientation, the depth parameters of the object with respect to the camera, and the focal length. We then use an algorithmic method based on moving frames à la Fels–Olver to obtain a fundamental set of invariants of these group actions. The invariants are used to define a set of equations determining the 3D object, thus providing a mathematical formulation of the problem where the irrelevant parameters do not appear.

Key words. structure from motion, group action, invariants, moving frame, pinpoint camera, orthographic camera

AMS subject classifications. 68T45, 53

DOI. 10.1137/S003613990340246X

1. Introduction.

This paper has two goals. Its first goal is to illustrate the potential of using the formalism of invariant theory in certain applications. This potential is, at this point, rather unexploited, and we hope to set a trend with these results. Its second goal is to provide new insights on the problem of structure from motion through a novel formulation in terms of group actions.

The problem of structure from motion is rather old and well studied. It consists of reconstructing an object from a set of pictures of this object (e.g., a movie). In this paper, we consider the case of objects represented by an ordered set of points in $\mathbb{R}^3$ and assume that the camera parameters (position and orientation of the camera, focal length) are unknowns.

The concept of invariance is of major importance in modern geometry. In the field of computer vision, invariants of classical groups have been used in the design of methods of object recognition and reconstruction for more than a decade (see [12, 13]). In particular, the invariants of the projective and affine transformation groups have been widely used [16, 21, 18]. However, invariant theory can also deal with a variety of other group actions such as the ones we encounter in the problem of structure from motion.

For our purposes, invariants are defined as real-valued functions on a manifold M which remain unchanged under the action (denoted by $*$) of a group G on M. The case of Lie group actions is particularly interesting. (The reader unfamiliar with the concept of Lie group actions may refer to [8] for an introduction.) When the Lie group action satisfies certain conditions, there exists a finite set of *fundamental* invariants $I_1, \ldots, I_N$, which are local coordinates for the quotient space M/G. In other words, for any point $z \in M$, there exists a neighborhood U of z, which can be written as

*Received by the editors February 25, 2003; accepted for publication (in revised form) July 7, 2003; published electronically April 21, 2004. This work was supported by NSF grants KDI BCS-9980091 and 0074276.

http://www.siam.org/journals/siap/64-4/40246.html

†Department of Engineering, Brown University, Providence, RI 02912 (bazin@lems.brown.edu).

‡Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany. Current address: Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2068 (boutin@math.purdue.edu, boutin@mis.mpg.de).

$U = U_1 \times U_2$, where U_1 is coordinatized by the value of the invariants and U_2 is coordinatized by some (or all) of the group parameters. A modern theory of moving frames recently developed by Fels and Olver [5, 6] provides us with a systematic way to obtain a set of fundamental invariants for any (regular) Lie group action.

One reason for interest in being able to obtain a set of fundamental invariants in a systematic manner is the following. Many problems involve unknown irrelevant parameters. Often, one needs to solve for these irrelevant parameters only because they are involved in the intermediate steps of the solution process, although they do not appear in the final solution. When the irrelevant parameters can be seen as group parameters transforming the other unknowns of the problem, using the coordinates provided by the invariants is a simple way to eliminate them. In these circumstances, the moving frame method is used as a computational method to eliminate unwanted unknowns in a set of equations.

For example, suppose that given the values of $x_1, x_2, y_1, y_2 \in \mathbb{R}$, one is interested in finding the values of the unknowns z_1 and $z_2 \in \mathbb{R}$. Assume that there exist parameters u and $v \in \mathbb{R}$ for which a set of equations of the type

$$\left.\begin{array}{rcl} z_1 & = & f_1(x_1, x_2, u), \\ z_2 & = & f_2(x_1, x_2, u), \\ z_1 & = & g_1(y_1, y_2, v), \\ z_2 & = & g_2(y_1, y_2, v) \end{array}\right\} \tag{1.1}$$

holds. Since we are not interested in the values of u and v, it would be desirable to eliminate these two variables from (1.1). Suppose that the functions f_1 and f_2 correspond to an action of $\mathbb{R}$ on $\mathbb{R}^2$ parameterized by u,

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = u * \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

If this group action satisfies certain conditions (to be explained in section 3), then there exists an invariant $I : \mathbb{R}^2 \to \mathbb{R}$ such that the equation

$$I(x_1, x_2) = I(z_1, z_2)$$

can be used in place of the first two equations of (1.1). Similarly, if g_1 and g_2 correspond to an action of $\mathbb{R}$ on $\mathbb{R}^2$ parameterized by v which satisfies the correct conditions, then we can find another invariant $J : \mathbb{R}^2 \to \mathbb{R}$ such that the equation

$$J(y_1, y_2) = J(z_1, z_2)$$

can be used in place of the last two equations of (1.1). We thus see that the solution (z_1, z_2) of our problem lies at the intersection of two orbits: these are the orbit through (x_1, x_2) under the group action defined by f_1 and f_2 and the orbit through (y_1, y_2) under the group action defined by g_1 and g_2.

The problem of structure from motion is one that involves many irrelevant unknowns. For example, the camera parameters used for taking each picture are irrelevant since we are merely interested in the structure of the object. It turns out that many of the irrelevant unknowns of this problem can be seen as group parameters acting on the other unknowns and eliminated using the moving frame method.

We begin our exposition in section 2 with a summary of the relevant theoretical aspects of the theory of moving frames and its application to the computation of (joint) invariants. In section 3, we consider the problem of structure from motion when the

pictures are taken using a pinpoint camera (perspective projections). We reformulate the problem in terms of a Lie group action and use a fundamental set of invariants of this group action to get rid of most of the irrelevant parameters. (As will be explained, it is impossible to get rid of all of them.) In the case of a fixed-focus camera, the invariants describe the object with nonlinear equations involving the object points and the camera centers; the camera orientation and the depth parameters of the object with respect to the camera are included in the group parameters and thus eliminated from the problem. Eliminating the camera orientation parameters in the case of a simultaneously translating and rotating camera is known to be difficult, unless vanishing points are identified. By contrast, the invariant framework easily takes care of this problem. In addition, the invariant formulation leads to a simple test to identify camera motions that are pure rotations.

The nonlinear equations given by the invariants can be solved with regular bundle adjustment techniques [20] where the nonlinear system of equations is viewed as an optimization problem in the unknown parameters. Efficient optimization techniques like the Levenberg–Marquardt algorithm are available to solve such problems. After presenting a practical example of three-dimensional (3D) object reconstruction using the invariant framework and such an optimization technique, we finish section 3 by showing how to deal with the case of a variable focus in a similar fashion.

Of course, approaches other than this *direct* one exist for structure from motion. For example, one can describe the perspective projection in projective space where it is expressed by simple matrix equations. With projective coordinates, projective relations and constraints such as the epipolar constraint or the trifocal tensor can be used to recover both the underlying shape and camera motion from a set of multilinear equations [9]. In section 4, we apply our invariant-based method to a formulation using projective coordinates. We obtain a set of invariants involving the projective coordinates of the object points and the camera centers, which are, in fact, much simpler than the invariants of the Euclidean approach used in section 3.

Other existing approaches use an affine approximation of the projection in order to simplify the problem. This leads to factorization techniques [19] based on the SVD. In section 5, we apply our method to the case of orthographic projections and obtain a set of linear invariants involving only the object points and the directions of the normals to the camera planes. Our equations can be solved with similar factorization techniques but involve fewer and simpler parameters than the ones in [19].

Our invariant-based approach actually applies to types of cameras other than the ones discussed here. For instance, the model used in section 3 works for any central projection, whether or not the image points lie on a plane. We hope that the sample cases presented here will convince the reader of the usefulness and versatility of this group theoretic approach to eliminating unknowns and inspire new applications of invariants.

2. Definitions and theoretical foundations. Let M be an m-dimensional smooth (Hausdorff) manifold and G be an r-dimensional Lie group. Denote by e the identity in G. Let $* : G \times M \to M$ be an action of G on M, i.e., a map $(g, z) \mapsto g * z$ such that $e * z = z$ for all $z \in M$ and $(gh) * z = g * (h * z)$ for all $z \in M$ and all $g, h \in G$.

DEFINITION 2.1. *An* invariant *is a function $I : M \to \mathbb{R}$ which remains unchanged under the action of the group. In other words,*

$$I(g * z) = I(z) \text{ for all } z \in M \text{ and all } g \in G.$$

A local invariant *is a function $I : U \to \mathbb{R}$ for some open subset $U \subset M$ such that*

$$I(g * z) = I(z) \text{ for all } z \in U \text{ and all } g \in G \text{ s.t. } g * z \in U.$$

DEFINITION 2.2. *We say that G acts* semiregularly *on M if all orbits have the same dimension. If, in addition, any point $p_0 \in M$ is surrounded by an arbitrarily small neighborhood whose intersection with the orbit through p_0 is connected, then we say that G acts* regularly.

The following theorem, due to Frobenius [7], is of central importance to our approach. A proof can be found in [14]. It provides us with a simple way of characterizing the orbits using invariants.

THEOREM 2.3 (Frobenius theorem). *If G acts on an open set $O \subset M$ semiregularly with s-dimensional orbits, then for all $p_0 \in O$ there exist $m - s$ functionally independent local invariants $I_1, \ldots, I_{m-s}$ defined on a neighborhood U of p_0 such that any other local invariant H defined near p_0 is a function $H = f(I_1, \ldots, I_{m-s})$. If G acts regularly on O, then we can choose $I_1, \ldots, I_{m-s}$ to be global invariants on O. In that case, two points $p_1, p_2 \in O$ are in the same orbit relative to G if and only if $I_i(p_1) = I_i(p_2)$ for all $i = 1, \ldots, m - s$.*

By *functional independence* of the (smooth) functions $I_1, \ldots, I_{m-s}$ on an open set O, we simply mean that the Jacobian matrix of $I_1, \ldots, I_{m-s}$ has maximal rank $m - s$ on an open and dense subset of O. The set $\{I_1, \ldots, I_{m-s}\}$ is often called a *complete fundamental set of invariants on O*. Note that a complete fundamental set of invariants is not unique.

As we shall consider actions on multiple points, we are interested in the case where $M = \mathcal{V} \times \mathcal{V} \times \cdots \times \mathcal{V}$ (n-times) $=: \mathcal{V}^{\times(n)}$ is the Cartesian product of n copies of a manifold $\mathcal{V}$.

DEFINITION 2.4. *We say that G acts* diagonally *on $\mathcal{V}^{\times(n)}$ if there exists an action $\cdot$ of G on $\mathcal{V}$ such that for any $g \in G$, any $n \in \mathbb{N}$, and any $z_1, \ldots, z_n \in \mathcal{V}$, the action $g * (z_1, \ldots, z_n)$ can be written as*

$$g * (z_1, \ldots, z_n) = (g \cdot z_1, \ldots, g \cdot z_n).$$

The group actions we will define for our object-camera systems are not diagonal actions. However, for each of these actions there is a normal subgroup H of G (the subgroup generating the translations along the rays of light) such that G/H acts diagonally. So for all practical purposes, we shall ultimately have to deal with diagonal group actions.

In our approach to structure from motion, invariants are used to obtain equations that must be satisfied by the object and the camera. The more invariants we have, the more equations need to be satisfied. We need enough equations to completely determine the object. Observe that the dimension of the orbit is bounded by the dimension of the group. So in the case of a diagonal action, taking more and more copies of $\mathcal{V}$ (i.e., more and more points) allows for the existence of as many invariants as necessary. The question that remains is as follows: How can we obtain an expression for these invariants? Thanks to a new formulation of Cartan's theory of moving frames [4, 5, 6], this problem can be solved in an algorithmic fashion. We now summarize some of the relevant aspects of the theory of moving frames, including how moving frames can be used as a tool to obtain a complete set of fundamental invariants.

DEFINITION 2.5. *A* (right) moving frame *is a map $\rho : M \to G$ which is (right) equivariant; i.e., $\rho(g * z) = \rho(z)g^{-1}$ for all $g \in G$ and $z \in M$.*

Unfortunately moving frames do not exist for all group actions.

THEOREM 2.6. *A moving frame exists if and only if the action of the group satisfies*

$$\{g \in G | \exists z \in M, g * z = z\} = \{e\},$$

where e denotes the identity in G. This property is called freeness *of the group action.*

Demanding freeness of the group action is very strong. It appears that, in order to be able to deal with the generic cases, we need to relax this condition a little bit.

DEFINITION 2.7. *A* local moving frame *is a map $\rho : M \to G$ such that $\rho(g * z) = \rho(z)g^{-1}$ for all $g \in N_e$, a neighborhood of the identity $e \in G$, and all $z \in M$.*

THEOREM 2.8. *A local moving frame exists if and only if there exists a neighborhood N_e of the identity in $e \in G$ such that*

$$\{g \in N_e | \exists z \in M, g * z = z\} = \{e\}$$

or, equivalently, if and only if for all $z \in M$, the dimension of the orbit through z is equal to r, the dimension of G. This property is called local freeness *of the group action.*

We are now interested in determining a condition on the action of G on M which guarantees that the diagonal action will be locally free on a sufficiently large number of copies of M.

DEFINITION 2.9. *We say that G acts on M* effectively *if*

$$\{g \in G | \quad g * p = p \text{ for all } p \in M\} = \{e\}.$$

We say that G acts on M locally effectively *if*

$$\{g \in G | \quad g * p = p \text{ for all } p \in M\} \text{ is a discrete subgroup of } G.$$

Many groups do not act effectively. However, given G acting noneffectively on M, we can consider $\tilde{G} = G/G_M$, where $G_M = \{g \in G | g * z = z \text{for all } z \in M\}$, which acts in essentially the same way as G except that it acts effectively. Unfortunately, effectiveness is not sufficient to guarantee that the diagonal action eventually becomes locally free.

DEFINITION 2.10. *We say that G acts* effectively on subsets of *M if, for any open subset $U \subset M$,*

$$\{g \in G | \quad g * p = p \text{ for all } p \in U\} = \{e\}.$$

We say that G acts locally effectively on subsets of *M if, for any open subset $U \subset M$,*

$$\{g \in G | \quad g * p = p \text{ for all } p \in M\} \text{ is a discrete subgroup of } G.$$

Observe that effectiveness on subsets implies effectiveness. The converse, of course, holds for all analytic group actions. However, this is not true in general (see [2] for a counterexample).

THEOREM 2.11 (see [2]). *If a group G acts on a manifold $\mathcal{V}$ locally effectively on subsets, then there exists $n \in N^+$ such that the induced diagonal action of G on $\mathcal{V}^{\times(n)}$ is locally free on an open and dense subset of $\mathcal{V}^{\times(n)}$. This is equivalent to saying that the orbit dimension is equal to the dimension of G on this open and dense subset. We denote by n_0 the minimal integer for which this is true.*

This means that any group action that is effective on subsets (e.g., any analytic group action, once the subgroup acting trivially is moded out) will be locally free on a sufficiently large number of copies of the manifold and a local moving frame will exist on this product.

We now explain how to construct a (local) moving frame and to obtain a complete fundamental set of invariants. A more detailed exposition can be found in [14, Chapter 8]. Let $g = (g_1, \dots, g_r)$ be local coordinates for G in a neighborhood of the identity. Suppose that G acts regularly on M. For simplicity, let us assume in addition that the orbits of G have the same dimension r as G itself. In other words, we are assuming that the action is locally free. Shortly after, we will explain how to deal with the case of merely regular actions using a simple variation of the following algorithm.

- *Step* 1. Write down the group transformation equations $\bar{x} = g * x$ explicitly.

$$\begin{cases} \bar{x}_1 &= f_1(g_1, \dots, g_r, x_1, \dots, x_m), \\ &\vdots \\ \bar{x}_m &= f_m(g_1, \dots, g_r, x_1, \dots, x_m). \end{cases}$$

- *Step* 2. Choose constants $c_1, \dots, c_r \in \mathbb{R}$ and set r of the transformed coordinates equal to those constants. For simplicity, we relabel the coordinates and write

$$\begin{cases} f_1(g_1, \dots, g_r, x_1, \dots, x_m) &= c_1, \\ &\vdots \\ f_r(g_1, \dots, g_r, x_1, \dots, x_m) &= c_r. \end{cases} \tag{2.1}$$

  These equations are called the *normalization equations.*
- *Step* 3. Solve the normalization equations for $g = (g_1, \dots, g_r)$. The solution $g = \rho(x)$ is a moving frame.
- *Step* 4. Compute the action of the moving frame on the remaining coordinates. The set of resulting functions,

$$\begin{cases} \bar{x}_{r+1}\big|_{g=\rho(x)} &= I_1(x_1, \dots, x_m), \\ &\vdots \\ \bar{x}_m\big|_{g=\rho(x)} &= I_{m-s}(x_1, \dots, x_m), \end{cases}$$

  is a complete fundamental set of local invariants.

The choice of constants in Step 2 is somewhat arbitrary; we are free to choose any numbers for which a solution to the normalization equations exists, provided that these constants define a cross-section (i.e., provided that the normalization equations define a submanifold which is transversal to the orbits). To simplify the solution process, it is usually a good idea to choose as many constants as possible to be zero.

If the action is not free but merely regular, we can still find a system of functionally independent local invariants. We proceed as follows. Let s be the dimension of the orbits of G ($s < r$). We solve the s equations $f_1(g, x) = c_1, \dots, f_s(g, x) = c_s$ for s of the group parameters and replace them in the remaining equations $\overline{x}_{s+1} = f_{s+1}(g, x), \dots, \overline{x}_m = f_m(g, x)$ to get the $m-s$ invariants. The other group parameters $g_{s+1}, \dots, g_r$ will not appear in the final expressions. This procedure is called a *partial moving frame normalization method.*

Equipped with these tools, obtaining invariants becomes a simple systematic procedure. We can thus feel free to consider any Lie group action imaginable and try to obtain its invariants. As we have seen, in theory, the invariants can always be found provided that the group action is locally effective on subsets (which we can always arrange in the case of analytic group actions). Of course, in practice, computational difficulties can be encountered in explicitly determining the invariants. Fortunately, the invariants are easily computed for the problem of structure from motion.

3. The case of a perspective camera. Let us assume that we are given t sets of n ordered points $p_1^\tau, \ldots, p_n^\tau \in \mathbb{R}^2$, $\tau = 1, \ldots, t$, which represent t pictures of a 3D unknown object made of n ordered points $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_n \in \mathbb{R}^3$. We would like to determine the points $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_n$ from these pictures. One possible way to try to solve this problem would be to define an equivalence relation between all the possible pictures of an object and to find functions depending on the picture points which are constant on each equivalence class. Characteristics of the object could be inferred from these functions, regardless of the camera position relative to the object. To define such an equivalence class, one could try to use the orbits of a group action; in other words, one could look for *invariants* of a group action that is transitive on the set of pictures of any given object, i.e., *view invariants*.

Unfortunately, as is commonly known in the vision community, *view invariants do not exist for* $3D$ *point sets of arbitrary size (in general position).* One can still build invariants for specific objects (for instance, planar sets of points, pencils of lines, etc.) but not for arbitrary shapes. The problem is that the set of pictures of any object intersects with the set of pictures of other objects. Observe that if a view invariant I takes a constant value c for all pictures of an object O, then I is also equal to c on the set of pictures of any object whose set of pictures intersects with the set of pictures of O. One can actually show [3] that any equivalence relation between all the pictures of each object defines a unique equivalence class on the space of pictures, and thus any view invariant is trivial. From a group point of view, this means that any group action that is transitive on the set of pictures of any object must be transitive on the set of *all* pictures.

Another way to try to solve this problem would be to use the reverse approach: try to characterize all the possible objects corresponding to a given picture. It would be useful to find functions which are constant on equivalence classes that include all the objects corresponding to a given picture. However, it is easy to see that such equivalence classes of objects are in one to one correspondence with the equivalence classes of pictures discussed above and thus that only one such equivalence class can exist. There are therefore no *object invariants*. Nevertheless, given a view of an object, one *can* infer information about the object, so there must be a way to overcome this difficulty.

The trick is to define an equivalence relation on a higher dimensional space, lifting the set of objects of different pictures to different *"heights"* in the extra dimensions. For this, we can use the three extra dimensions provided by the camera center position. More precisely, we can construct a Lie group action on the object points and the camera center which summarizes what is unknown about the object-camera system given a picture of an object and knowing the mechanism used by the camera. Invariants of this group action prove to be sufficient for solving the problem of recovering the object coordinates in $\mathbb{R}^3$.

So let us think for a moment about the process of taking a picture. This process involves, first, the placement of a camera in space. Then, particles of light start from

each point of the object and travel on a straight line in the direction of the camera center, leaving their trace on a film, i.e., on the intersection of the picture plane and the travel lines. So, to the picture-camera system placed somewhere in $\mathbb{R}^3$, there corresponds a set of n straight lines in $\mathbb{R}^3$ representing the paths of light going from the object to the camera center (sometimes referred to as a "ray bundle").

This process can be seen as a result of the action of a group composed of an action of the special Euclidean group $SE(3)$ (i.e., the group of rotations and translations in $\mathbb{R}^3$) and an action of $\mathbb{R}^n$ on the camera center and the image points (in three dimensions). Here is how.

Given is a two-dimensional (2D) image depicting n points $p_1, \dots, p_n \in \mathbb{R}^2$. We assume this picture was taken by a camera with fixed internal parameters. These parameters can be calibrated beforehand so that the focal length is $\mathcal{F} = 1$[1] and the 2D image coordinates match the 3D coordinates as defined below. We embed the picture-camera system in $\mathbb{R}^3$ by setting the camera center to be $\tilde{p}_0 = (0, 0, 0)$ and the picture points $\tilde{p}_i$'s to be $\tilde{p}_i = p_i \times \mathcal{F}$. This is, in general, not the actual position in which the picture was taken. However, there exists a rigid transformation $g \in SE(3)$ acting diagonally on $(\mathbb{R}^3)^{\times n}$ such that $g * (\tilde{p}_0, \tilde{p}_1, \dots, \tilde{p}_n) = (\mathfrak{P}_0, \mathfrak{P}_1, \dots, \mathfrak{P}_n)$ corresponds to the actual position of the picture-camera system at the moment where the picture was taken. Once the picture points are in this position, we know that we can move each of them independently along each ray of light so as to go back to its source on the object. This way, the picture can be mapped onto the object.

In summary, the mapping is given by the transformation

$$\bar{P}_0 = RP_0 + T, \tag{3.1}$$

$$\bar{P}_i = R(P_i + \lambda_i(P_i - P_0)) + T \text{ for } i = 1, \dots, n, \tag{3.2}$$

with $R \in SO(3)$ a rotation, $T \in \mathbb{R}^3$ a translation, and $\lambda_i \in \mathbb{R}$ a factor of depth, applied to $P_0 = \tilde{p}_0$ and $P_i = \tilde{p}_i$ for $i = 1, \dots, n$. As one can check, this mapping is actually a group action: we have an action of $\mathbb{R}^n$ (parameterized by the λ's) commuting with an action of $SE(3)$ (parameterized by rotations R and translations T). Therefore, this defines an action of the $(6 + n)$-dimensional Lie group $SE(3) \times \mathbb{R}^n$ on the $(3n + 3)$-dimensional manifold $(\mathbb{R}^3)^{\times(n+1)}$.

We would like to determine where $\mathfrak{P}_0$ and the $\mathcal{O}_i$'s lie. Given a picture, it is of course impossible to determine the camera center and object points $(\mathfrak{P}_0, \mathcal{O}_1, \dots, \mathcal{O}_n)$. However, we know to which orbit under the action of $SE(3) \times \mathbb{R}^n$ they belong, since they belong to the same orbit as the (embedding of the) picture-camera system!

Assuming that the picture points are distinct, then the group action is regular and the orbits are six-dimensional for $n = 1$ and $(6+n)$-dimensional as soon as $n \geq 2$. Therefore, by Theorem 2.3, there are $2n - 3$ fundamental invariants whenever $n \geq 2$ and these invariants can be used to characterize the orbits. We follow the steps of the moving frame normalization method to obtain them. We set

$$\begin{aligned} \bar{P}_0 &= (0, 0, 0)^T, \\ (0, 1, 0)\bar{P}_1 &= 0, \\ (0, 0, 1)\bar{P}_1 &= 0, \\ (0, 0, 1)\bar{P}_2 &= 0, \\ \text{and } (1, 0, 0) \cdot \bar{P}_i &= 1 \text{ for all } i = 1, \dots, n. \end{aligned}$$

[1]The value is arbitrary. It simply fixes the overall scale of the 3D reconstruction.

Solving for the group parameters, we obtain

$$
\begin{aligned}
T &= -RP_0, \\
R &= R_1 R_2 R_3, \\
R_1 &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{f}{\sqrt{f^2+g^2}} & \frac{g}{\sqrt{f^2+g^2}} \\ 0 & -\frac{g}{\sqrt{f^2+g^2}} & \frac{f}{\sqrt{f^2+g^2}} \end{pmatrix}, \\
R_2 &= \begin{pmatrix} \frac{\sqrt{x_1^2+y_1^2}}{\sqrt{x_1^2+y_1^2+z_1^2}} & 0 & \frac{z_1}{\sqrt{x_1^2+y_1^2+z_1^2}} \\ 0 & 1 & 0 \\ \frac{-z_1}{\sqrt{x_1^2+y_1^2+z_1^2}} & 0 & \frac{\sqrt{x_1^2+y_1^2}}{\sqrt{x_1^2+y_1^2+z_1^2}} \end{pmatrix}, \\
R_3 &= \begin{pmatrix} \frac{x_1}{\sqrt{x_1^2+y_1^2}} & \frac{y_1}{\sqrt{x_1^2+y_1^2}} & 0 \\ -\frac{y_1}{\sqrt{x_1^2+y_1^2}} & \frac{x_1}{\sqrt{x_1^2+y_1^2}} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \\
\lambda_i &= \frac{1}{(R(P_i-P_0))_x} - 1,
\end{aligned}
\tag{3.3}
$$

where $f = \frac{-y_1x_2+x_1y_2}{\sqrt{x_1^2+y_1^2}}$, $g = \frac{z_2(x_1^2+y_1^2)-z_1(x_1x_2+y_1y_2)}{\sqrt{x_1^2+y_1^2}\sqrt{x_1^2+y_1^2+z_1^2}}$, $(x_1, y_1, z_1)^T = P_1 - P_0$, and $(x_2, y_2, z_2)^T = P_2 - P_0$. These group parameters define a moving frame MF. Substituting the moving frame into the transformation equations, we get

$$
\begin{aligned}
\bar{P}_0\big|_{MF} &= (0,0,0)^T, \\
\bar{P}_1\big|_{MF} &= (1,0,0)^T, \\
\bar{P}_2\big|_{MF} &= \begin{pmatrix} 1 \\ \frac{f\sqrt{x_1^2+y_1^2+z_1^2}(x_1y_2-x_2y_1)+g[z_2(x_1^2+y_1^2)-z_1(x_1x_2+y_1y_2)]}{(x_1x_2+y_1y_2+z_1z_2)\sqrt{x_1^2+y_1^2}\sqrt{f^2+g^2}} \\ 0 \end{pmatrix}, \\
\bar{P}_i\big|_{MF} &= \begin{pmatrix} 1 \\ \frac{f\sqrt{x_1^2+y_1^2+z_1^2}(x_1y_i-x_iy_1)+g[z_i(x_1^2+y_1^2)-z_1(x_1x_i+y_1y_i)]}{(x_1x_i+y_1y_i+z_1z_i)\sqrt{x_1^2+y_1^2}\sqrt{f^2+g^2}} \\ \frac{g\sqrt{x_1^2+y_1^2+z_1^2}(x_iy_1-x_1y_i)+f[z_i(x_1^2+y_1^2)-z_1(x_1x_i+y_1y_i)]}{(x_1x_i+y_1y_i+z_1z_i)\sqrt{x_1^2+y_1^2}\sqrt{f^2+g^2}} \end{pmatrix}
\end{aligned}
$$

for all $i = 3, \ldots, n$, where $(x_i, y_i, z_i) = P_i - P_0$. Each component of these vectors is an invariant of the group action.

These expressions have an easy geometric interpretation. Since $\sqrt{f^2+g^2} =$

$\frac{\|(x_1,y_1,z_1)\times(x_2,y_2,z_2)\|}{\sqrt{x_1^2+y_1^2+z_1^2}}$, we can rewrite the above system as

$$\begin{aligned}
\bar{P}_0\big|_{MF} &= (0,0,0)^T,\\
\bar{P}_1\big|_{MF} &= (1,0,0)^T,\\
\bar{P}_2\big|_{MF} &= \begin{pmatrix} 1 \\ \frac{\|(x_1,y_1,z_1)\times(x_2,y_2,z_2)\|}{(x_1,y_1,z_1)\cdot(x_2,y_2,z_2)} \\ 0 \end{pmatrix},\\
\bar{P}_i\big|_{MF} &= \begin{pmatrix} 1 \\ \frac{[(x_1,y_1,z_1)\times(x_i,y_i,z_i)]\cdot[(x_1,y_1,z_1)\times(x_2,y_2,z_2)]}{[(x_1,y_1,z_1)\cdot(x_i,y_i,z_i)]\,\|(x_1,y_1,z_1)\times(x_2,y_2,z_2)\|} \\ \frac{(x_i,y_i,z_i)\cdot[(x_2,y_2,z_2)\times(x_1,y_1,z_1)]\|(x_1,y_1,z_1)\|}{[(x_1,y_1,z_1)\cdot(x_i,y_i,z_i)]\,\|(x_1,y_1,z_1)\times(x_2,y_2,z_2)\|} \end{pmatrix},
\end{aligned}$$

where $\cdot$ represents the scalar product between two vectors.

We now see that the components of $\bar{P}_2\big|_{MF}$ and $\bar{P}_i\big|_{MF}$ are sine or cosine of angles between the directions spanned by $\overline{P_1P_0}$, $\overline{P_2P_0}$, $\overline{P_iP_0}$ and the directions orthogonal to them. These are clearly invariant by translation, rotation, and motion along the projection lines. As a fundamental set, we simply pick the only $2n-3$ nonconstant invariants:

$$\begin{aligned}
I_2 &= \frac{\|(x_1,y_1,z_1)\times(x_2,y_2,z_2)\|}{(x_1,y_1,z_1)\cdot(x_2,y_2,z_2)},\\
I_i &= \frac{[(x_1,y_1,z_1)\times(x_i,y_i,z_i)]\cdot[(x_1,y_1,z_1)\times(x_2,y_2,z_2)]}{[(x_1,y_1,z_1)\cdot(x_i,y_i,z_i)]\,\|(x_1,y_1,z_1)\times(x_2,y_2,z_2)\|},\\
J_i &= \frac{(x_i,y_i,z_i)\cdot[(x_2,y_2,z_2)\times(x_1,y_1,z_1)]\|(x_1,y_1,z_1)\|}{[(x_1,y_1,z_1)\cdot(x_i,y_i,z_i)]\,\|(x_1,y_1,z_1)\times(x_2,y_2,z_2)\|}
\end{aligned}$$

for $i=3,\dots,n$.

Each picture taken defines a point in $\mathbb{R}^3\times(\mathbb{R}^3)^{\times(n)}$ and therefore determines an orbit of our group action. Each orbit is characterized by the set of $2n-3$ equations given by the invariants. More precisely, indexing the pictures with the discrete parameter $\tau=1,\dots,t$, we have

$$\begin{aligned}
I_i(P_0^\tau,P_1,\dots,P_n) &= \alpha_i^\tau \text{ for } i=2,\dots,n,\\
J_j(P_0^\tau,P_1,\dots,P_n) &= \beta_j^\tau \text{ for } j=3,\dots,n
\end{aligned}$$

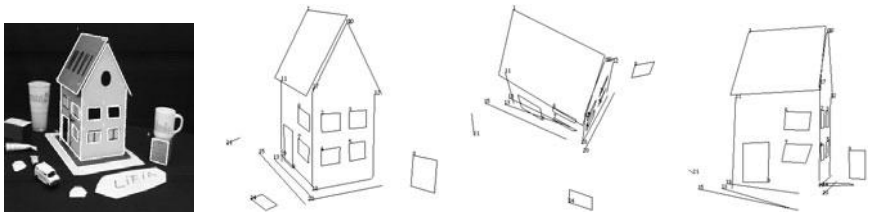
for appropriate constants α_i^τ's and β_j^τ's. These constants are prescribed by the pictures: since the picture-camera system itself belongs to the orbits, we have

$$\begin{aligned}
\alpha_i^\tau &= I_i(\tilde{p}_0^\tau,\tilde{p}_1^\tau,\dots,\tilde{p}_n^\tau),\\
\beta_j^\tau &= J_j(\tilde{p}_0^\tau,\tilde{p}_1^\tau,\dots,\tilde{p}_n^\tau).
\end{aligned}$$

We are interested in solving the equations

$$\begin{aligned}
I_i(\mathfrak{P}_0^\tau,\mathcal{O}_1,\dots,\mathcal{O}_n) &= \alpha_i^\tau \text{ for } i=2,\dots,n,\\
J_j^\tau(\mathfrak{P}_0^\tau,\mathcal{O}_1,\dots,\mathcal{O}_n) &= \beta_j^\tau \text{ for } j=3,\dots,n
\end{aligned}$$

for $\tau=1,\dots,t$. We have a system of $(2n-3)t$ (nonlinear) equations with $3n+3t$ unknowns, the solution of which is determined up to a rotation and translation of the 3D camera-object system as a whole, which we can fix arbitrarily, thus eliminating



a b c d

FIG. 3.1. *A reconstruction example:* (a) *The first of six images, with line and rectangle features drawn;* (b) *a similar view of the* 3*D reconstructed features;* (c) *a top view;* (d) *a side view of the reconstruction.*

six variables.[2] For $n > 3$ and $t \geq \frac{3n-6}{2n-6}$, the number of equations is greater than the number of unknowns, so we can try to solve them.

Experiments with real video images have been performed (see Figure 3.1) using a sequential nonlinear optimization technique based on the Levenberg–Marquardt algorithm [15]. The points used as picture points $(\tilde{p}_0^\tau, \tilde{p}_1^\tau, \ldots, \tilde{p}_n^\tau)$ are the endpoints of lines and rectangles drawn on Figure 3.1(a). These points have been obtained in successive images with a simple tracking procedure [1]. We computed the values of all the invariants $(\alpha_i^\tau, \beta_j^\tau)$ using these points and solved the $(2n-3)t$ nonlinear equations for the unknowns $(\mathfrak{P}_0^\tau, \mathcal{O}_1, \ldots, \mathcal{O}_n)$. The solution gave us the reconstructed 3D object, namely, the set of lines and rectangles defined by $(\mathcal{O}_1, \ldots, \mathcal{O}_n)$. Although the bottom and left elements are not perfectly replaced due to noise in the input picture points, the reconstructed object is visually correct in any view. In particular, there is no global distortion, as one would fear in the case of projective reconstructions. The computations take only a few minutes.

Observe that our camera-system does not take into account the angle of the camera; the orientation of the image plane was only included in the group parameters (not on the space acted on) and thus factored out of the problem in the invariant formulation. Besides the advantage of not having to solve for this unwanted unknown, we also obtain the following lemma.

LEMMA 3.1. *The motion of the camera between two pictures is a pure rotation (i.e., a rotation around the center of projection* P_0*) if and only if the values of the invariants* $\{I_i, J_j | i = 2, \ldots, n,\ j = 3, \ldots, n\}$ *evaluated on any corresponding points in the two views are equal.*

Proof. Invariance of our invariants under pure rotations is obvious from the construction of the invariants.

To prove that equality of our invariants evaluated on all corresponding points guarantees that the camera motion is a pure rotation, observe that the first invariant I_2 is the tangent of the angle between the lines $\overline{\mathfrak{P}_0\mathcal{O}_1}$ and $\overline{\mathfrak{P}_0\mathcal{O}_2}$. Its value remains constant for fixed $\mathcal{O}_1, \mathcal{O}_2$ only if $\mathfrak{P}_0$ moves along a circle around the $\overline{\mathcal{O}_1\mathcal{O}_2}$ axis. This holds for all possible choices of $\mathcal{O}_1$ and $\mathcal{O}_2$, so the camera center must lie somewhere on the intersection of a set of circles, which can be taken to intersect at merely one point to guarantee that the camera center does not move. □

Tomasi and Kanade [19] identify two problems related to using the traditional *direct* approach to structure from motion in a noisy context. First there is the fact

[2]However, we should keep in mind that the choice of these variables will affect the numerical resolution [10].

that, when the camera motion is small, the effects of a camera rotation and translation are hard to distinguish. Second, obtaining the shape by comparing depths is sensitive to noise, since the depth can be considerably larger than the dimensions of the shape. Note that both these difficulties are bypassed by our approach, since the depths and the rotation of the camera do not appear in our equations. We also provide a new formulation in Euclidean space that involves quantities independent of any choice of world coordinate system, removing the so-called *gauge problem*.

Unfortunately, the nonlinearity of the invariants is a serious drawback. As one can tell from the orbit structure of this group action, this is inherent to the problem as formulated. One might want to ask whether there exist coordinates which lead to simpler expressions for the invariants. For example, we could use an inductive moving frame construction [11] in order to obtain invariants of one subgroup of $SE(3) \times \mathbb{R}^n$, say, either $SE(3)$ or $\mathbb{R}^n$, and use these invariants as new coordinates on which the remaining group coordinates are acting. The resulting invariants are actually much simpler in these coordinates. However, it turns out that *each* of these new coordinates would involve the camera center. One would thus need to define a new set of coordinates for *each* picture so there would always be more unknown coordinates than equations. These invariants are thus not useful for recovering the structure from a set of pictures, although they could be good tools for analyzing the motion of an object taken by a fixed or even rotating camera.

Another way of obtaining simpler equations is to work in projective coordinates. However, as a trade-off, projective coordinates require using more variables. We will discuss this other method in the next section, but first we generalize the Euclidean coordinate approach to the case of a variable focal length.

3.1. Letting the focal distance vary. It is a bit more complicated to set up the group transformation equations in the case where the focal length is allowed to vary from one image to the next. One way to do this is the following. Consider P_M, the closest point to P_0 on the image plane, i.e., the embedding of the middle point of the picture. Changing the focal distance corresponds to transforming P_M into a point P'_M with a real parameter α according to the rule

$$P'_M = P_M + \alpha(P_M - P_0).$$

The induced action on the P_i's can be taken as a rotation about the center of camera P_0 which preserves the distance to the camera center. More precisely, each P_i is moved to a new point P'_i in such a way that $\|P'_i - P_0\| = \|P_i - P_0\|$ and that its transformed picture point is $\tilde{p}'_i = \tilde{p}_i + \alpha(P_M - P_0)$. Since the picture point is given by $p_i = P_0 + \frac{P_i - P_0}{(P_i - P_0)\cdot(P_M - P_0)}$, we find that

$$P'_i = P_0 + \|P_i - P_0\| \quad \frac{\frac{P_i - P_0}{(P_i - P_0)\cdot(P_M - P_0)} + \alpha(P_M - P_0)}{\left\| \frac{P_i - P_0}{(P_i - P_0)\cdot(P_M - P_0)} + \alpha(P_M - P_0) \right\|}.$$

Combining translations of the P_i's along the line P_iP_0 together with rotations and translations of the line arrangement as a whole, we get the following $(n+7)$-dimen-

sional Lie group action:

$$\begin{aligned}
\bar{P}_0 &= RP_0 + T,\\
\bar{P}_M &= R\left(P_M + \alpha(P_0 - P_M)\right) + T,\\
\bar{P}_i &= R\left(P_0 + (1+\lambda_i)\|P_i - P_0\| \frac{\frac{P_i - P_0}{(P_i - P_0)\cdot(P_M - P_0)} + \alpha(P_M - P_0)}{\left\|\frac{P_i - P_0}{(P_i - P_0)\cdot(P_M - P_0)} + \alpha(P_M - P_0)\right\|}\right) + T.
\end{aligned}$$

Observe that the change of focal length parameterized by α commutes with the action of each λ_i on P_i because it preserves the norm $\|P_i - P_0\|$.

We apply the previous moving frame normalization technique by setting

$$\begin{aligned}
\bar{P}_0 &= (0,0,0)^T,\\
\bar{P}_M &= (1,0,0)^T,\\
\bar{P}_1 \cdot (0,0,1) &= 0,\\
\text{and } \bar{P}_i \cdot (1,0,0) &= 1 \text{ for all } i = 1,\dots,n.
\end{aligned}$$

The corresponding group parameters are similar to (3.3) for R and T, except that $(x_1, y_1, z_1)^T$ and $(x_2, y_2, z_2)^T$ must be replaced by $(u_1, v_1, w_1)^T = P_M - P_0$ and $(u_2, v_2, w_2)^T = P_1 - P_0$. The other parameters are

$$\alpha = \frac{1}{\|P_M - P_0\|} - 1,$$

$$\lambda_i = \frac{\left\|\frac{P_i - P_0}{(P_i - P_0)\cdot(P_M - P_0)} + \alpha(P_M - P_0)\right\|}{\|P_i - P_0\|\|P_M - P_0\| + \frac{\|P_i - P_0\|}{\|P_M - P_0\|}} - 1 \text{ for all } i = 1,\dots,n.$$

Substituting these group parameters into the equations for the $\bar{P}_i$'s, we obtain the following complete fundamental set of invariants:

$$I_1 = \frac{\|(P_M - P_0)\times(P_1 - P_0)\|}{(P_1 - P_0)\cdot(P_M - P_0)\left(1 + \|P_M - P_0\| - \|P_M - P_0\|^2\right)},$$

$$I_i = \frac{(P_M - P_0)\times(P_i - P_0)\cdot(P_M - P_0)\times(P_1 - P_0)}{\|(P_M - P_0)\times(P_1 - P_0)\|(P_i - P_0)\cdot(P_M - P_0)\left(1 + \|P_M - P_0\| - \|P_M - P_0\|^2\right)},$$

$$J_i = \frac{(P_i - P_0)\cdot\left[(P_1 - P_0)\times(P_M - P_0)\right]\|P_M - P_0\|}{\|(P_M - P_0)\times(P_1 - P_0)\|(P_i - P_0)\cdot(P_M - P_0)\left(1 + \|P_M - P_0\| - \|P_M - P_0\|^2\right)}$$

for $i = 2,\dots,n$.

Observe that solving for P_M and P_0 implies solving for the focal length. Therefore, the focal length has not been removed in this formulation. Having to solve for the focal length is undesirable since it can induce numerical instabilities. We can actually completely include the focal length in the group parameters by letting $v = \frac{P_M - P_0}{\|P_M - P_0\|}$ and $\gamma = \alpha\|P_M - P_0\|^2$. Then γ can be seen as a new group parameter $\gamma \in \mathbb{R}_{\neq -1}$ in the group action given by

$$\begin{aligned}
\bar{P}_0 &= RP_0 + T,\\
\bar{v} &= Rv,\\
\bar{P}_i &= R\left(P_0 + (1+\lambda_i)\|P_i - P_0\| \frac{P_i - P_0 + \left(\gamma(P_i - P_0)\cdot v\right)v}{\|P_i - P_0 + \left(\gamma(P_i - P_0)\cdot v\right)v\|}\right) + T.
\end{aligned}$$

One can check that the group action parameterized by γ is compatible with the group structure of $\mathbb{R}_{\neq -1}$ with group multiplication $\circ$ given by

$$\gamma_1 \circ \gamma_2 = \gamma_1 + \gamma_2 + \gamma_1\gamma_2 \text{ for all } \gamma_1, \gamma_2 \in \mathbb{R}_{\neq -1}.$$

We have an action of the $(7+n)$-dimensional Lie group $SE(3) \times \mathbb{R}_{\neq -1} \times \mathbb{R}^n$ on a $(3n+5)$-dimensional space; there are $2n-2$ fundamental invariants. To obtain these invariants, we start by setting $\bar{P}_0 = (0,0,0)^T$ and solve for T. We then set $\bar{v} = (1,0,0)^T$ and the third component of $\bar{P}_1$ to zero and solve for the rotation matrix R. We skip the details of these computations since they are very similar to the previous cases. Substituting these group parameters into the other transformation equations, we obtain

$$\bar{P}_i = (1+\lambda_i)\|P_i - P_0\| \begin{pmatrix} F_i \cdot v \\ \frac{(v \times F_i)\cdot(v \times F_1)}{\|v \times F_1\|} \\ \frac{F_i \cdot (v \times F_1)}{\|v \times F_1\|} \end{pmatrix} \text{ for } i = 2, \ldots, n,$$

$$\bar{P}_1 = (1+\lambda_1)\|P_1 - P_0\| \begin{pmatrix} F_1 \cdot v \\ \|v \times F_1\| \\ 0 \end{pmatrix},$$

where F_i represents the fraction

$$F_i = \frac{P_i - P_0 + (\gamma(P_i - P_0)\cdot v)v}{\|P_i - P_0 + (\gamma(P_i - P_0)\cdot v)v\|} \text{ for } i = 1, \ldots, n.$$

We then set the first component of each $\bar{P}_i$ to one for $i = 1, \ldots, n$ and solve for the λ_i's. (For this, we need to assume that $(P_i - P_0)\cdot v \neq 0$.) Substituting these λ_i's into the transformation equations, we get

$$\bar{P}_i = \begin{pmatrix} 1 \\ \frac{(v \times F_i)\cdot(v \times F_1)}{\|v \times F_1\| F_i \cdot v} \\ \frac{F_i \cdot (v \times F_1)}{\|v \times F_1\| F_i \cdot v} \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{(v \times (P_i - P_0))\cdot(v \times (P_1 - P_0))}{\|v \times (P_1 - P_0)\|(1+\gamma)(P_i - P_0)\cdot v} \\ \frac{(P_i - P_0)\cdot(v \times (P_1 - P_0))}{\|(v \times (P_1 - P_0))\|(1+\gamma)(P_i - P_0)\cdot v} \end{pmatrix} \text{ for } i = 2, \ldots, n,$$

$$\bar{P}_1 = \begin{pmatrix} 1 \\ \frac{\|v \times F_1\|}{F_1 \cdot v} \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{\|v \times (P_1 - P_0)\|}{(1+\gamma)(P_1 - P_0)\cdot v} \\ 0 \end{pmatrix}.$$

Provided that $P_1 - P_0$ is not parallel to v, we can finish obtaining the moving frame by setting the second component of $\bar{P}_1$ to one and solving for γ. Substituting this γ into the two nonconstant components of each P_i, we obtain the following two invariants:

$$I_i = \frac{v \cdot (P_1 - P_0)(v \times (P_i - P_0)) \cdot (v \times (P_1 - P_0))}{v \cdot (P_i - P_0)\|v \times (P_1 - P_0)\|^2},$$

$$J_i = \frac{v \cdot (P_1 - P_0)(P_i - P_0) \cdot (v \times (P_1 - P_0))}{v \cdot (P_i - P_0)\|v \times (P_1 - P_0)\|^2}$$

for $i = 2, \ldots, n$. In a similar way as for the case of a fixed focal length, when the number of pictures t and the number of object points n are large enough, these invariants provide a number of equations which are, in most cases, sufficient to solve for $P_1, \ldots, P_n$ and P_0^τ, v^τ for $\tau = 1, \ldots, t$.

4. Using projective coordinates. The projective space $\mathbb{P}^3$ is the set $\{(x,y,z,w) \in \mathbb{R}^4 \setminus \{(0,0,0,0)\}\}$ modulo multiplication by a scalar multiple in $\mathbb{R}\setminus\{0\}$. If $(x,y,z,w) \in \{\mathbb{R}^4 \setminus \{(0,0,0,0)\}\}$, then the coset of (x,y,z,w) in $\mathbb{P}^3$ is denoted by $(x : y : z : w)$. Consider the chart U of $\mathbb{P}^3$ defined by $\{(x : y : z : w) \in \mathbb{P}^3 | w \neq 0\}$. The map $\phi : U \to \mathbb{R}^3$ defined by

$$\phi(x : y : z : w) = \left(\frac{x}{w}, \frac{y}{w}, \frac{z}{w}\right)$$

provides coordinates for the chart U. In other words, Euclidean coordinates of $\mathbb{R}^3$ are local coordinates for a piece of $\mathbb{P}^3$. One way to obtain simpler invariants is to work directly in $\mathbb{R}^4\setminus\{(0,0,0,0)\}$ (i.e., in projective coordinates) by choosing representatives (x_0,y_0,z_0,w_0) and (x_i,y_i,z_i,w_i) in $\{(x,y,z,w) \in \mathbb{R}^4 | w \neq 0\}$ for the camera center and the object points, respectively.

Consider the following action of $SE(3) \times \mathbb{R}^n$ on $n+1$ copies of $\mathbb{R}^4$:

$$\begin{pmatrix} \bar{x}_0 \\ \bar{y}_0 \\ \bar{z}_0 \\ \bar{w}_0 \end{pmatrix} = \begin{pmatrix} & R & & T \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{x}_0 \\ \bar{y}_0 \\ \bar{z}_0 \\ \bar{w}_0 \end{pmatrix},$$

$$\begin{pmatrix} \bar{x}_i \\ \bar{y}_i \\ \bar{z}_i \\ \bar{w}_i \end{pmatrix} = \begin{pmatrix} & R & & T \\ 0 & 0 & 0 & 1 \end{pmatrix} \left[\begin{pmatrix} x_i \\ y_i \\ z_i \\ w_i \end{pmatrix} + \lambda_i \left(\begin{pmatrix} x_i \\ y_i \\ z_i \\ w_i \end{pmatrix} - \begin{pmatrix} x_0 \\ y_0 \\ z_0 \\ w_0 \end{pmatrix} \right) \right]$$
$$\text{for } i = 1, \ldots, n,$$

where $R \in SO(3)$ is a 3×3 rotation matrix, $T \in \mathbb{R}^3$ represents a translation, and the $\lambda_i \in \mathbb{R}$ are the depth parameters. Since lines through the origin are mapped to lines through the origin, this induces an action on $\mathbb{P}^3$. In local coordinates ϕ for the chart U, this corresponds exactly to the action defined by (3.1) and (3.2). Assuming that the object points are pairwise distinct (and, of course, also distinct from the camera center), then, as soon as the number of object points $n \geq 2$, the orbits are $(6+n)$-dimensional both on $(\mathbb{R}^4)^{\times(n+1)}$ and on $(\mathbb{P}^3)^{\times(n+1)}$. There are therefore $3n-2$ fundamental invariants in projective coordinates.

To obtain a fundamental set of invariants, we start by setting $\omega_i = 1$ for all i's, and $\bar{P}_0 = (0,0,0)^T$. Provided that all $w_i \neq w_0$, the corresponding group parameters are $\lambda_i = \frac{1-w_i}{w_i-w_0}$ and $T = -\frac{1}{w_0}R$. For simplicity, we let $\beta_i = \frac{1-w_i}{w_i-w_0}$. We then set the second and third components of $\bar{P}_1$ to zero and the third component of $\bar{P}_2$ to zero and solve for the rotation matrix R to finish obtaining the moving frame. Substituting the moving frame into the transformation equations, we obtain

$$\begin{aligned}
(\bar{x}_0, \bar{y}_0, \bar{z}_0, \bar{w}_0)|_{MF} &= (0,0,0,w_0), \\
(\bar{x}_1, \bar{y}_1, \bar{z}_1, \bar{w}_1)|_{MF} &= (\|Q_1\|, 0, 0, 1), \\
(\bar{x}_2, \bar{y}_2, \bar{z}_2, \bar{w}_2)|_{MF} &= \left(\frac{Q_2 \cdot Q_1}{\|Q_1\|}, \frac{\|Q_2 \times Q_1\|}{\|Q_1\|}, 0, 1\right), \\
(\bar{x}_i, \bar{y}_i, \bar{z}_i, \bar{w}_i)|_{MF} &= \begin{pmatrix} \frac{Q_i \cdot Q_1}{\|Q_1\|} \\ \frac{(Q_1 \times Q_i) \cdot (Q_1 \times Q_2)}{\|Q_1\| \|Q_1 \times Q_2\|} \\ \frac{Q_i \cdot (Q_1 \times Q_2)}{\|Q_1 \times Q_2\|} \\ 1 \end{pmatrix}^T \\
&\text{for } i = 3, \ldots, n,
\end{aligned}$$

where $\cdot$ represents the scalar product between two vectors and $Q_i = (1+\beta_i)P_i - (\beta_i + \frac{1}{w_0})P_0$. As a complete set of fundamental invariants, we can take, in addition to the coordinate w_0, the functions

$$H_i = Q_i \cdot Q_1 \text{ for } i = 1, \dots, n,$$
$$I_i = (Q_1 \times Q_i) \cdot (Q_1 \cdot Q_2) \text{ for } i = 2, \dots, n,$$
$$J_i = Q_i \cdot (Q_1 \times Q_2) \text{ for } i = 3, \dots, n.$$

The invariants are thus functions of the object points P_i's, the β_i's, and the camera center projective coordinates P_0 and w_0. For every picture, the parameter w_0 can be fixed arbitrarily to any value other than zero. Given t pictures, the unknowns are thus the P_i's, β_i^τ's, and P_0^τ for $i = 1, \dots, n$ and $\tau = 1, \dots, n$, while the invariants provide $t(3n-3)$ equations. With enough points and enough pictures, we obtain more equations than unknowns.

5. The case of an orthographic camera. The orthographic camera is an approximation of the perspective camera. In this model, we assume that the camera center lies at infinity, and so the rays of light are parallel to each other.

Let $v = (v_x, v_y, v_z)^T$ be the unit direction vector of the rays of light and let $P_1, \dots, P_n$ represent the object points in $\mathbb{R}^3$. Any picture of the object provides some information about the structure of the object. What remains unknown is the orientation and the position of the camera at the moment when the picture was taken, as well as the distance from the camera plane to each object point. The following action of $SE(3) \times \mathbb{R}^n$ on $\{(v, P_1, \dots, P_n) \in (\mathbb{R}^3)^{\times(n+1)}$ such that $|v| = 1\}$ summarizes what is unknown about the object given a picture:

$$\bar{v} = Rv,$$
$$\bar{P}_i = R(P_i + \lambda_i v) + T \text{ for } i = 1, \dots, n,$$

where $R \in SO(3)$ is a rotation matrix, $T \in \mathbb{R}^3$ represents a translation, and the real numbers $\lambda_1, \dots, \lambda_n$ are the depth parameters. More precisely, given a picture $p_1, \dots, p_n$, the orbit passing through

$$P_i = (p_i, 0) \text{ for } i = 1, \dots, n,$$
$$v = (0, 0, 1)$$

(i.e., the embedding of the picture in $\mathbb{R}^3$) under this group action corresponds to all possible 3D objects that could have been used to take this picture.

To obtain the invariants of this group action, we set $\bar{P}_1 = (0, 0, 0)^T$, the first component of each $\bar{P}_i$ to zero, the second and third components of $\bar{P}_1$ to zero, and the third component of $\bar{P}_1$ to zero. We use the partial moving frame normalization method and, in order to obtain a moving frame, solve for all parameters except λ_1, which does not appear in the final expressions. Substituting this moving frame in the

group transformation equations, we obtain

$$\begin{aligned}
\bar{v}\big|_{MF} &= (1,0,0)^T,\\
\bar{P}_1\big|_{MF} &= (0,0,0)^T,\\
\bar{P}_2\big|_{MF} &= \begin{pmatrix} 1 \\ \|(P_2-P_1)-[(P_2-P_1)\cdot v]v\| \\ 0 \end{pmatrix},\\
\bar{P}_i\big|_{MF} &= \begin{pmatrix} 1 \\ (P_i-P_1)\cdot \frac{(P_2-P_1)-[(P_2-P_1)\cdot v]v}{\|(P_2-P_1)-[(P_2-P_1)\cdot v]v\|} \\ (P_i-P_1)\cdot\left(v\times \frac{(P_2-P_1)-[(P_2-P_1)\cdot v]v}{\|(P_2-P_1)-[(P_2-P_1)\cdot v]v\|}\right) \end{pmatrix} \text{ for } i=1,\dots,n.
\end{aligned}$$

A fundamental set of invariants is given by the nonconstant components of these vectors. Observe that some of these expressions are fractions. However, their denominator is actually one of the invariants of the fundamental set. We can thus simply get rid of the denominator and take the following functions as our fundamental set of invariants:

$$\begin{aligned}
I_2 &= \|(P_2-P_1)-[(P_2-P_1)\cdot v]v\|,\\
I_i &= (P_i-P_1)\cdot[(P_2-P_1)-[(P_2-P_1)\cdot v]v] \text{ for } i=3,\dots,n,\\
J_i &= (P_i-P_1)\cdot[v\times(P_2-P_1)] \text{ for } i=3,\dots,n.
\end{aligned}$$

Given pictures $p_1^\tau,\dots,p_n^\tau\in\mathbb{R}^2$, we let v^τ for $\tau=1,\dots,n$ be the direction vectors of the rays of light of the camera. The object points $P_1,\dots,P_n$ and direction vectors v^τ thus satisfy the matrix equation

$$\begin{pmatrix} (P_2-P_1)-[(P_2-P_1)\cdot v^\tau]v^\tau \\ v^\tau\times(P_2-P_1) \end{pmatrix}\begin{pmatrix} P_2-P_1, & P_3-P_1, & \dots, & P_n-P_1 \end{pmatrix} = \begin{pmatrix} \alpha_2^\tau, & \alpha_3^\tau, & \dots, & \alpha_n^\tau \\ \beta_2^\tau, & \beta_3^\tau, & \dots, & \beta_n^\tau \end{pmatrix},$$

where the α's and β's are constants prescribed by the pictures. For solving these equations, we can make a change of variable and let the two entries of the leftmost matrix be two unknown parameters m_1^τ and m_2^τ subject to the condition $|m_1^\tau|+|m_2^\tau|\le|P_2-P_1|$. We obtain a factorization equation, like the one introduced by Tomasi and Kanade [19], with a different formulation. Note that our system involves only the normal to the camera plane, i.e., two parameters, while theirs involves all three parameters specifying the orientation of the camera.

6. Conclusion. This paper presented applications of a systematic technique invented by Fels and Olver for building invariants of a Lie group action. We started by summarizing this technique. We then showed how to formulate the problem of structure from motion in three different settings (Euclidean coordinates, projective coordinates, and orthographic projections) in terms of Lie group actions. In each setting, the group parameters included unknown unwanted parameters of the problem. These parameters were removed from the equations by reformulating the problem using invariants of these group actions.

The orbit structure and the invariants of the group action provide interesting insights on the geometry of the projections. They also provide a formalization for similar results obtained more empirically [17]. For solving the structure from motion

problem, our results relate to several well-known techniques but eliminate additional unnecessary parameters, sometimes difficult to remove otherwise. Further work is now ongoing to make such invariant systems more computationally attractive for various practical applications.

Acknowledgments. Both authors would like to thank David Cooper for support, encouragement, and stimulating discussions, as well as David Mumford and Jean Ponce for their useful comments.

REFERENCES

[1] P. L. Bazin and J. M. Vézien, *Tracking geometric primitives in video streams*, in Proceedings of the Fourth Irish Machine Vision and Image Processing Conference, Belfast, 2000, pp. 43–50.

[2] M. Boutin, *On orbit dimensions under a simultaneous Lie group action on n copies of a manifold*, J. Lie Theory, 12 (2002), pp. 191–203.

[3] J. B. Burns, R. S. Weiss, and E. M. Riseman, *The non-existence of general-case view-invariants*, in Geometric Invariance in Computer Vision, MIT Press, Cambridge, MA, 1992, pp. 120–131.

[4] É. Cartan, *Leçons sur la géométrie projective complexe. La théorie des groupes finis et continus et la géométrie différentielle traitées par la méthode du repère mobile. Leçons sur la théorie des espaces à connexion projective*, Les Grands Classiques Gauthier-Villars [Gauthier-Villars Great Classics], Éditions Jacques Gabay, Sceaux, 1992. Reprint of the editions of 1931 and 1937.

[5] M. Fels and P. J. Olver, *Moving coframes.* I. *A practical algorithm*, Acta Appl. Math., 51 (1998), pp. 161–213.

[6] M. Fels and P. J. Olver, *Moving coframes.* II. *Regularization and theoretical foundations*, Acta Appl. Math., 55 (1999), pp. 127–208.

[7] G. Frobenius, *Über das Pfaff'sche problem*, J. Reine Angew. Math., 82 (1877), pp. 230–315.

[8] V. V. Gorbatsevich, A. L. Onishchik, and E. B. Vinberg, *Foundations of Lie Theory and Lie Transformation Groups*, Springer-Verlag, Berlin, 1997.

[9] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, UK, 2001.

[10] K. Kanatani, *Gauge-based reliability analysis of* 3*-d reconstruction from two uncalibrated perspective views*, in Proceedings of the Fifteenth International Conference on Pattern Recognition, Barcelona, 2000, pp. 76–79.

[11] I. A. Kogan, *Inductive construction of moving frames*, in The Geometrical Study of Differential Equations (Washington, DC, 2000), Contemp. Math. 285, AMS, Providence, RI, 2001, pp. 157–170.

[12] J. L. Mundy and A. Zisserman, eds., *Geometric Invariance in Computer Vision*, Artificial Intelligence, MIT Press, Cambridge, MA, 1992.

[13] J. L. Mundy, A. Zisserman, and D. Forsyth, eds., *Workshop on Applications of Invariance in Computer Vision*, Lecture Notes in Comput. Sci. 825, Springer-Verlag, New York, 1994.

[14] P. J. Olver, *Classical Invariant Theory*, London Math. Soc. Stud. Texts 44, Cambridge University Press, Cambridge, UK, 1999.

[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed., Cambridge University Press, Cambridge, UK, 1992.

[16] L. Quan, *Invariants of six points and projective reconstruction from three uncalibrated images*, IEEE Trans. Pattern Anal. Mach. Intell., 17 (1995), pp. 34–46.

[17] C. Rother and S. Carlsson, *Linear multi view reconstruction and camera recovery*, in Proceedings of the International Conference on Computer Vision, Vancouver, Canada, IEEE, Los Alamitos, CA, 2001, pp. 42–51.

[18] G. Sparr, *Euclidean and affine structure/motion for uncalibrated cameras from affine shape and subsidiary information*, in Proceedings of the SMILE Workshop, Freiburg, Germany, 1998, pp. 187–207.

[19] C. Tomasi and T. Kanade, *Shape and motion from image streams under orthography: A factorization method*, Int. J. Comput. Vision, 9 (1992), pp. 137–154.

[20] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, *Bundle adjustment: A modern synthesis*, in Proceedings of Vision Algorithms: Theory and Practice, Corfu, Greece, 1999, pp. 278–294.

[21] I. Weiss, 3-*D curve reconstruction from uncalibrated cameras*, in Proceedings of the International Conference on Pattern Recognition (ICPR), Vol. I, IEEE, Los Alamitos, CA, 1996, pp. 323–327.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1175–1185

CONGESTION REDUX*

J. M. GREENBERG†

Abstract. In this paper we analyze a class of second-order traffic models and show that these models support stable oscillatory traveling waves typical of the waves observed on a congested roadway. The basic model has trivial or constant solutions where cars are uniformly spaced and travel at a constant equilibrium velocity that is determined by the car spacing. The stable traveling waves arise because there is an interval of car spacing for which the constant solutions are unstable. These waves consist of a smooth part where both the velocity and spacing between successive cars are increasing functions of a Lagrange mass index. These smooth portions are separated by shock waves that travel at computable negative velocity.

Key words. conservation laws, traffic congestion, follow-the-leader

AMS subject classification. 35

DOI. 10.1137/S0036139903431737

1. Introduction. In the last several years a number of authors [1, 2, 3, 4, 5, 6, 7, 8, 9] have advanced "higher order" traffic models in an attempt to characterize strong permanent waves which appear in congested traffic. At the continuum level all of these authors have worked with models of the following form:

$$\frac{\partial s}{\partial t} - \frac{\partial u}{\partial m} = 0 \tag{1.1}$$

and

$$\epsilon \frac{\partial u}{\partial t} = \epsilon P'(s) \frac{\partial u}{\partial m} + V(s) - u. \tag{1.2}$$

Here $t \geq 0$ is time, m is a Lagrangian mass coordinate which gives the car index, and $\epsilon > 0$ has the interpretation of a relaxation time. The velocity of the mth car at time t is $u(m,t)$, and $s(m,t) \geq L > 0$ is a measure of the spacing between successive cars. Finally, the function $s \to V(s)$ has the interpretation of an "equilibrium" velocity, and the term $\epsilon P'(s)\frac{\partial u}{\partial m}$ appearing in (1.2) is typically referred to as the anticipatory acceleration. All authors assume that $P'(s) \geq 0$ on $s \geq L$. The parameter $L > 0$ has the interpretation of the length of a car on the roadway.

The trajectory of the mth car is given as the solution of

$$\frac{\partial x}{\partial t} = u \quad \text{and} \quad x(m,0) = x_0(m), \tag{1.3}$$

where $x_0(m)$ is the position of the mth car at $t = 0$. $s(m,t)$ is related to $x(m,t)$ by

$$s(m,t) = \frac{\partial x}{\partial m}(m,t) \tag{1.4}$$

and measures the spacing between successive cars.

*Received by the editors July 21, 2003; accepted for publication (in revised form) October 31, 2003; published electronically April 21, 2004.
http://www.siam.org/journals/siap/64-4/43173.html
†U.S. Office of Naval Research, International Field Office, 223 Old Marylebone Road, London NW1 5TH, UK and Carnegie Mellon University, Department of Mathematical Sciences, Pittsburgh, PA 15213 (jgreenberg@onrglobal.navy.mil).

The hypothesis that $P'(s) \geq 0$ implies that the system (1.1) and (1.2) is hyperbolic with wave speeds $c = -P'(s) \leq 0$ and $c = 0$ and thus information propagates from right to left. This observation implies that when constructing finite difference schemes for (1.1) and (1.2) the appropriate spatial differences should be downwind, i.e., that

$$s(m,t) \doteq \frac{x(m+\Delta m,t) - x(m,t)}{\Delta m} \tag{1.5}$$

and

$$\frac{\partial u}{\partial m}(m,t) \doteq \frac{u(m+\Delta m,t) - u(m,t)}{\Delta m}. \tag{1.6}$$

If one chooses to discretize (1.1)–(1.4) spatially, keep time continuous, and, moreover, choose $\Delta m = 1$ (recalling that cars are really discrete entities), one is led to the classic follow-the-leader system

$$\frac{dx_m}{dt} = u_m \tag{1.7}$$

and

$$\epsilon \frac{du_m}{dt} = \epsilon P'(x_{m+1} - x_m)(u_{m+1} - u_m) + V(x_{m+1} - x_m) - u_m \tag{1.8}$$

studied by traffic engineers. On the other hand, if one lets

$$\rho(x,t) = \frac{1}{s(m,t)} \quad \text{and} \quad v(x,t) = u(m,t) \tag{1.9}$$

when

$$x = x(m,t), \tag{1.10}$$

one finds that as functions of x and t the functions ρ and v satisfy

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v) = 0 \tag{1.11}$$

and

$$\epsilon \left(\frac{\partial v}{\partial t} + (v + \rho \mathcal{R}_{,\rho}(\rho)) \frac{\partial v}{\partial x} \right) = W(\rho) - v, \tag{1.12}$$

where

$$\mathcal{R}(\rho) \stackrel{def}{=} P(1/\rho) \quad \text{and} \quad W(\rho) \stackrel{def}{=} V(1/\rho). \tag{1.13}$$

Of course

$$\rho^2 \mathcal{R}_{,\rho}(\rho) = -P'(s = 1/\rho) \leq 0 \quad \text{and} \quad \rho^2 W_{,\rho}(\rho) = -V'(s = 1/\rho) \leq 0. \tag{1.14}$$

References [2, 3, 4, 5] dealt primarily with the case where $P(\cdot)$ and $V(\cdot)$ were monotone increasing on $s \geq L$ and had the following additional properties:

$$V(L^+) = 0 \quad \text{and} \lim_{s \to \infty} V(s) = v_\infty \tag{1.15}$$

and

$$0 < V'(s) \le P'(s),\ P''(s) < 0 \quad \text{and} \quad V''(s) < 0,\ L < s < \infty. \tag{1.16}$$

In that series of papers the authors established a variety of results about the system (1.1)–(1.4) and its discrete counterpart (1.7)–(1.8), notably that the constant solutions

$$s(m,t) \equiv s_0 > L \quad \text{and} \quad u(m,t) \equiv V(s_0) \tag{1.17}$$

were stable in the L_∞ norm.

Bando et al. [6] considered the discrete system (1.7)–(1.8) when $P'(\cdot) \equiv 0$ and found that the steady solutions (1.7)–(1.8) were linearly unstable if $0 < V'(s_0)$ was large enough and linearly stable otherwise. The continuous system (1.1)–(1.2) with $P'(\cdot) \equiv 0$ supports a stronger conclusion, namely, that all steady solutions are linearly unstable and this is a defect in that model. In that same paper, Bando and his coauthors also exhibited large amplitude oscillatory solutions to (1.7) and (1.8) in the case where

$$V(s) = \frac{v_\infty \left(\tanh\left(\frac{s-rL}{\delta}\right) + \tanh\left(\frac{(r-1)L}{\delta}\right)\right)}{\left(1 + \tanh\left(\frac{(r-1)L}{\delta}\right)\right)}, s \ge L \tag{1.18}$$

when

$$r > 1, \quad v_\infty > 0, \quad \text{and} \quad \delta > 0. \tag{1.19}$$

These solutions are reminiscent of the strong permanent waves seen in congested traffic. The authors' calculations gave no indication of the propagation speeds of these waves.

Finally, Greenberg, Klar, and Rascle [7] considered the system (1.1)–(1.2) when

$$P(s) = v_\infty \left(1 - L/s\right), \quad 0 < L \le s \tag{1.20}$$

and

$$V(s) = \begin{cases} \mu v_\infty \left(1 - L/s\right), & L \le s < s_* \\ \\ v_\infty \left(1 - L/s\right), & s_* < s < \infty, \end{cases} \tag{1.21}$$

where $0 < \mu < 1$ and $v_\infty > 0$. The trivial equilibria for this model are

$$s = s_0 \quad \text{and} \quad u \equiv V(s_0) \quad \text{when} \quad s_0 \ne s_* \tag{1.22}$$

and

$$s = s_* \quad \text{and any} \quad u = u_* \text{ in the interval } \left(\mu v_\infty \left(1 - L/s_*\right),\ v_\infty \left(1 - L/s_*\right)\right). \tag{1.23}$$

The former equilibria are stable and the latter unstable. In [7] the authors established the existence of stable periodic traveling waves (the ring-road scenario) of large amplitude which propagate with speed $c = -P'(s_*)$. These waves were functions of $\xi = m - ct$ and were composed of a smooth increasing portion satisfying

$$s(-m_a^+) = s_a, \quad s(0) = s_*, \quad \text{and} \quad s(M_a^-) = S_a. \tag{1.24}$$

The numbers $s_a < s_* < S_a$ were not arbitrary. They satisfied

$$\frac{P(S_a) - P(s_a)}{S_a - s_a} = P'(s_*), \tag{1.25}$$

and the numbers m_a and M_a satisfied

$$k\,(m_a + M_a) = M \quad \text{and} \quad k\int_{m_a}^{M_a} s(\xi)d\xi = l. \tag{1.26}$$

Here M represents the number of cars on the ring-road, l is the length of the ring-road, and $k \geq 1$ is an integer which gives the number of increasing segments per period. These waves have jump discontinuities at the points $\{m_a \pm n\,(m_a + M_a)\}_{n=0}^{\infty}$ and (1.25) guarantees that the Rankine–Hugoniot conditions for (1.1)–(1.2) hold across the discontinuities. These waves also satisfy the Lax entropy condition across the shocks, namely, the condition that $S_a > s_a$.

Our goal in the remainder of this paper is to show that the results of [7] were no fluke; that is, they were not an artifact of the jump discontinuity in the equilibrium velocity function defined in (1.21) but rather were generic. In the remainder of this paper we shall limit ourselves to the analysis of (1.1)–(1.2) when $P(\cdot)$ and $V(\cdot)$ are both increasing on $[L, \infty)$ and satisfy the normalization conditions

$$P(L^+) = V(L^+) = 0 \quad \text{and} \quad \lim_{s\to\infty} V(s) = v_\infty > 0. \tag{1.27}$$

We shall assume that $V'(\cdot)$ has an isolated single maximum at $s_* > L$, that

$$V''(s) > 0, \quad L \leq s < s_* \quad \text{and } V''(s) < 0, \quad s_* < s < \infty, \tag{1.28}$$

that the difference $(P' - V')\,(\cdot)$ has two isolated zeros at points s_1 and s_2 satisfying $L < s_1 < s_* < s_2 < \infty$, and, finally, that $(P' - V')(\cdot) > 0$ on $(L, s_1) \cup (s_2, \infty)$.

In section 2 we shall give a simple argument showing that for s_0 in (s_1, s_2), the constant solution defined in (1.17) is unstable. We shall also show that if the initial data for s lies in this interval, then s approximately evolves via a convective backwards heat equation, thus confirming the instability of the constant solutions. This latter result will be established by using a Chapman–Enskog expansion of the solutions of (1.1) and (1.2). In section 3 we shall show how to construct the large amplitude periodic traveling wave solutions to (1.1)–(1.2) reminiscent of the waves seen in congested traffic. These solutions are similar in structure to those obtained in [7]. Section 4 will be devoted to numerical simulations. Here we shall limit ourselves to

$$P(s) = \lambda(1 - L/s), \quad L \leq s, \tag{1.29}$$

and $V(\cdot)$ given by (1.18). We shall demonstrate that for nonconstant initial data taking on values in the unstable interval (s_1, s_2), solutions converge to traveling waves. These simulations will be run on the follow-the-leader model (1.7)–(1.8). Comprehensive surveys on this vast subject may be found in Helbing [8] and Nagel, Wagner, and Woesler [9].

2. Linear stability of (1.17). We look for solutions of (1.1)–(1.2) of the form

$$s = s_0 + \delta_1 A \quad \text{and} \quad u = V(s_0) + \delta_1 W, \tag{2.1}$$

where $0 < \delta_1 \ll 1$. To leading order in δ_1 we find that A satisfies

$$\epsilon\left(\frac{\partial^2 A}{\partial t^2} - P'(s_0)\frac{\partial^2 A}{\partial t \partial m}\right) = V'(s_0)\frac{\partial A}{\partial m} - \frac{\partial A}{\partial t}. \tag{2.2}$$

If we look for solutions of (2.2) of the form

$$A = \exp(ikm + \lambda t), \tag{2.3}$$

we find that λ and k satisfy

$$\epsilon\lambda^2 + (1 - ik\epsilon P'(s_0))\,\lambda - ikV'(s_0) = 0. \tag{2.4}$$

Moreover, if we write $\lambda = \alpha + i\beta$ (with α and β real), we obtain

$$\epsilon(\alpha^2 - \beta^2) + \alpha + k\epsilon P'(s_0)\beta = 0 \tag{2.5}$$

and

$$2\epsilon\alpha\beta + \beta - k\epsilon P'(s_0)\alpha - kV'(s_0) = 0. \tag{2.6}$$

If we restrict our attention to the case where $0 < \epsilon \ll 1$, we find one root goes as

$$\lambda_1 = -\frac{1}{\epsilon} + ik(P' - V')(s_0) + 0(\epsilon) \tag{2.7}$$

and the other as

$$\lambda_2 = ikV'(s_0) - \epsilon k^2 V'(s_0)(P' - V')(s_0) + 0(\epsilon^2) \tag{2.8}$$

and it is the latter identity which allows us to conclude that the system is linearly stable when $(P' - V')(s_0) > 0$ and linearly unstable when $(P' - V')(s_0) < 0$.

A similar conclusion may be reached if we apply a Chapman–Enskog procedure to (1.1) and (1.2) when $0 < \epsilon \ll 1$. Specifically, we seek solutions to (1.1) and (1.2) where u is of the form

$$u = U^1 = u^0 + \epsilon u^1 \tag{2.9}$$

and u^0 and u^1 are independent of ϵ and functionals of s. Insertion of the ansatz (2.9) into (1.2) yields

$$\begin{aligned} &u^0 = V(s), u^1 = V'(s)\left(P'(s) - V'(s)\right)\frac{\partial s}{\partial m} \quad \text{and} \\ &U^1 = V(s) + \epsilon V'(s)\left(P'(s) - V'(s)\right)\frac{\partial s}{\partial m}. \end{aligned} \tag{2.10}$$

Then s is determined by solving

$$\frac{\partial s}{\partial t} = \frac{\partial}{\partial m}\left(V(s) + \epsilon V'(s)\left(P'(s) - V'(s)\right)\frac{\partial s}{\partial m}\right). \tag{2.11}$$

This latter equation has a strong maximum principle so long as the initial data for s satisfies either

$$L \le s(m,0) < s_1 \quad \text{for all } m \tag{2.12}$$

or

$$s_2 \le s(m,0) < \infty \quad \text{for all } m \tag{2.13}$$

because in either of these cases the diffusion coefficient, $V'(s)(P'(s) - V'(s))$, is positive. On the other hand, when $s_1 < s < s_2$, the diffusion coefficient is negative and this yields explosive growth of the solution, confirming the instability of the constant solution (1.17) when $s_1 < s_0 < s_2$.

3. Large amplitude periodic traveling waves. In this section we seek solutions to (1.1) and (1.2) that are functions of

$$\xi = m + ct, \quad c > 0, \tag{3.1}$$

which are periodic in ξ with periodic M, the number of cars on the ring-road. The conversation structure of (1.1) implies that the $s(\cdot)$ component of the solution satisfies

$$\int_0^M s(\xi)d\xi = l, \tag{3.2}$$

where l is the length of the ring-road.

Insertion of the ansatz (3.1) into (1.1) implies that $u(\cdot)$ and $s(\cdot)$ satisfy

$$u(\xi) = u_\# + c(s(\xi) - s_\#), \tag{3.3}$$

and we insist that

$$u_\# = V(s_\#) \quad \text{and} \quad s(0) = s_\# \in (s_1, s_2). \tag{3.4}$$

The relations (3.3) and (3.4) further imply that

$$\epsilon c\,(c - P'(s))\,\frac{ds}{d\xi} = (V(s) - V(s_\#) - c(s - s_\#))\,. \tag{3.5}$$

We seek a solution to (3.4) and (3.5) which is increasing on $-m_a < \xi < M_a$, where $-m_a < 0 < M_a$. For speeds $0 < c < V'(s_\#)$, we see that the right-hand side of (3.5) satisfies

$$\text{sign } (V(s) - V(s_\#) - c(s - s_\#)) = \text{sign } (s - s_\#) \tag{3.6}$$

for $|s - s_\#|$ small enough, and thus to obtain an increasing solution to (3.4) and (3.5) on some interval containing $\xi = 0$ in its interior we are compelled to choose

$$c = P'(s_\#). \tag{3.7}$$

This choice of c, together with the hypothesis that $P''(\cdot) < 0$, guarantees that

$$\text{sign } (P'(s_\#) - P'(s)) = \text{ sign } (s - s_\#), \tag{3.8}$$

and thus, with this choice of c, we are guaranteed a solution of (3.4) and (3.5) defined in some interval $-\tilde{m}_a < \xi < \tilde{M}_a$, where $-\tilde{m}_a < 0 < \tilde{M}_a$. Moreover, this solution satisfies

$$\frac{ds}{d\xi}(0) = \frac{-\left(V'(s_\#) - P'(s_\#)\right)}{\epsilon P'(s_\#)P''(s_\#)} > 0 \tag{3.9}$$

for $s_1 < s_\# < s_2$.

We shall now refine the observations of the preceding paragraphs. If

$$V(L) - V(s_2) - P'(s_2)(L - s_2) > 0, \tag{3.10}$$

we let $\bar{s}$ in (s_1, s_2) be the unique solution of

$$V(L) - V(\bar{s}) - P'(\bar{s})(L - \bar{s}) = 0, \tag{3.11}$$

whereas, if

$$V(L) - V(s_2) - P'(s_2)(L - s_2) \leq 0, \tag{3.12}$$

we let

$$\bar{s} = s_2. \tag{3.13}$$

In either case, for any $s_\#$ in $(s_1, \bar{s})$ we let $L < s_-(s_\#) < s_\# < s_+(s_\#)$ be the other two solutions of

$$V(s_\pm) - V(s_\#) - P'(s_\#)(s_\pm - s_\#) = 0. \tag{3.14}$$

We of course have

$$V(s) - V(s_\#) - P'(s_\#)(s - s_\#) < 0, \quad s_-(s_\#) < s < s_\#, \tag{3.15}$$

and

$$V(s) - V(s_\#) - P'(s_\#)(s - s_\#) > 0, \quad s_\# < s < s_+(s_\#). \tag{3.16}$$

For any s_a in $(s_-(s_\#), s_\#)$ we now let $S(s_a) > s_\#$ be the unique solution of

$$\frac{P(S(s_a)) - P(s_a)}{S(s_a) - s_a} = P'(s_\#) \tag{3.17}$$

and note that

$$\frac{dS(s_a)}{ds_a} = \frac{(P'(s_\#) - P'(s_a))}{(P'(s_\#) - P'(S(s_a)))} < 0. \tag{3.18}$$

We also let $\underline{s}(s_\#)$ be the smallest value of $s_a \geq s_-(s_\#)$ such that $S(s_a) \leq s_+(s_\#)$ and for any s_a in $(\underline{s}(s_\#), s_\#)$ we let

$$-m_a = \epsilon P'(s_\#) \int_{s_a}^{s_\#} \frac{(P'(r) - P'(s_\#))\, dr}{(V(r) - V(s_\#) - P'(s_\#)(r - s_\#))} < 0 \tag{3.19}$$

and

$$M_a = \epsilon P'(s_\#) \int_{s_\#}^{S(s_a)} \frac{(P'(s_\#) - P'(r))dr}{(V(r) - V(s_\#) - P'(s_\#)(r - s_\#))} > 0. \tag{3.20}$$

We note that one of the integrals (3.19) or (3.20) or both diverge as $s_a \to \underline{s}(s_\#)^+$. For any ξ in $(-m_a, M_a)$, the solution to (3.4) and (3.5) is given by the quadrature formula

$$\epsilon P'(s_\#) \int_{s_\#}^{s(\xi)} \frac{(P'(s_\#) - P'(r))\, dr}{(V(r) - V(s_\#) - P'(s_\#)(r - s_\#))} = \xi, \tag{3.21}$$

and the solution is extended to $(-\infty, \infty)$ by insisting that the periodicity condition

$$s\left(\xi \pm n(m_a + M_a)\right) = s(\xi), \quad n = 0, 1, \dots, \tag{3.22}$$

holds. As constructed, the solution has jump discontinuities as the points $M_a \pm n(m_a + M_a)$, $n = 0, 1, \dots$, and (3.17), (3.19), and (3.20) guarantee that the Rankine–Hugoniot condition for (1.1) and (1.2) holds across these discontinuities. The Lax

entropy condition that $s^-\left(M_a \pm n(m_a + M_a)\right) > s^+\left(M_a \pm n(m_a + M_a)\right)$ is also guaranteed since

$$s^-(M_a \pm n(m_a + M_a)) = S(s_a) > s_a = s^+(M_a \pm n(m_a + M_a)). \tag{3.23}$$

What remains to be shown is that for integers $k = 1, 2, \ldots$ we can choose s_a in $(\underline{s}(s_\#), s_\#)$ and $s_\#$ in $(s_1, \overline{s})$ so that

$$k(m_a + M_a) = M \tag{3.24}$$

and

$$\int_0^M s(\xi)d\xi = l. \tag{3.25}$$

The integer k represents the number of increasing segments per period.

We start by analyzing (3.24). Equations (3.19) and (3.20) imply that solving (3.24) is equivalent to solving

$$k\epsilon P'(s_\#)\int_{s_a}^{S(s_a)} \frac{(P'(s_\#) - P'(r))\,dr}{(V(r) - V(s_\#) - P'(s_\#)(r - s_\#))} \stackrel{def}{=} F(s_\#, s_a) = M. \tag{3.26}$$

We observe for any $s_\#$ in $(s_1, \overline{s})$ that

$$F(s_\#, s_\#) = 0, \tag{3.27}$$

$$\begin{aligned}\frac{\partial F}{\partial s_a}(s_\#, s_a) = k\epsilon P'(s_\#)\Bigg(&\frac{(P'(s_\#) - P'(S(s_a))S'(s_a))}{(V(S(s_a)) - V(s_\#) - P'(s_\#)\,(S(s_a) - s_\#))} \\ &+ \frac{(P'(s_a) - P'(s_\#))}{(V(s_a) - V(s_\#) - P'(s_\#)(s_a - s_\#))}\Bigg)\end{aligned} \tag{3.28}$$

for any s_a in $(\underline{s}(s_\#), s_\#)$, and, finally, that

$$\lim_{s_a \to \underline{s}(s_\#)^+} F(s_\#, s_a) = +\infty. \tag{3.29}$$

Then (3.27)–(3.29) guarantee that for each $s_\#$ in $(s_1, \overline{s})$ there is a unique number $s_a(s_\#)$ in $(\underline{s}(s_\#), s_\#)$ satisfying (3.26). Thus, solving (3.24) and (3.25) is equivalent to finding an $s_\#$ in $(s_1, \overline{s})$ such that

$$k\epsilon P'(s_\#)\int_{s_a(s_\#)}^{S(s_a(s_\#))} \frac{(P'(s_\#) - P'(r))\,r dr}{(V(r) - V(s_\#) - P'(s_\#)(r - s_\#))} = l. \tag{3.30}$$

The last identity is a consequence of (3.25) and the fact that on $(-m_a, M_a)$

$$\frac{d\xi}{dr} = \frac{k\epsilon P'(s_\#)\,(P'(s_\#) - P'(r))}{(V(r) - V(s_\#) - P'(s_\#)(r - s_\#))}. \tag{3.31}$$

If we exploit the fact that $s_a(s_\#)$ satisfies (3.26), we find that solving (3.30) is equivalent to solving

$$Ms_\# + k\epsilon P'(s_\#)\int_{s_a(s_\#)}^{S(s_a(s_\#))} \frac{(P'(s_\#) - P'(r))\,(r - s_\#)dr}{(V(r) - V(s_\#) - P'(s_\#)(r - s_\#))} = l. \tag{3.32}$$

To get some idea about the range of the function defined by the left-hand side of (3.32) we note that

$$\text{sign}\left(\frac{(P'(s_\#)-P'(r))\,(r-s_\#)}{(V(r)-V(s_\#))-P'(s_\#)(r-s_\#)}\right)=\text{sign}\;(r-s_\#) \tag{3.33}$$

and that for r close to $s_\#$

$$\frac{(P'(s_\#)-P'(r))\,(r-s_\#)}{(V(r)-V(s_\#)-P'(s_\#)(r-s_\#))}\sim\frac{-P''(s_\#)(r-s_\#)}{(V'-P')\,(s_\#)}. \tag{3.34}$$

So long as $s_1<s_\#<\overline{s}$ we have $s_-(s_\#)<s_a(s_\#)$ and $S(s_a(s_\#))<s_+(s_\#)$ and the integrand $\frac{(P'(s_\#)-P'(r))(r-s_\#)}{(V(r)-V(s_\#)-P'(s_\#)(r-s_\#))}$ is nonsingular. In this case the function defined by the left-hand side of (3.32) is approximately given by

$$Ms_\#-\frac{k\epsilon P'(s_\#)P''(s_\#)(S(s_a(s_\#))-s_a(s_\#))(S(s_a(s_\#))+s_a(s_\#)-2s_\#)}{2(V'-P')(s_\#)}. \tag{3.35}$$

This last identity is instructive, especially in the situation where $P''(\cdot)$ is approximately constant. In that case $S(s_a(s_\#))+s_a(s_\#)-2s_\#$ is approximately zero and thus the function defined by (3.35) approximately reduces to $Ms_\#$. Equation (3.30) then approximately becomes

$$Ms_\#=l. \tag{3.36}$$

This sort of analysis on the function defined by the left-hand side of (3.30) is all we could manage with the degree of generality allowed on the functions $P(\cdot)$ and $V(\cdot)$. Though not particularly sharp it gives a fair indication of when (3.24) and (3.25) are solvable.

4. Simulations. All computations in this section were run with the follow-the-leader model (1.7) and (1.8) when

$$P(s)=\lambda\left(1-\frac{L}{s}\right),\quad L\le s, \tag{4.1}$$

and

$$V(s)=v_\infty\frac{\left(\tanh\left(\frac{s-rL}{\delta}\right)+\tanh\left(\frac{(r-1)L}{\delta}\right)\right)}{\left(1+\tanh\left(\frac{(r-1)L}{\delta}\right)\right)}. \tag{4.2}$$

The specific parameters used were

$$L=15\text{ feet}, \tag{4.3}$$

$$\lambda=150\text{ feet/sec }=102.2727\ldots\text{ mph}, \tag{4.4}$$

$$v_\infty=100\text{ feet/sec }=68.1818\ldots\text{ mph}, \tag{4.5}$$

$$\delta=15\text{ feet}, \tag{4.6}$$

and

$$r=3. \tag{4.7}$$

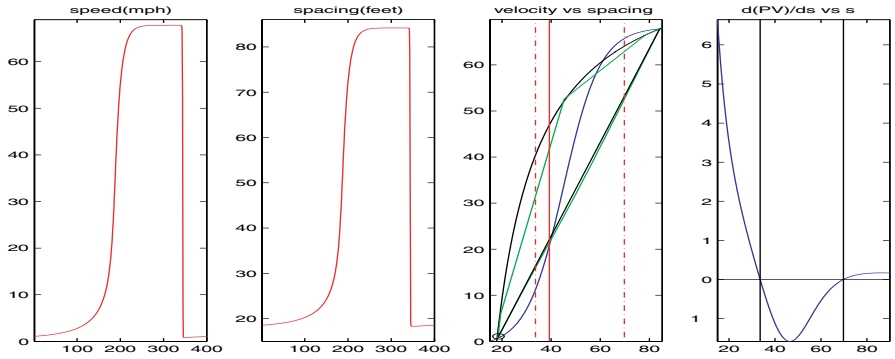


FIG. 1.

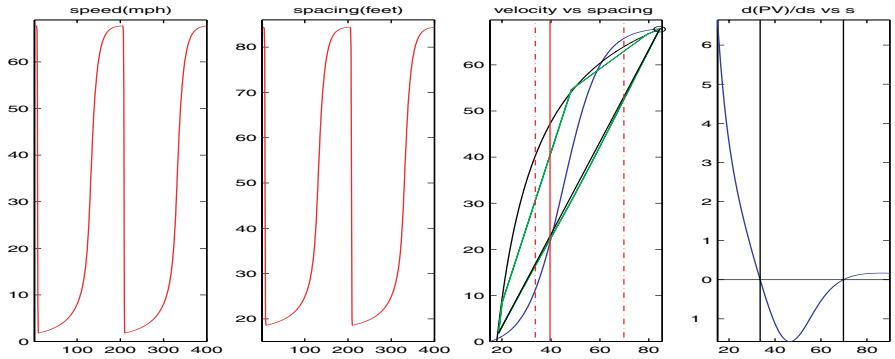


FIG. 2.

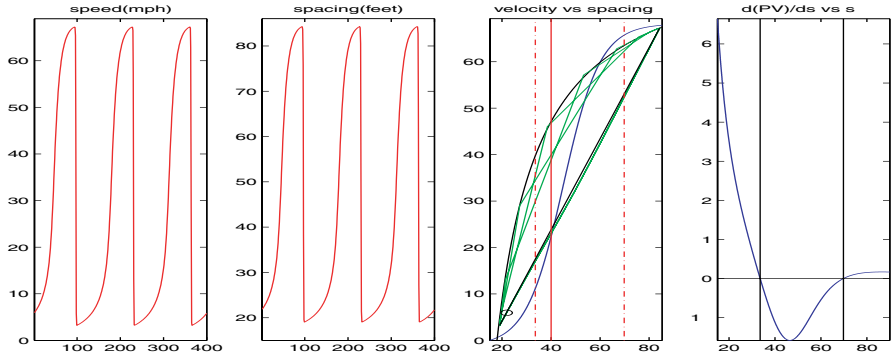


FIG. 3.

For initial data, we choose three sets of data

$$x_m^{(k)}(0) = 45m + 30 \sum_{j=0}^{m-1} \sin\left(\frac{kj\pi}{200}\right) \tag{4.8}$$

and

$$u_m^{(k)}(0) = 35 \text{ feet/sec} \tag{4.9}$$

for $m = 0, \pm 1, \pm 2, \dots$ and $k = 1, 2$, and 3. The observation that

$$x_{m+400}^{(k)}(0) = x_m^{(k)}(0) + 18000 \tag{4.10}$$

implies that we may interpret the data as initial data for a ring-road with 400 cars which is of length 18000 feet.

For our choice of parameter values the unstable region for $(P' - V')(\cdot)$ is the interval $33.59625\ldots < s < 69.8215$ and our data has initial car spacings

$$s_m^{(k)}(0) = x_{m+1}^{(k)}(0) - x_m^{(k)}(0) \tag{4.11}$$

which lie in that interval. A graph of $s \to (P' - V')(s)$ is shown in the fourth panel of Figures 1–3. Simulations were run with relaxation times

$$\epsilon = 1,\ 5,\ \text{and } 10. \tag{4.12}$$

We show the spatially periodic solutions at time $t = 1$ hour when $\epsilon = 10$ seconds. Figures 1, 2, and 3 correspond to the initial data indexed by $k = 1, 2$, and 3, respectively. The solution indexed by each particular k has k discontinuities per period after one hour. Run over a longer period, they all revert to a solution with one discontinuity per period.

The first two frames in each figure are self-explanatory. In the third frame of each figure we plot the curve $m \to (s_m = x_{m+1} - x_m, u_m)$. This curve is shown in green. The blue curve is the equilibrium curve $s \to (s, V(s))$ and the black concave curve is a suitably normalized image of $P(\cdot)$. The circle -o- is the image of (s_1, u_1).

REFERENCES

[1] C.F. Daganzo, *Requiem for second-order fluid approximations of traffic flow*, Transportation Res. Part B, 29 (1995), pp. 277–286.

[2] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[3] J.M. Greenberg, *Extensions and amplifications of a traffic model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.

[4] B. Argall, E. Cheleskin, J.M. Greenberg, C. Hinde, and P.-J. Lin, *A rigorous treatment of a follow-the-leader traffic model with traffic lights present*, SIAM J. Appl. Math., 63 (2002), pp. 149–168.

[5] A. Aw, A. Klar, T. Materne, and M. Rascle, *Derivation of continuum traffic flow models from microscopic follow-the-leader models*, SIAM J. Appl. Math., 63 (2002), 259–278.

[6] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, and Y. Sugiyama, *Dynamical model at traffic congestion and numerical simulation*, Phys. Rev. E(3), 57 (1995), pp. 1035–1042.

[7] J.M. Greenberg, A. Klar, and M. Rascle, *Congestion on multilane highways*, SIAM J. Appl. Math, 63 (2003), pp. 818–833.

[8] D. Helbing, *Traffic and related self-driven many-particle systems*, Rev. Modern Phys., 73 (2001), pp. 1067–1141.

[9] K. Nagel, P. Wagner, and R. Woesler, *Still flowing: Approaches to traffic flow and traffic jam modeling*, Oper. Res., 51 (2003), pp. 681–710.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1186–1197

SURFACE GREEN'S FUNCTIONS FOR AN INCOMPRESSIBLE, TRANSVERSELY ISOTROPIC ELASTIC HALF-SPACE*

RICHARD S. CHADWICK†, BRETT SHOELSON†, AND HONGXUE CAI†

Abstract. The surface displacements produced by normal and tangential point loads applied to the surface of an incompressible, transversely isotropic material are considered when anisotropy is produced by a single family of fibers oriented perpendicular to the surface normal. Three elastic constants (two shear moduli and a fiber modulus) characterize the linear elasticity of such a material. The problems are solved analytically in two-dimensional Fourier transform space, and explicit surface displacement formulae are given for the inverses in physical space. Simple relations are given as asymptotic expansions for weak anisotropy. Computed surface displacement patterns are illustrated, and the application of the results to atomic force microscopy is discussed.

Key words. anisotropic material, asymptotic analysis, elastic material, Green's functions

AMS subject classifications. 74B05, 74G05, 74E10, 74L15

DOI. 10.1137/S0036139903425338

1. Introduction.

The present problem is motivated by recent interest in using the atomic force microscope (AFM) as a microindenter of biological samples with the intent of determining the local elastic moduli of the tissue (Dimitriadis et al. (2002), Chadwick (2002)). The condition of material incompressibility can hold for biological samples when the time required to deform the tissue is small compared to the time required for water to flow out of it. Also, structural fibers do not penetrate biological membranes, which typically form the surfaces that are probed by the AFM. It seems, therefore, that the present problem is the simplest relevant departure from an isotropic half-space. A recent account of Green's functions for anisotropic materials is given by Pan and Yuan (2000). While the case of point loading of a compressible, transversely isotropic elastic half-space with fibers oriented in the direction of the surface normal has been well studied for normal loads (Lekhnitskii (1963), Willis (1966), Conway, Farnham, and Ku (1967), Green and Zerna (1968)) and tangential loads (Chen (1965), Turner (1980)), the present problem has apparently not received much specific attention. Here we use the stress-strain relation of Spencer (1984), the Fourier transform formalism developed by Willis (1966), and the inversion method developed by Barnett and Lothe (1975) for general anisotropic bodies. Spencer's stress-strain formalism has the advantages of avoiding a sometimes delicate limiting process needed to treat the incompressible case, and also having a more physical interpretation of the elastic constants, compared to the more general formalism used by Willis (1966), for example. Also, Willis, being interested in the Hertzian problem, did not need to invert the point force solutions. Barnett and Lothe (1975) developed their inversion method in conjunction with the Stroh formalism for crystal symmetry

*Received by the editors March 24, 2003; accepted for publication (in revised form) August 25, 2003; published electronically April 21, 2004.

http://www.siam.org/journals/siap/64-4/42533.html

†Section on Auditory Mechanics, National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Bethesda, MD 20892 (chadwick@helix.nih.gov, shoelson@helix.nih.gov, hongxuec@helix.nih.gov).

classes, as they describe and extend. The Stroh formalism, with hexagonal symmetry, can also be used to study problems of biological interest such as we consider here.

2. Formulation. We consider a half-space (x, y, z) with the elastic material in the domain $z < 0$. Let the fibers run in a direction specified by the unit vector $\overrightarrow{a}$. The stress-strain relation of the material is given by Spencer (1984),

$$\sigma_{ij} = -pI_{ij} + 2\mu_T e_{ij} + E_f a_k e_{kl} a_l a_i a_j + 2(\mu_L - \mu_T)(a_i a_k e_{kj} + e_{ik} a_k a_j), \tag{2.1}$$

where p is an isotropic pressure term required for an incompressible material; I_{ij} is the unit tensor; e_{ij} is the strain tensor $\frac{1}{2}(\partial u_i/\partial x_j + \partial u_j/\partial x_i)$, where the u_i are displacements; and E_f, μ_T, and μ_L are the fiber modulus and the shear moduli normal and parallel to the fiber direction, respectively. Introducing the displacement components (u, v, w) and taking $\overrightarrow{a} : (0, 1, 0)$, so that fibers run in the y direction, the stress components are

$$\sigma_{xx} = -p + 2\mu_T \frac{\partial u}{\partial x}, \quad \sigma_{yy} = -p + (E_f + 4\mu_L - 2\mu_T)\frac{\partial v}{\partial y}, \quad \sigma_{zz} = -p + 2\mu_T \frac{\partial w}{\partial z}, \tag{2.2}$$

$$\sigma_{xy} = \mu_L \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right), \quad \sigma_{zy} = \mu_L \left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z} \right), \quad \sigma_{xz} = \mu_T \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right). \tag{2.3}$$

Substitution of (2.2), (2.3) into the equations of equilibrium $\partial \sigma_{ij}/\partial x_j = 0$, and using the condition of incompressibility $\partial u_i/\partial x_i = 0$, yields the following system:

$$(2\mu_T - \mu_L)\frac{\partial^2 u}{\partial x^2} + \mu_L \frac{\partial^2 u}{\partial y^2} + \mu_T \frac{\partial^2 u}{\partial z^2} + (\mu_T - \mu_L)\frac{\partial^2 w}{\partial x \partial z} - \frac{\partial p}{\partial x} = 0, \tag{2.4}$$

$$\mu_L \frac{\partial^2 v}{\partial x^2} + (E_f + 3\mu_L - 2\mu_T)\frac{\partial^2 v}{\partial y^2} + \mu_L \frac{\partial^2 v}{\partial z^2} - \frac{\partial p}{\partial y} = 0, \tag{2.5}$$

$$\mu_T \frac{\partial^2 w}{\partial x^2} + \mu_L \frac{\partial^2 w}{\partial y^2} + (2\mu_T - \mu_L)\frac{\partial^2 w}{\partial z^2} + (\mu_T - \mu_L)\frac{\partial^2 u}{\partial x \partial z} - \frac{\partial p}{\partial z} = 0, \tag{2.6}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0. \tag{2.7}$$

To complete the formulation we consider three separate surface stress conditions:

$$\sigma_{xz}(x, y, 0) = \delta(x)\delta(y), \quad \sigma_{yz}(x, y, 0) = 0, \quad \sigma_{zz}(x, y, 0) = 0, \tag{2.8}$$

$$\sigma_{xz}(x, y, 0) = 0, \quad \sigma_{yz}(x, y, 0) = \delta(x)\delta(y), \quad \sigma_{zz}(x, y, 0) = 0, \tag{2.9}$$

$$\sigma_{xz}(x, y, 0) = 0, \quad \sigma_{yz}(x, y, 0) = 0, \quad \sigma_{zz}(x, y, 0) = -\delta(x)\delta(y), \tag{2.10}$$

where $\delta(x_i)$ denotes the Dirac delta function. In each case all stress components must decay as $z \to -\infty$. The conditions specified by (2.8) represent a point shear force acting in a cross-fiber direction (see Problem 1), those specified by (2.9) represent a point shear force acting in the fiber direction (Problem 2), and those specified by (2.10) represent a compressive point normal force (Problem 3).

3. Solutions in Fourier transform space. Following Willis (1966), we take Fourier transforms in the x and y directions. Thus

$$\overline{U}(\xi,\eta,z)=\frac{1}{2\pi}\int\int_{-\infty}^{\infty}u(x,y,z)e^{[i(\xi x+\eta y)]}dxdy, \tag{3.1}$$

with similar expressions for $\overline{V}$, $\overline{W}$, and $\overline{P}$, the transforms of v, w, and p. Transforming the set (2.4)–(2.7) yields a system of coupled ordinary differential equations in z. Assuming a solution form $\overline{U}(\xi,\eta,z)=U(\xi,\eta)\exp[i\lambda z]$, and similarly for the other transformed variables, gives the homogeneous algebraic system

$$\begin{aligned}
-[(2-\beta)\xi^2+\beta\eta^2+\lambda^2]U+(1-\beta)\xi\lambda W+\frac{i\xi P}{\mu_T}&=0,\\
-[\beta\xi^2+(\alpha+3\beta-2)\eta^2+\beta\lambda^2]V+\frac{i\eta P}{\mu_T}&=0,\\
(1-\beta)\xi\lambda U-[\xi^2+\beta\eta^2+(2-\beta)\lambda^2]W-\frac{i\lambda P}{\mu_T}&=0,\\
i\xi U+i\eta V-i\lambda W&=0,
\end{aligned} \tag{3.2}$$

which has solutions, provided that the determinant of the matrix $\underline{M}$ of coefficients of U, V, W, and P is zero. Here we have introduced the ratios of elastic moduli: $\alpha=E_f/\mu_T$ and $\beta=\mu_L/\mu_T$. The determinant is a sixth-order polynomial in λ that turns out to be bicubic and factorable:

$$(\lambda^2+\xi^2+\beta\eta^2)[\beta\lambda^4+(\alpha\eta^2+2\beta\eta^2+2\beta\xi^2)\lambda^2+\beta\eta^4+(\alpha+2\beta)\xi^2\eta^2+\beta\xi^4]=0. \tag{3.3}$$

Of the six roots, which occur in complex conjugate pairs, we choose those that have negative imaginary parts to ensure decay of stress as $z\to-\infty$. Thus, from the first factor we obtain

$$\lambda_1=-i\sqrt{\xi^2+\beta\eta^2}, \tag{3.4}$$

while from the second, which is biquadratic, we find

$$\begin{aligned}
\lambda_2&=-i\sqrt{\xi^2+\eta^2\left[1+\frac{\alpha}{2\beta}(1+\phi)\right]},\\
\lambda_3&=-i\sqrt{\xi^2+\eta^2\left[1+\frac{\alpha}{2\beta}(1-\phi)\right]},
\end{aligned} \tag{3.5}$$

where $\phi=\sqrt{1+4\frac{\beta}{\alpha}}$. Note that each of the three radicands is positive for positive elastic moduli. Corresponding to each of these roots are three vectors $\overrightarrow{b_i}:(U_i,V_i,W_i,P_i)$, that are the eigenvectors corresponding to the zero eigenvalues of the matrices $\underline{M}(\lambda_i)$, $i=1,2,3$. These vectors have the following components: $\overrightarrow{b_1}:(\lambda_1/\xi,0,1,0)$, $\overrightarrow{b_2}:(U_2,V_2,(\xi U_2+\eta V_2)/\lambda_2,1)$, and $\overrightarrow{b_3}:(U_3,V_3,(\xi U_3+\eta V_3)/\lambda_3,1)$, where

$$\begin{aligned}
V_2&=\frac{i}{\eta[2(\beta-1)+\frac{1}{2}\alpha(1-\phi)]},\\
U_2&=\frac{\beta\xi V_2[2(\beta-1)+\alpha(1-\phi)]}{\eta[2\beta(\beta-1)-\alpha(1+\phi)]},
\end{aligned} \tag{3.6}$$

and

$$V_3 = \frac{i}{\eta[2(\beta-1)+\frac{1}{2}\alpha(1+\phi)]}, \tag{3.7}$$
$$U_3 = \frac{\beta\xi V_3[2(\beta-1)+\alpha(1+\phi)]}{\eta[2\beta(\beta-1)-\alpha(1-\phi)]}.$$

The general solution for the transformed variables $\overrightarrow{\Psi} : (\overline{U}, \overline{V}, \overline{W}, \overline{P})$ is the superposition of the three eigenvectors, i.e.,

$$\Psi_j(\xi,\eta,z) = \sum_{n=1}^{3} A_n b_{nj} e^{i\lambda_n z}, \quad j = 1,2,3,4, \tag{3.8}$$

where $\Psi_1 = \overline{U}$, $\Psi_2 = \overline{V}$, ... , and b_{nj} denotes the jth component of the eigenvector $\overrightarrow{b_n}$. The A_n are arbitrary functions of (ξ,η) that will be evaluated by applying the three stress boundary conditions appropriate for each problem. Using (2.2), (2.3), (3.1), and (3.8), the transformed surface stresses are

$$\overline{\sigma_{xz}}(\xi,\eta,0) = i\mu_T \sum_{n=1}^{3} A_n(\lambda_n b_{n1} - \xi b_{n3}), \tag{3.9}$$
$$\overline{\sigma_{yz}}(\xi,\eta,0) = i\mu_L \sum_{n=1}^{3} A_n(\lambda_n b_{n2} - \eta b_{n3}),$$
$$\overline{\sigma_{zz}}(\xi,\eta,0) = \mu_T \sum_{n=1}^{3} A_n(2i\lambda_n b_{n3} - b_{n4}).$$

The solution of Problem 1 is obtained by solving the linear system (3.9) for A_1, A_2, A_3 when the left-hand sides are set to $1/(2\pi)$, 0 , 0, respectively. Similarly, for Problems 2 and 3 the left-hand sides are set to 0, $1/(2\pi)$, 0, and 0, 0, $-1/(2\pi)$, respectively.

4. Fourier inversions. We now adopt the more concise notation $G_{ij}(x,y)$ or $G_{ij}(r,\varphi)$ in polar coordinates to denote the Green's tensor that represents the surface displacement component in the ith direction ($i = 1,2,3$) due to a point force in the jth problem or direction ($j = 1,2,3$), and $\bar{G}_{ij}(\xi,\eta)$ or $\bar{G}_{ij}(\rho,\theta)$ to denote the corresponding transforms. Willis (1966), Barnett and Lothe (1975), and Pan and Yuan (2000) have shown for compressible general anisotropic materials that the $\bar{G}_{ij}$ are homogeneous of degree -1 in ρ, i.e., $\bar{G}_{ij}(\rho,\theta) = \rho^{-1}\bar{g}_{ij}(\theta)$. It is easily seen that this result carries over to the present incompressible case. This is important because it enables a great simplification of the inversions into physical space:

$$G_{ij}(x,y) = \frac{1}{2\pi}\int\int_{-\infty}^{\infty} \bar{G}_{ij}(\xi,\eta) e^{-i(\xi x+\eta y)} d\xi d\eta, \tag{4.1}$$
$$G_{ij}(r,\varphi) = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} \rho^{-1}\bar{g}_{ij}(\theta) e^{-i\rho r\cos(\theta-\varphi)} \rho d\rho d\theta,$$

where polar coordinates $\xi = \rho\cos\theta$, $\eta = \rho\sin\theta$, and $x = r\cos\varphi$, $y = r\sin\varphi$ have been introduced. The radial integration is accomplished via the identity used by Barnett and Lothe (1975):

$$\int_0^{\infty} e^{-i\rho r\cos(\theta-\varphi)} d\rho = \pi\delta(r\cos(\theta-\varphi)) - iP\frac{1}{r\cos(\theta-\varphi)}, \tag{4.2}$$

where P denotes the principal value. Furthermore, Barnett and Lothe introduced the functional identity

$$\delta(r\cos(\theta-\varphi)) = \frac{\delta(\theta-\varphi\pm\frac{\pi}{2})}{|r\sin(\theta-\varphi)|}. \tag{4.3}$$

These two identities lead directly to the inversion formula

$$G_{ij}(r,\varphi) = \frac{1}{4\pi\mu_T r}\bar{g}_{ij}(\theta)|_{\varphi\pm\frac{\pi}{2}} - \frac{i}{4\pi^2\mu_T r}P\int_0^{2\pi}\frac{\bar{g}_{ij}(\theta)}{\cos(\theta-\varphi)}d\theta. \tag{4.4}$$

The first term of the inversion formula is simply the sum of two ninety-degree rotations of the functions in transform space, while the second term is a principal value integral. In the present problem we note the following symmetry properties of the $\bar{g}_{ij}(\theta)$ in terms of their Fourier series expansions on $0 \le \theta \le 2\pi$: $\bar{g}_{11}, \bar{g}_{22}$, and $\bar{g}_{33}$ have cosine series expansions in even multiples of θ; $\bar{g}_{12}$ and $\bar{g}_{21}$ have sine series expansions in even multiples of θ; $\bar{g}_{13}$ and $\bar{g}_{31}$ have cosine series expansions in odd multiples of θ; and $\bar{g}_{23}$ and $\bar{g}_{32}$ have sine series expansions in odd multiples of θ. The principal-value integrals vanish for those functions that have an expansion in even multiples of θ, while the sum of the two ninety-degree rotations are additive for those functions. Thus G_{11}, G_{22}, G_{33}, G_{12}, and G_{21} simply reduce to

$$G_{ij}(r,\varphi) = \frac{1}{2\pi\mu_T r}\bar{g}_{ij}\left(\varphi+\frac{\pi}{2}\right). \tag{4.5}$$

For the remaining off-diagonal terms $\bar{g}_{13}$, $\bar{g}_{31}$, $\bar{g}_{23}$, and $\bar{g}_{32,}$which have expansions in odd multiples of θ, the sum of the two ninety-degree rotations cancels, and their principal-value integrals can be evaluated using the following formulae given by Barnett and Lothe (1975):

$$P\int_0^{2\pi}\frac{\cos(2n+1)}{\sin(\theta-\theta_0)}d\theta = -2\pi\sin(2n+1)\theta_0, \tag{4.6}$$

$$P\int_0^{2\pi}\frac{\sin(2n+1)}{\sin(\theta-\theta_0)}d\theta = 2\pi\cos(2n+1)\theta_0, \tag{4.7}$$

where $n = 0, 1, 2, \ldots$. Thus G_{13} and G_{23} reduce to the sums

$$G_{13}(r,\varphi) = \frac{1}{2\pi\mu_T r}\sum_{n=0}^{\infty} c_n \sin\left[\left(2n+1\right)(\varphi+\frac{\pi}{2})\right] \tag{4.8}$$

$$G_{23}(r,\varphi) = -\frac{1}{2\pi\mu_T r}\sum_{n=0}^{\infty} s_n \cos\left[\left(2n+1\right)(\varphi+\frac{\pi}{2})\right] \tag{4.9}$$

where the c_n are the Fourier cosine coefficients of $-i\bar{g}_{13}$, and the s_n are the Fourier sine coefficients of $-i\bar{g}_{23}$. We will see in the following sections that $G_{12} = G_{21,}$ $G_{13} = G_{31}$, and $G_{32} = G_{23}$.

5. Asymptotic results for small anisotropy. Although the exact expressions for the transformed surface displacements are quite unwieldy and are not displayed here, they can be easily obtained using Mathematica (Wolfram Research, Chicago, IL). As an independent check on the exact method given in section 4, it turns out that Mathematica can invert the transformed surface displacements term by term when

they are expanded in a Taylor series about $\alpha = 0$ and $\beta = 1$ for small anisotropy. Here we summarize the results, where an expanded surface displacement in transform space is immediately followed by its corresponding inverse transform into physical space. The method for obtaining these inverses is described in the appendix.

PROBLEM 1.

$$
\begin{aligned}
2\pi\mu_T U(\xi,\eta) &= \frac{\xi^2+2\eta^2}{2(\xi^2+\eta^2)^{3/2}} - (\beta-1)\frac{\eta^2}{2(\xi^2+\eta^2)^{3/2}} - \alpha\frac{5\xi^2\eta^4}{16(\xi^2+\eta^2)^{7/2}} + \cdots, \\
2\pi\mu_T u(x,y,0) &= \frac{y^2+2x^2}{2(y^2+x^2)^{3/2}} - (\beta-1)\frac{x^2}{2(y^2+x^2)^{3/2}} - \alpha\frac{5y^2x^4}{16(y^2+x^2)^{7/2}} + \cdots, \\
2\pi\mu_T V(\xi,\eta) &= -\frac{\xi\eta}{2(\xi^2+\eta^2)^{3/2}} + (\beta-1)\frac{\xi\eta}{2(\xi^2+\eta^2)^{3/2}} + \alpha\frac{\xi\eta^3(6\xi^2+\eta^2)}{16(\xi^2+\eta^2)^{7/2}} + \cdots, \\
2\pi\mu_T v(x,y,0) &= \frac{yx}{2(y^2+x^2)^{3/2}} - (\beta-1)\frac{yx}{2(y^2+x^2)^{3/2}} - \alpha\frac{yx^3(6y^2+x^2)}{16(y^2+x^2)^{7/2}} + \cdots, \\
2\pi\mu_T W(\xi,\eta) &= i(\beta-1)\frac{\xi\eta^2}{2(\xi^2+\eta^2)^2} + i\alpha\frac{\xi\eta^4}{8(\xi^2+\eta^2)^3} + \cdots, \\
2\pi\mu_T w(x,y,0) &= (\beta-1)\frac{x(x^2-y^2)}{4(y^2+x^2)^2} + \alpha\frac{x(3x^4-6x^2y^2-y^4)}{64(y^2+x^2)^3} + \cdots.
\end{aligned} \tag{5.1}
$$

PROBLEM 2.

$$
\begin{aligned}
2\pi\mu_T U(\xi,\eta) &= -\frac{\xi\eta}{2(\xi^2+\eta^2)^{3/2}} + (\beta-1)\frac{\xi\eta}{2(\xi^2+\eta^2)^{3/2}} + \alpha\frac{\xi\eta^3(6\xi^2+\eta^2)}{16(\xi^2+\eta^2)^{7/2}} + \cdots, \\
2\pi\mu_T u(x,y,0) &= \frac{yx}{2(y^2+x^2)^{3/2}} - (\beta-1)\frac{yx}{2(y^2+x^2)^{3/2}} - \alpha\frac{yx^3(6y^2+x^2)}{16(y^2+x^2)^{7/2}} + \cdots, \\
2\pi\mu_T V(\xi,\eta) &= \frac{2\xi^2+\eta^2}{2(\xi^2+\eta^2)^{3/2}} - (\beta-1)\frac{2\xi^2+\eta^2}{2(\xi^2+\eta^2)^{3/2}} - \alpha\frac{\eta^2(8\xi^4+4\xi^2\eta^2+\eta^4)}{16(\xi^2+\eta^2)^{7/2}} + \cdots, \\
2\pi\mu_T v(x,y,0) &= \frac{2y^2+x^2}{2(y^2+x^2)^{3/2}} - (\beta-1)\frac{2y^2+x^2}{2(y^2+x^2)^{3/2}} - \alpha\frac{x^2(8y^4+4y^2x^2+x^4)}{16(y^2+x^2)^{7/2}} + \cdots, \\
2\pi\mu_T W(\xi,\eta) &= -i(\beta-1)\frac{\xi^2\eta}{2(\xi^2+\eta^2)^2} - i\alpha\frac{\xi^2\eta^3}{8(\xi^2+\eta^2)^3} + \cdots, \\
2\pi\mu_T w(x,y,0) &= -(\beta-1)\frac{y(y^2-x^2)}{4(y^2+x^2)^2} + \alpha\frac{y(3x^4-6x^2y^2-y^4)}{64(y^2+x^2)^3} + \cdots.
\end{aligned} \tag{5.2}
$$

PROBLEM 3.

$$
\begin{aligned}
2\pi\mu_T U(\xi,\eta) &= i(\beta-1)\frac{\xi\eta^2}{2(\xi^2+\eta^2)^2} + i\alpha\frac{\xi\eta^4}{8(\xi^2+\eta^2)^3} + \cdots, \\
2\pi\mu_T u(x,y,0) &= (\beta-1)\frac{x(x^2-y^2)}{4(y^2+x^2)^2} + \alpha\frac{x(3x^4-6x^2y^2-y^4)}{64(y^2+x^2)^3} + \cdots, \\
2\pi\mu_T V(\xi,\eta) &= -i(\beta-1)\frac{\xi^2\eta}{2(\xi^2+\eta^2)^2} - i\alpha\frac{\xi^2\eta^3}{8(\xi^2+\eta^2)^3} + \cdots, \\
2\pi\mu_T v(x,y,0) &= (\beta-1)\frac{y(x^2-y^2)}{4(y^2+x^2)^2} + \alpha\frac{y(3x^4-6x^2y^2-y^4)}{64(y^2+x^2)^3} + \cdots,
\end{aligned} \tag{5.3}
$$

$$2\pi\mu_T W(\xi,\eta) = -\frac{1}{2(\xi^2+\eta^2)^{1/2}} + (\beta-1)\frac{\eta^2}{2(\xi^2+\eta^2)^{3/2}} + \alpha\frac{\eta^4}{16(\xi^2+\eta^2)^{5/2}} + \cdots,$$

$$2\pi\mu_T w(x,y,0) = -\frac{1}{2(y^2+x^2)^{1/2}} + (\beta-1)\frac{x^2}{2(y^2+x^2)^{3/2}} + \alpha\frac{x^4}{16(y^2+x^2)^{5/2}} + \cdots.$$

Notice that the formulae reduce to the well-known results for an incompressible isotropic material (cf. Landau and Lifshitz (1975, pp. 29–30)) when $\alpha = 0$ and $\beta = 1$.

6. Symmetry properties of Green's tensor. Reverting back to the $G_{ij}(x,y)$ notation, we see from the preceding asymptotic formulae the surprising result that $G_{ij}(x,y) = G_{ji}(x,y)$ for $i \neq j$. This is also true for the *exact* transformed displacements, according to logical comparisons in Mathematica. Should we be surprised by this result, and how general is it? The Green's tensor is symmetric for an isotropic half-space and infinite medium (Landau and Lifshitz (1975, pp. 29–30)), as well as for a half-space with cubic crystal symmetry (Portz and Maradudin (1977)). However, other authors do not mention this symmetry for general anisotropic half-space problems. Courant and Hilbert (1962, pp. 393–394) indicate that a Green's tensor has this symmetry property when the differential system is self-adjoint. As we shall show, this symmetry depends upon whether or not a reciprocal theorem exists in the form

$$\int\int_S \overrightarrow{\sigma}\cdot\overrightarrow{u^*}\,dS = \int\int_S \overrightarrow{\sigma^*}\cdot\overrightarrow{u}\,dS,$$

where $\overrightarrow{u}$ and $\overrightarrow{u^*}$ are the respective surface displacements resulting from two different surface stress vectors $\overrightarrow{\sigma}$ and $\overrightarrow{\sigma^*}$, applied to an elastic body having a surface S. Such a reciprocal relation has been attributed to Betti for elastic systems and Rayleigh for more general dynamic systems (see Timoshenko (1930, pp. 351–355)). A similar relation for incompressible creeping viscous flow was established by Lorentz (see Happel and Brenner (1965, pp. 85–87)). The reciprocal theorem can be shown to hold not only for the elastic solid considered here, but also for a general linear anisotropic elastic solid having the stress-strain relation $\sigma_{ij} = c_{ijkl}e_{kl}$. The geometry of the body can also be arbitrary. To see this, consider the quadratic form $\sigma_{ij}e^*_{ij} = c_{ijkl}e_{kl}e^*_{ij} = c_{klij}e^*_{ij}e_{kl} = \sigma^*_{kl}e_{kl} = \sigma^*_{ij}e_{ij}$, noting that the switch $c_{ijkl} = c_{klij}$ is permissible since the strain energy $\frac{1}{2}\,c_{ijkl}e_{ij}e_{kl}$ must be invariant to the switch. Consider further, from the definition of strain, $\sigma_{ij}e^*_{ij} = \frac{1}{2}\,\sigma_{ij}\partial u^*_i/\partial x_j + \frac{1}{2}\,\sigma_{ij}\partial u^*_j/\partial x_i = \frac{1}{2}\,\sigma_{ij}\partial u^*_i/\partial x_j + \frac{1}{2}\,\sigma_{ji}\partial u^*_j/\partial x_i = \sigma_{ij}\partial u^*_i/\partial x_j$, where symmetry of the stress tensor $\sigma_{ij} = \sigma_{ji}$ has been used. Therefore $\sigma_{ij}e^*_{ij}$ can be written in the divergence form $\sigma_{ij}e^*_{ij} = \partial(\sigma_{ij}u^*_i)/\partial x_j$, since $\partial\sigma_{ij}/\partial x_j = 0$ is the condition of static equilibrium. Therefore $\partial(\sigma_{ij}u^*_i)/\partial x_j = \partial(\sigma^*_{ij}u_i)/\partial x_j$, which is the condition of self-adjointness. Integrating over the volume of the body and applying the divergence theorem gives the surface integral form of the reciprocal relation. The symmetry of Green's tensor follows from this relation when the stress vectors are due to point forces. A breakdown of the reciprocal theorem evidently occurs in problems involving elastic inclusions and inhomogeneities (Eshelby (1961)) and lattice defects (Eshelby (1956)). In such cases the difference between the two sides of the reciprocal theorem is identified with an interaction energy.

7. Surface displacement patterns. We calculate surface displacements using the formulae given in section 4. Surface displacement patterns are displayed graphically as superpositions of contour plots for the displacement component normal to

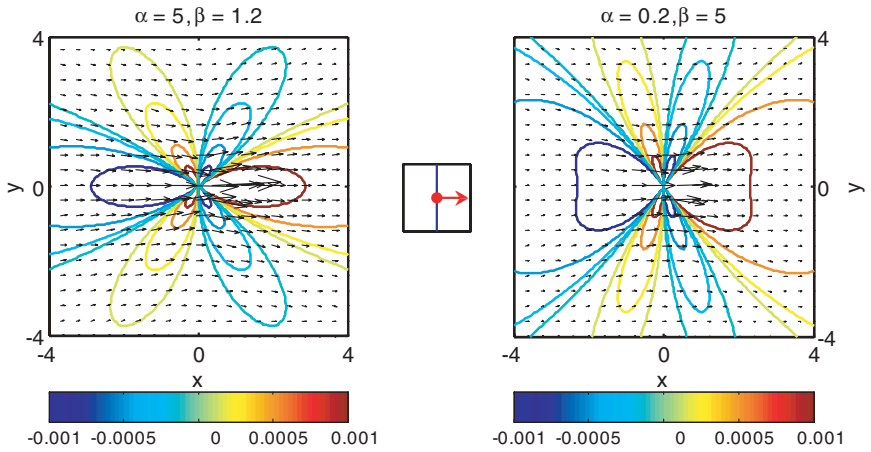


FIG. 1. *In Problem* 1, *a shear point force is applied at the origin, with the force oriented normal to the y-directed family of fibers. The center inset shows the fiber orientation, and the force direction is shown by the red arrow. The contours represent w, the z-directed deformation of the surface of the material, where the in-plane displacements are shown as arrows.*

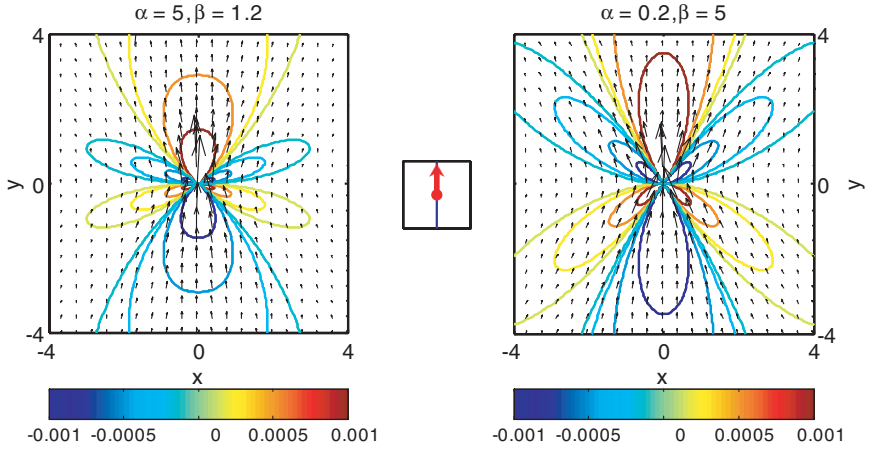


FIG. 2. *In Problem* 2, *a shear point force is applied at the origin, oriented in the direction of the fibers.*

the surface, together with quiver plots for the in-plane displacements. Patterns for Problems 1–3 are shown in Figures 1–3, respectively. Each figure comprises two panels. In the left panel, $\alpha = 5$ and $\beta = 1.2$; in this case the fiber modulus E_f is 5 times larger than the transverse shear modulus μ_T, while the longitudinal shear modulus μ_L is 1.2 times larger than μ_T. In the right panel, $\alpha = 0.2$, $\beta = 5$; in this case E_f is 5 times smaller than μ_T, while μ_L is 5 times larger than μ_T. The left panels therefore depict surface displacements induced by surface point forces in a material in which

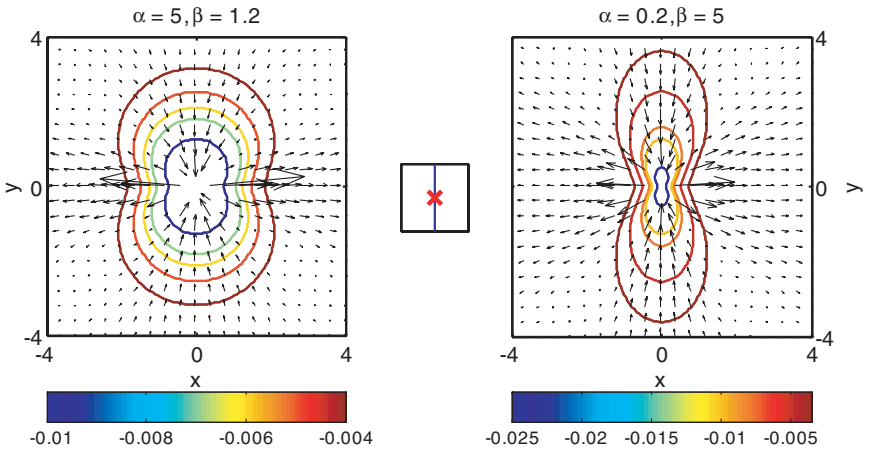


FIG. 3. *In Problem* 3*, a normal point load is applied at the origin. The red "x" in the center inset depicts a force directed into the surface of the half-space.*

fibers are relatively stiff compared to the surrounding matrix, while the right panels show the response of a material whose matrix is relatively stiff in a direction normal to relatively compliant fibers. Figure 4 comprises a triad of panels representing Problems 1–3 for the set of parameters $\alpha = 10$, $\beta = 0.2$. These parameters could be representative of a biological gel-like tissue having embedded, directed collagenous fibers.

8. Discussion. It is of interest to outline how the Green's functions found in this paper might be used to extract the three elastic moduli of isolated small samples of biological material using the AFM. The AFM piezo/cantilever system can be used to apply normal or tangential loads, the magnitudes of which can be determined by measuring the deflections induced by the bending or twisting of calibrated cantilevers using the laser/photodetector system of the instrument. The AFM can also map topographic variations of the sample surface on a nanometer scale. The following strategies can be used to sequentially determine μ_T, μ_L, and E_f. Since elements of the Green's tensor $G_{ij}(x, y)$ represent surface displacements in the ith direction that result from a unit point force at the origin acting in the jth direction, the surface displacements resulting from a force $\vec{f}$ acting at the origin are $u_i(x, y) = G_{ij}(x, y)f_j$. By applying a known tangential force of magnitude f_1 in the cross-fiber direction, and noting that $G_{11}(0, y)$ does not involve μ_L or E_f (cf. section 5), we can estimate the transverse shear modulus μ_T by fitting measured surface displacements $u_1(0, y)$ to a function of $1/y$. If the tangential force were applied by the AFM cantilever, then $u_1(0, y)$ could be measured by optical imaging of microspheres, for example. Alternatively, if the tangential force were applied by another calibrated probe, then the AFM cantilever would be available for scanning the topography before and after application of the force. Subtraction and postprocessing of two topographic images would then provide an estimate of $u_1(0, y)$, provided that the signal-to-noise ratio were large enough. Similarly, $G_{11}(x, 0)$ does not involve E_f, so the longitudinal shear modulus μ_L can be estimated by fitting measured surface displacements $u_1(x, 0)$ to a function of $1/x$. Finally, applying a known tangential force of magnitude f_2 in

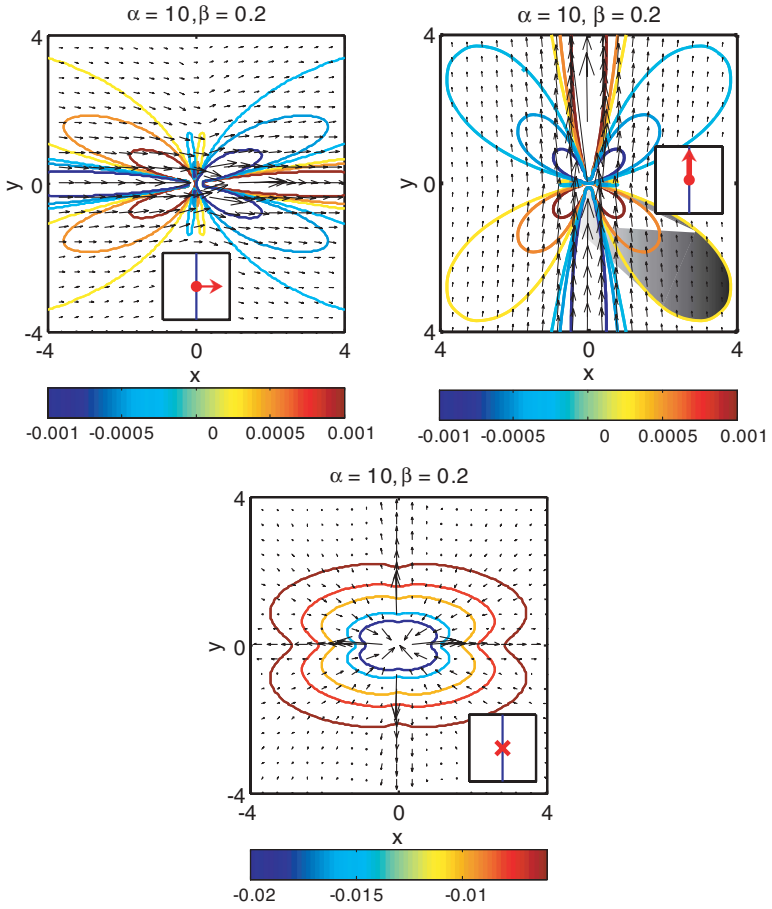


FIG. 4. *Surface displacement patterns for Problems* 1–3. *Parameters are thought to be representative of a biological gel-like material having directed, embedded collagenous fibers.*

the fiber direction and fitting measured values of $u_2(x,0)$ to a function of $1/x$ could provide an estimate of E_f via the $G_{22}(x,0)$ function. Other strategies using only a normal force of magnitude f_3 can be devised using $G_{33}(0,y)$, $G_{13}(x,x)$, etc.

Finally, it is worth estimating the distance from the probe in terms of the contact radius a, at which the point-force solution should provide a good approximation to the contact problem. For an isotropic, incompressible material, the ratio of the normal displacement resulting from Hertzian contact (outside the contact region) to that from a normal point force is given by

$$\frac{3}{4}\zeta\left[(2-\zeta^2)\sin^{-1}\frac{1}{\zeta}+\sqrt{\zeta^2-1}\right]=1+\frac{1}{10\zeta^2}+\cdots \qquad \text{as } \zeta\to\infty,$$

where $\zeta = r/a$. From this relationship, we see that the point-force solution is very quickly approached only a few contact radii away from the origin.

Appendix. Evaluation of some inverse Fourier transforms. Mathematica can invert two key transform types, from which all of the results in section 5 can be

obtained. One is the remarkable result that transformed functions having the form

$$F_m(\xi,\eta) = \frac{\eta^{2m}}{(\xi^2+\eta^2)^{m+\frac{1}{2}}}, \quad m = 0, 1, \dots, \tag{A.1}$$

have inverse transforms in x, y space that are simply rotations of $\pi/2$ radians of the function itself; i.e.,

$$f_m(x,y) = F_m^{-1}(\xi,\eta) = \frac{x^{2m}}{(y^2+x^2)^{m+\frac{1}{2}}}, \quad m = 0, 1, \dots. \tag{A.2}$$

Inverse transforms of

$$F_{mn}(\xi,\eta) = \frac{\eta^{2m-n}\xi^n}{(\xi^2+\eta^2)^{m+\frac{1}{2}}}, \quad m = 0, 1, \dots, n = 1, 2, \dots, \tag{A.3}$$

can be easily evaluated by recursively integrating $f_m(x,y)$ with respect to y and differentiating the result with respect to x, e.g.,

$$f_{31}(x,y) = \frac{\partial}{\partial x}\int \frac{x^6}{(y^2+x^2)^{\frac{7}{2}}}dy = -\frac{yx^5}{(y^2+x^2)^{\frac{7}{2}}}, \tag{A.4}$$

$$f_{32}(x,y) = -\frac{\partial}{\partial x}\int \frac{yx^5}{(y^2+x^2)^{\frac{7}{2}}}dy = \frac{y^2x^4}{(y^2+x^2)^{\frac{7}{2}}}, \quad \text{etc.}$$

Note that these inverses also exhibit a simple rotation property. Inverse transforms of

$$H_{mn}(\xi,\eta) = \frac{\eta^{2m-1-n}\xi^n}{(\xi^2+\eta^2)^m}, \quad m = 2, 3, \dots, \ n = 1, 2, \tag{A.5}$$

can be found directly by Mathematica:

$$h_{mn}(x,y) = H_{mn}^{-1}(\xi,\eta) = \frac{poly_{2m-1}(x,y)}{(y^2+x^2)^m}, \quad m = 2, 3, \dots, \ n = 1, 2, \tag{A.6}$$

where $poly_{2m-1}(x,y)$ is a polynomial, homogeneous of degree $2m-1$, e.g.,

$$h_{21}(x,y) = \frac{-i}{2}\frac{x(x^2-y^2)}{(y^2+x^2)^2}, \qquad h_{31}(x,y) = \frac{-i}{8}\frac{x(3x^4-6x^2y^2-y^4)}{(y^2+x^2)^3}, \tag{A.7}$$

$$h_{32}(x,y) = \frac{i}{8}\frac{y(3x^4-6x^2y^2-y^4)}{(y^2+x^2)^3}.$$

Acknowledgment. The authors are extremely grateful to an anonymous reviewer whose comments lead to a great improvement in this study.

REFERENCES

D. M. Barnett and J. Lothe (1975), *Line force loadings on anisotropic half-spaces and wedges*, Phys. Norvegica, 8, pp. 13–22.

R. S. Chadwick (2002), *Axisymmetric indentation of a thin incompressible elastic layer*, SIAM J. Appl. Math, 62, pp. 1520–1530.

W. T. Chen (1965), *Stresses in a transversely isotropic elastic cone under an asymmetric force at its vertex*, Z. Angew. Math. Phys., 16, pp. 337–343.

H. D. Conway, K. A. Farnham, and T. C. Ku (1967), *The indentation of a transversely isotropic half-space by a rigid sphere*, J. Appl. Mech. ASME, pp. 491–492.

R. Courant and D. Hilbert (1962), *Methods of Mathematical Physics*, Vol. 1, Interscience, New York.

E. K. Dimitriadis, F. Horkay, M. Maresca, B. Kachar, and R. S. Chadwick (2002), *Determination of elastic moduli of thin layers of soft material using the atomic force microscope*, Biophysical J., 82, pp. 2798–2810.

J. D. Eshelby (1956), *The continuum theory of lattice defects*, Progress in Solid State Physics, 3, pp. 79–144.

J. D. Eshelby (1961), *Elastic inclusions and inhomogeneities*, in Progress in Solid Mechanics, Vol. II, I. N. Sneddon and R. Hills, eds., North–Holland, Amsterdam, pp. 89–139.

A. E. Green and W. Zerna (1968), *Theoretical Elasticity*, Oxford University Press, London.

J. Happel and H. Brenner (1965), *Low Reynolds Number Hydrodynamics*, Prentice–Hall, New York.

L. D. Landau and E. M. Lifshitz (1975), *Theory of Elasticity*, Pergamon Press, Oxford, UK.

S. G. Lekhnitskii (1963), *Theory of Elasticity of an Anisotropic Elastic Body*, Holden–Day, San Francisco.

E. Pan and F. G. Yuan (2000), *Three-dimensional Green's functions in anisotropic bimaterials*, Int. J. Solids Structures, 37, pp. 5329–5351.

K. Portz and A. A. Maradudin (1977), *Surface contribution to the low-temperature specific heat of a cubic crystal*, Phys. Rev. B, 16, pp. 3535–3540.

A. J. M. Spencer (1984), *Constitutive theory for strongly anisotropic solids*, in Continuum Theory of the Mechanics of Fibre-Reinforced Composites, A. J. M. Spencer, ed., Springer-Verlag, New York, pp. 1–32.

S. Timoshenko (1930), *Strength of Materials*, D. van Nostrand, New York.

J. R. Turner (1980), *Contact on a transversely isotropic half space, or between two transversely isotropic bodies*, Int. J. Solids Structures, 16, pp. 409–419.

J. R. Willis (1966), *Hertzian contact of anisotropic bodies*, J. Mech. Phys. Solids, 14, pp. 163–175.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1198–1223

POINT DYNAMICS IN A SINGULAR LIMIT OF THE KELLER–SEGEL MODEL 1: MOTION OF THE CONCENTRATION REGIONS*

J. J. L. VELÁZQUEZ†

Abstract. The purpose of this paper is to study a singular perturbation limit of a Keller–Segel system that generates blow-up in finite time. The main question that is addressed is the description of the evolution of the solutions of this problem beyond the blow-up time for the limit problem if a suitable parameter $\varepsilon > 0$ approaches zero. This problem is studied using matched asymptotic expansions. The resulting limit solution can be described beyond the blow-up time by means of the motion of a set of points whose dynamics is coupled with a parabolic-elliptic system of equations.

Key words. chemotaxis, singular perturbations, Keller–Segel model

AMS subject classifications. 35K45, 35B25, 92B05

DOI. 10.1137/S0036139903433888

1. Introduction.

Chemotactic aggregation has received a lot of attention in recent decades, both from experimental and theoretical viewpoints (cf., for instance, [1, 4, 5, 7, 8, 10, 11, 13, 20, 28, 34, 37, 39, 42, 44, 45]).

A particular model that has been extensively used in several studies of chemotaxis is the well-known Keller–Segel one (cf. [28]). A property of the Keller–Segel system that has been considered often is the fact that for suitable choices of the chemotactic function, solutions of this problem might blow-up in finite time (cf. [3, 17, 18, 24, 25, 26, 27, 32, 33]). More detailed references about recent results concerning blow-up in chemotaxis models can be found in [5, 16, 47] and references therein.

Blow-up is a property that cannot be expected to take place for the magnitudes that describe the behavior of biological or physical systems. Usually, in many problems in applied mathematics blow-up occurs only for some approximation of the real problem, and it indicates the presence not of a real singularity, but rather of a change in the orders of magnitude of the values of some quantity that characterizes the state of a system. For instance, this is a well-established fact in combustion theory, where blow-up just means that the values of the temperature and other physical magnitudes rise several orders of magnitude when ignition takes place (cf. [30]). Blow-up usually takes place in physical or biological models if they are approximations of more realistic models, usually containing some small parameter, say $\varepsilon > 0$, that cannot exhibit singular behaviors unless this parameter is set to zero. Suppose that for $\varepsilon = 0$ the limit problem can develop singularities in finite time. The behavior of the complete model for $\varepsilon > 0$ usually is similar to that of the limit model away from the singularities. However, the features of the problem with $\varepsilon > 0$ but small are usually very different from those of the limit problem near the singularities. The presence of blow-up just indicates that the approximations that led to that simpler model where blow-up takes place are not valid anymore near the singularity and that the whole dynamics of the complete model needs to be taken into account there.

*Received by the editors June 29, 2003; accepted for publication (in revised form) September 11, 2003; published electronically May 5, 2004.
http://www.siam.org/journals/siap/64-4/43388.html
†Departamento de Matemática Aplicada, Facultad de Matemáticas, Universidad Complutense, Madrid 28040, Spain (JJ_velazquez@mat.ucm.es).

The goal of this paper is to study by means of matched asymptotic expansions the behavior of the solutions of a system of partial differential equations of the Keller–Segel type. The problem under consideration will contain a parameter $\varepsilon \geq 0$. If $\varepsilon = 0$ the problem exhibits blow-up in finite time. It will be shown that the original model, in a suitable asymptotic limit, can describe in a rather detailed manner the formation and motion of some regions where the mass of the cells is concentrated. In particular a set of equations describing the motion of these points will be obtained. Since the solutions of the considered model blow-up if the singular perturbation parameter is set to zero, the behavior of the solutions derived here can be considered as some kind of "continuation beyond blow-up" for the solutions of the original system.

The precise problem that will be considered in this paper is the following:

$$\frac{\partial u}{\partial t} = \Delta u - \nabla(G_\varepsilon(u)\nabla v),\ x \in \mathbb{R}^2,\ t > 0, \tag{1.1}$$

$$\Delta v + u = 0,\ x \in \mathbb{R}^2,\ t > 0, \tag{1.2}$$

where u denotes the concentration of the organism and v is the concentration of the chemical secreted by it. In this particular version of the system a term v_t in the right-hand side of (1.2) has been neglected. This assumption is usually made in the study of the Keller–Segel model because diffusion for the organism is usually much slower than diffusion of the chemical that can be assumed to be at equilibrium. We will make the following choice of chemotactic function:

$$G_\varepsilon(u) = \frac{1}{\varepsilon} Q(\varepsilon u), \tag{1.3}$$

where $\varepsilon > 0$ is a small parameter, and the function $Q(\xi)$ is an increasing function satisfying

$$Q(s) = s - \alpha s^2 + \dots \quad \text{as } s \to 0, \tag{1.4}$$

$$Q(s) \sim L \quad \text{as } s \to \infty, \tag{1.5}$$

where $L > 0$ is a given number. A typical example would be $Q(s) = \frac{s}{1+s}$. In other words, instead of assuming that the chemotactic function $G_\varepsilon(u)$ increases without limit as the concentration of the organism becomes high, it will be assumed that the mobility of the organism saturates to a constant value. Another choice of $G_\varepsilon(u)$ certainly would be possible, but we will use this particular one just for the sake of simplicity.

The idea of replacing the linear function u by a function $G_\varepsilon(u)$ preventing cell density collapse has been used in [21]. More complicated models that would avoid blow-up also can be found in [6]. A rigorous study of the continuation of the solutions for the Keller–Segel model beyond the blow-up time in radial cases has been undertaken in [41].

The choice of the function $G_\varepsilon(u)$ in (1.1) is not motivated by some strong biological reason, but it has been basically made on mathematical grounds, to avoid collapse of the solutions.

System (1.1)–(1.5) is a particular case of the Keller–Segel model. Let us remark that for $\varepsilon = 0$ the system (1.1)–(1.5) formally becomes

$$\frac{\partial u}{\partial t} = \Delta u - \nabla(u\nabla v),\ x \in \mathbb{R}^2,\ t > 0, \tag{1.6}$$

$$\Delta v + u = 0,\ x \in \mathbb{R}^2,\ t > 0. \tag{1.7}$$

It is known that solutions of the problem (1.6), (1.7) might blow-up in a finite time $T > 0$ (cf. [9, 17, 27, 32]). As long as u, v remain bounded, the limit from (1.1)–(1.5) to (1.6), (1.7) does not pose any serious mathematical problem. However, the situation becomes mathematically more interesting for times $t > T$, because the solutions of (1.1)–(1.5) are globally defined in time, something that does not occur for the solutions of (1.6), (1.7). It is then natural to ask what happens to the solutions of (1.1)–(1.5) as $\varepsilon \to 0$ and $t > T$. This is the problem that will be addressed in this paper using asymptotic expansions.

Aggregation processes in real biological systems are complex processes involving many diverse ingredients. One of the most extensively studied problems is the aggregation of *Dictyostelium discoideum*. Many characteristics of that aggregation process are not included in simple models such as (1.1)–(1.5); one such characteristic is that the chemical does not propagate in a simple diffusive manner as indicated by (1.2), but rather by means of a sophisticated chemical oscillatory process (cf. [31, 38, 46]). On the other hand, aggregation of *Dictyostelium discoideum* does not occur by means of continuous densities u as occurs in (1.1)–(1.5), but on the contrary cells aggregate in some characteristic streams where cells move towards the center (cf. [38]). In any case, simple continuous models like (1.1)–(1.5) have been extensively used and provide some understanding of some features of the aggregation process in *Dictyostelium discoideum* and other organisms (cf., for instance, [22, 23, 29]).

The plan of this article is as follows. Section 2 describes the structure of some steady states associated to (1.1)–(1.5). Section 3 provides a description of the motion of regions with high densities of u that will be called concentration regions. In section 3, a system of equations that describes the dynamics of such a concentration region is derived. The well-posedness of such a system of equations has been proved in [48]. Finally, in the appendix a rigorous proof of some asymptotic properties of the steady states that have been formally derived in section 2 is provided.

In a second part of this paper (cf. [49]) the relation between the results of the present paper and the blow-up mechanism for (1.6), (1.7) is considered. The second paper studies how the blow-up process for (1.6), (1.7) is stopped for times close enough to the blow-up time if (1.6), (1.7) are replaced by (1.1), (1.2). In particular, it is seen that blow-up for (1.6), (1.7) yields for (1.1), (1.2) the formation of concentration regions that evolve later according to the equations derived in the present paper.

2. Steady states. In this section, as a preliminary step, we study radial steady state solutions of (1.1), (1.2). More precisely we consider the radial solutions of the system

$$\Delta \bar{u} - \nabla(G_\varepsilon(\bar{u})\nabla \bar{v}) = 0, \ x \in \mathbb{R}^2, \tag{2.1}$$

$$\Delta \bar{v} + \bar{u} = 0, \ x \in \mathbb{R}^2, \tag{2.2}$$

where $G_\varepsilon(u)$ is as in (1.3)–(1.5), and where from now on bars will be introduced above the functions that denote steady states. The scaling structure of $G_\varepsilon(u)$ in (1.3) suggests introducing the new set of variables

$$\bar{u} = \frac{1}{\varepsilon}\bar{U}, \tag{2.3}$$

$$x = \sqrt{\varepsilon} y \tag{2.4}$$

that transforms (2.1), (2.2) into

$$\Delta_y \bar{U} - \nabla_y (Q(\bar{U}) \nabla_y \bar{v}) = 0, \ y \in \mathbb{R}^2, \tag{2.5}$$

$$\Delta_y \bar{v} + \bar{U} = 0, \ y \in \mathbb{R}^2. \tag{2.6}$$

Let us write $r = |y|$. The main result obtained in this section is given in the following theorem.

THEOREM 2.1. *Suppose that $Q(\cdot) \in C^1(\mathbb{R}^+)$ is an increasing function satisfying* (1.4). *Let us assume also that $\frac{Q(s)}{s}$ is a decreasing function. Then, for each $M > 8\pi$ there exists a unique radial solution (up to rigid displacements) such that*

$$\int_{\mathbb{R}^2} \bar{U}(y; M) d^2 y = M.$$

The function $\bar{U}(y; M) = \bar{U}(r; M)$ is decreasing on r and its asymptotic behavior as $r \to \infty$ is given by

$$\bar{U}(r; M) \sim \frac{k(M)}{r^{\frac{M}{2\pi}}} \ \text{ as } \ r \to \infty.$$

Proof of Theorem 2.1. For radial solutions, integration of (2.5) in $\mathbb{R}^2$ yields

$$r \frac{\partial \bar{U}}{\partial r} - Q(\bar{U}) r \frac{\partial \bar{v}}{\partial r} = c, \tag{2.7}$$

where c is a real constant to be determined. If we assume that $\bar{U}$ has a finite amount of mass in $\mathbb{R}^2$ it would follow that $c = 0$.

On the other hand, we can write (2.6) for radial solutions as

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \bar{v}}{\partial r} \right) + \bar{U} = 0. \tag{2.8}$$

Plugging (2.7) (with $c = 0$) into (2.8) we obtain

$$\frac{1}{r} \frac{\partial}{\partial r} \left(\frac{r}{Q(\bar{U})} \frac{\partial \bar{U}}{\partial r} \right) + \bar{U} = 0 \tag{2.9}$$

that has to be solved with the additional condition

$$\frac{\partial \bar{U}}{\partial r}(0) = 0. \tag{2.10}$$

Assume that

$$\bar{U}(0) = \lambda, \ \lambda > 0. \tag{2.11}$$

Problem (2.9)–(2.11) can be analyzed using standard ODE theory. We will denote from now on as $\bar{U}(r; \lambda)$ the unique solution of (2.9)–(2.11), although by convenience we will drop the dependence on λ if it is clear from the context. Notice that we can rewrite (2.9) as

$$\frac{1}{Q(\bar{U})} \frac{\partial^2 \bar{U}}{\partial r^2} - \frac{Q'(\bar{U})}{(Q(\bar{U}))^2} \left(\frac{\partial \bar{U}}{\partial r} \right)^2 + \frac{1}{Q(\bar{U})} \frac{\partial \bar{U}}{\partial r} + \bar{U} = 0; \tag{2.12}$$

examining the sign of $\frac{\partial^2 \bar{U}}{\partial r^2}$ at the possible points where $\frac{\partial \bar{U}}{\partial r} = 0$, we see that $\bar{U}$ is decreasing as long as it remains positive. We then have three possibilities. Either $\bar{U}$ vanishes at some $r_0 > 0$, or $\lim_{r\to\infty} \bar{U}(r) = \bar{U}_0 > 0$, or $\lim_{r\to\infty} \bar{U}(r) = 0$. In order to exclude the possibility of $\bar{U}$ vanishing at a finite value $r_0 > 0$ we argue as follows. Since $\bar{U} \le \lambda$ for $0 \le r \le r_0$, by multiplying (2.9) by r and integrating we obtain

$$r\frac{\partial}{\partial r}[S(\bar{U})] \ge -\frac{\lambda r^2}{2}, \tag{2.13}$$

where $S(\bar{U}) = -\int_{\bar{U}}^1 \frac{d\xi}{Q(\xi)}$. Notice that $S(\bar{U}) = \log(\bar{U}) + O(1)$ as $\bar{U} \to 0^+$. Dividing (2.13) by r and integrating in the interval $[0, r]$, we arrive at

$$S(\bar{U}(r)) \ge S(\lambda) - \frac{\lambda r^2}{4}. \tag{2.14}$$

Therefore if $\bar{U}(r_0) = 0$, the left-hand side of (2.14) converges to $-\infty$ as $r \to r_0^+$ and the right-hand side remains bounded, yielding a contradiction.

Suppose then that $\lim_{r\to\infty} \bar{U}(r) = \bar{U}_0 > 0$. Thus for r large enough (2.9) could be approximated as

$$\frac{\partial}{\partial r}\left(\frac{r}{Q(\bar{U}_0)}\frac{\partial \bar{U}}{\partial r}\right) + \bar{U}_0 r = 0.$$

Integrating this equation twice one would obtain the following asymptotics for U to the leading order:

$$\bar{U} \sim -\frac{\bar{U}_0 r^2}{4},$$

but, since this asymptotics implies that $\bar{U}$ changes sign for large r we would derive a contradiction again. Therefore, the only possibility that has been left is $\bar{U}(r) > 0$ for any $r > 0$ and $\lim_{r\to\infty} \bar{U}(r) = 0$.

We can then compute in a more detailed manner the asymptotics of $\bar{U}(r)$ as $r \to \infty$. Notice that to the leading order (2.9) can be written for $r \to \infty$ as

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial(\bar{V})}{\partial r}\right) + e^{\bar{V}} = 0, \tag{2.15}$$

where we have written $\bar{V} = \log(\bar{U})$. Equation (2.15), which is known as Emden's equation and also as Bratu's equation, can be transformed into an autonomous differential equation by means of the change of variables

$$\bar{V} = -2\log(r) + \bar{W},\ r = e^{\xi}$$

that transforms (2.15) into

$$\frac{\partial^2 \bar{W}}{\partial \xi^2} + e^{\bar{W}} = 0. \tag{2.16}$$

This equation can be explicitly solved. Standard analysis of the Hamiltonian system (2.16) shows that its solutions behave asymptotically as

$$\bar{W} \sim -\theta\xi + C + o(1) \ \text{ as } \ \xi \to \infty,$$

where θ, C are real constants, and $\theta > 0$. In the original set of variables this asymptotics reads as

$$\bar{U} \sim \frac{k(\lambda)}{r^{\beta(\lambda)}} \text{ as } r \to \infty, \tag{2.17}$$

where λ is as in (2.11) and $\beta(\lambda) = 2 + \theta > 2$.

It is possible to derive a relation between $\beta(\lambda)$ and the mass of the function $\bar{U}$. Integration of (2.9) in a ball $B_R(0)$ yields

$$\int_{B_R(0)} \bar{U} d^2x = -\int_{\partial B_R(0)} \frac{1}{Q(\bar{U})} \frac{\partial \bar{U}}{\partial r} dS_x. \tag{2.18}$$

Using (1.4) and (2.17) we obtain the asymptotics

$$\frac{1}{Q(\bar{U})} \frac{\partial \bar{U}}{\partial r} \sim -\beta(\lambda) \frac{1}{r} \text{ as } r \to \infty.$$

Therefore, taking the limit $R \to \infty$ in (2.18) we obtain

$$\int_{\mathbb{R}^2} \bar{U} d^2x = 2\pi\beta(\lambda). \tag{2.19}$$

This shows that the coefficient $\beta(\lambda) > 2$ is uniquely determined by the mass of $\bar{U}$.

It turns out that for a rather large class of functions $Q(\xi)$ the function $\beta(\lambda)$ (or, equivalently, the mass of $\bar{U}$) is increasing with respect to λ. For instance, this property is satisfied if the following assumption holds:

$$\frac{Q(s)}{s} \text{ is decreasing with respect to } s. \tag{2.20}$$

For example, the function $Q(s) = \frac{s}{1+s}$ satisfies this assumption. Suppose that (2.20) is satisfied. We define

$$H(\xi, \lambda) \equiv \lambda Q\left(\frac{\xi}{\lambda}\right) = \xi \frac{Q(s)}{s},$$

where $s = \frac{\xi}{\lambda}$. We remark that under the assumption (2.20) $H(\xi, \lambda)$ is increasing on λ for each ξ fixed. If we assume also that $Q(s)$ is differentiable, (2.20) would imply that $Q'(s) - \frac{Q(s)}{s} < 0$ for $s > 0$. Combining this inequality with (1.4), the following inequalities hold:

$$Q(s) < s \text{ for } s > 0, \tag{2.21}$$

$$Q'(s) < 1 \text{ for } s > 0. \tag{2.22}$$

We will assume also that $Q(s)$ is an increasing function. Then

$$Q'(s) > 0 \text{ for } s > 0. \tag{2.23}$$

In order to conclude the proof of Theorem 2.1 we need some information concerning the asymptotics of $M(\lambda)$ for $\lambda \to 0^+$ and $\lambda \to \infty$. The following result will conclude the proof of the theorem.

PROPOSITION 2.2. *Suppose that* (1.3), (2.20) *hold. Let us denote as* $U(r,\lambda)$ *the solution of* (2.9)–(2.11). *Then the function* $M(\lambda) \equiv \int_{\mathbb{R}^2} U(r,\lambda) d^2x$ *is strictly increasing on* λ. *Moreover,*

$$\lim_{\lambda\to 0^+} M(\lambda) = 8\pi, \ \lim_{\lambda\to\infty} M(\lambda) = +\infty. \tag{2.24}$$

A rigorous proof of Proposition 2.2 will be given in the appendix. Using formal asymptotics we will now derive the behavior of the solutions of (2.9)–(2.11) as $\lambda \to 0^+$ and $\lambda \to \infty$ since some of these asymptotic properties will be used later.

In the case $\lambda \to 0^+$ we remark that the monotonicity of $\bar{U}(r;\lambda)$ on r as well as (2.11) shows that $\lim_{\lambda\to 0^+} \bar{U}(r,\lambda) = 0$. We can then approximate to the leading order $Q(\bar{U})$ by $\bar{U}$ in (2.9). More precisely, we introduce new variables φ, ξ by means of

$$\bar{U} = \frac{\lambda}{8}\varphi, \ r = \frac{\sqrt{8}}{\sqrt{\lambda}}\xi. \tag{2.25}$$

The factors $\sqrt{8}$, 8 have been introduced here just to obtain some of the formulae in forthcoming computations normalized in a convenient manner. Plugging (2.25) in (2.9) and taking the limit $\lambda \to 0^+$, (2.9) becomes

$$\frac{1}{\xi}\frac{\partial}{\partial\xi}\left(\frac{\xi}{\varphi_0}\frac{\partial\varphi_0}{\partial\xi}\right) + \varphi_0 = 0, \tag{2.26}$$

where $\varphi_0(\xi;\lambda) = \lim_{\lambda\to 0^+} \varphi(\xi;\lambda)$.

This equation has to be complemented with the boundary conditions (cf. (2.10), (2.11))

$$\varphi_0(0) = 8, \ \frac{\partial\varphi_0}{\partial\xi}(0) = 0. \tag{2.27}$$

The solution of (2.26), (2.27) is given by

$$\varphi_0 = \frac{8}{(1+\xi^2)^2}. \tag{2.28}$$

Then

$$\bar{U}(r;\lambda) \sim \frac{\lambda}{(1+\frac{\lambda r^2}{8})^2} \ \text{ as } \ \lambda \to 0^+ \tag{2.29}$$

whence the first formula from (2.24) follows. We will later need, in the study of the mechanism of the formation of the concentration regions, the first higher correction to (2.29). To this end, we just expand U, or, equivalently, φ in powers of λ. By assumption $Q(s)$ satisfies (1.4). We then look for φ in the form

$$\varphi = \varphi_0 + \frac{\alpha\lambda}{8}\varphi_1 + \dots, \tag{2.30}$$

where φ_0 is given by (2.28) and satisfies (2.26), (2.27). Standard computations show that φ_1 satisfies

$$\frac{1}{\xi}\frac{\partial}{\partial\xi}\left(\frac{\xi}{\varphi_0}\frac{\partial\varphi_1}{\partial\xi}\right) - \frac{1}{\xi}\frac{\partial}{\partial\xi}\left(\frac{\xi\varphi_1}{(\varphi_0)^2}\frac{\partial\varphi_0}{\partial\xi}\right) + \varphi_1 = -\frac{1}{\xi}\frac{\partial}{\partial\xi}\left(\xi\frac{\partial\varphi_0}{\partial\xi}\right), \tag{2.31}$$

$$\varphi_1(0) = 0. \tag{2.32}$$

A particular homogeneous solution of (2.31) can be obtained using the fact that there exists a transformation group of rescalings associated to (2.26). More precisely, given $\varphi_0(\xi)$, a solution of (2.26), it is possible to compute a one-parameter family of solutions given by $\varphi_{0,\lambda}(\xi) = \lambda\varphi_0(\sqrt{\lambda}\xi)$. Differentiating this formula, as well as the equation (2.26) with respect to λ at the particular value $\lambda = 1$, we obtain that the following function solves the homogeneous part of (2.31):

$$\varphi_{1,h}(\xi) = \left[\frac{16}{(1+\xi^2)^3} - \frac{8}{(1+\xi^2)^2}\right]. \tag{2.33}$$

The analysis of (2.31), (2.32) becomes simpler using as variable the mass of φ. This is commonly done in the study of radial solutions of the Keller–Segel model. Let us define

$$m(\xi) = \int_0^\xi \eta\varphi_1(\eta)d\eta. \tag{2.34}$$

Using (2.31) it then follows after some computations that m solves

$$\frac{1}{\varphi_0}\frac{\partial^2 m}{\partial\xi^2} - \left(\frac{\varphi_{0,\xi}}{\varphi_0^2} + \frac{1}{\varphi_0\xi}\right)\frac{\partial m}{\partial\xi} + m = -\xi\varphi_{0,\xi}. \tag{2.35}$$

A solution of the homogeneous equation associated to (2.35) can be obtained by applying the transformation (2.34) to (2.33). After multiplying the resulting solution by a constant due to the linearity of the problem we obtain

$$m_h(\xi) = \frac{\xi^2}{(1+\xi^2)^2}.$$

The solution of (2.35) can then be obtained using the standard variation of constants method. Looking for solutions of (2.35) in the form $m(\xi) = m_h(\xi)f(\xi)$ and using the fact that the boundary condition (2.32) combined with (2.31) implies $m(\xi) = O(\xi^2)$ as $\xi \to 0^+$, we obtain, after some computations,

$$m(\xi) = \frac{16}{3}\frac{\xi^2}{(1+\xi^2)^3}[\xi^4 + 4\xi^2 + 2(\xi^2+1)\log(\xi^2+1)]. \tag{2.36}$$

Using (2.34) and (2.36) it would be possible to compute $\varphi_1(\xi)$. Notice, however, that it will be more relevant for us to compute just the asymptotics of $M(\lambda)$ as $\lambda \to 0^+$. The definition (2.34) implies that the total contribution of $\varphi_1(\xi)$ to the mass $M(\lambda)$ is $\frac{\pi\alpha m(\infty)}{4}$ (cf. (2.30)). Using (2.36) it then follows that

$$M(\lambda) = 8\pi + \frac{4\pi\alpha\lambda}{3} + o(\lambda) \text{ as } \lambda \to 0^+. \tag{2.37}$$

The computed asymptotics (2.29), (2.37) will be useful in describing the formation of regions with high values of u.

Let us remark that the asymptotics of the solutions of (2.9)–(2.10) can be derived using standard asymptotic methods. It turns out that

$$M(\lambda) \sim \frac{2\pi a\lambda\nu_1}{\sqrt{L}} \text{ as } \lambda \to \infty, \tag{2.38}$$

where $a = |\frac{\partial J_0}{\partial r}(\nu_1)| = 0.51914750\ldots$, and ν_1 is the first root of the Bessel function J_0. □

3. Singular points dynamics. In this section we will formally derive a limit problem that describes the evolution of some solutions of (1.1)–(1.5) as $\varepsilon \to 0^+$.

Notice that the analysis of the steady states in section 2 shows the existence of stationary solutions of (1.1)–(1.2) having their mass concentrated in regions of size $\sqrt{\varepsilon}$ and an amount of mass $M \in (8\pi, +\infty)$. Our goal is to describe asymptotically solutions of (1.1)–(1.5) containing finite amounts of mass of u that are concentrated near some particular points $x_j(t)$. These regions will be termed from now on as concentration regions or singular points. In the neighborhoods of these singular points, solutions will be described by means of the steady states described in section 2. Although there are several differences at the technical level some of the ideas used in this section are closely related to the study of the so-called spike dynamics in reaction-diffusion equations that has been studied by several authors in both steady and evolution settings (cf. [2, 35, 36, 40, 50, 51]).

It is interesting to remark also that the evolution laws for the concentration regions have some analogies with the motion of a set of point vortices, a problem that has been thoroughly studied (cf. [43]). The analogies and differences of the resulting limit problem with the problem of vortex dynamics will be discussed in more detail below.

3.1. Global existence of solutions if $\varepsilon > 0$. The solutions of (1.1)–(1.5), in contrast to those of (1.6), (1.7), are globally defined in time in the following proposition.

PROPOSITION 3.1. *For any $u_0(\cdot) \in L^1(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2) \cap C^\alpha(\mathbb{R}^2)$, $u_0 \geq 0$, $0 < \alpha < 1$, for any $Q(\cdot) \in C^1(\mathbb{R}^+)$ satisfying (1.4), (1.5), and for any $\varepsilon > 0$, there exists a unique solution $u(\cdot) \in C([0,\infty], L^\infty(\mathbb{R}^2)) \cap C^\infty((0,\infty) \times \mathbb{R}^2)$ of (1.1)–(1.3) satisfying $u(x,t) = u_0(x)$ with v uniquely determined by means of the formula*

$$v(x,t) = -\frac{1}{2\pi} \int_{\mathbb{R}^2} \log(|x-y|) u(y,t) d^2y. \tag{3.1}$$

Remark 3.2. The choice of a particular v is required because for each given u, the solution of (1.2) is not uniquely determined without suitable growth estimates at infinity for v.

Proof of Proposition 3.1. Local existence can be obtained using classical methods of semigroup theory (cf. [15]). Indeed, (1.1), (1.2) can be rewritten as the fixed point problem

$$u(\cdot) = e^{t\Delta} u_0(\cdot) - \int_0^t e^{(t-s)\Delta} [\nabla (G_\varepsilon(u) \nabla v(\cdot, s))] ds \tag{3.2}$$

with v given by (3.1) and where $e^{t\Delta}$ is the heat semigroup in the whole plane that is given by the formula

$$e^{t\Delta} f(x) = \frac{1}{4\pi t} \int_{\mathbb{R}^2} e^{-\frac{(x-y)^2}{4t}} f(y) d^2y.$$

Classical potential theory (cf. [14]) shows that

$$\|\nabla v(\cdot, s)\|_{L^\infty(\mathbb{R}^2)} \leq C \|u(\cdot, s)\|_{L^\infty(\mathbb{R}^2)}. \tag{3.3}$$

On the other hand, regularity theory for parabolic equations (cf. [12]) yields

$$\left\|\nabla (e^{t\Delta} f(\cdot))\right\|_{L^\infty(\mathbb{R}^2)} \leq \frac{C}{\sqrt{t}} \|f(\cdot)\|_{L^\infty(\mathbb{R}^2)}. \tag{3.4}$$

Estimates (3.3), (3.4) can be used as it is usual in semigroup theory (cf. [15]) to obtain existence and uniqueness of a solution of (3.2) in $C([0,\infty], L^\infty(\mathbb{R}^2) \cap C^\alpha(\mathbb{R}^2))$. By classical regularity theory this solution is a classical solution (cf. [12]).

In order to show global existence notice that as long as u, solution of (1.1)–(1.3), is bounded in $L^\infty(\mathbb{R}^2)$, we have

$$\int_{\mathbb{R}^2} u(x,t)d^2x = \int_{\mathbb{R}^2} u_0(x)d^2x. \tag{3.5}$$

Using potential theory it follows that for this range of times

$$\|\nabla v(\cdot,t)\|_{L^\infty(\mathbb{R}^2)} \le C \int_{\mathbb{R}^2} u_0(x)d^2x. \tag{3.6}$$

On the other hand, using the boundedness Q, Q', as well as (1.2), we obtain

$$|\nabla(G_\varepsilon(u)\nabla v)| \le C\left(|\nabla v| + \frac{1}{\varepsilon}u\right).$$

The global boundedness of u, v (cf. (3.5), (3.6)) as well as regularity theory for parabolic equations then implies

$$\|u(\cdot,t)\|_{L^\infty(\mathbb{R}^2)} \le \frac{C}{\varepsilon}\int_{\mathbb{R}^2} u_0(x)d^2x,\ t>0 \tag{3.7}$$

whence the global existence of a classical solution follows. □

Remark 3.3. Estimate (3.7) is optimal in its ε dependence as will be checked in the asymptotic formulae computed in the next subsection.

3.2. Derivation of the equations of motion of the concentration regions. We now compute the "outer limit" that describes the solutions of (1.1)–(1.5) away from the concentration regions. Assuming for the moment that a given solution can be described by a family of concentration regions placed at some points $x = x_j(t)$, $j = 1, 2, \dots N$, having, respectively, masses $M_j(t)$, we can approximate u in the region $|x - x_j(t)| \gg \sqrt{\varepsilon}$ as

$$u \approx \sum_{j=1}^{N} M_j(t)\delta(x - x_j(t)) + u_{reg}(x,t), \tag{3.8}$$

where $M_j(t) > 8\pi,\ j = 1, 2, \dots N$, and $u_{reg}(x,t)$ is a bounded function that will be described in more detail later. To fix ideas we will assume that $u(\cdot,0) \in L^1(\mathbb{R}^2)$ and that problem (1.1)–(1.5) is solved in $\mathbb{R}^2 \times \mathbb{R}^+$. Moreover, as indicated in Proposition 3.1, we will determine uniquely v, solution of (1.2), by means of (3.1). Using (3.8), (3.1) it would follow that

$$v(x,t) \approx -\frac{1}{2\pi}\sum_{j=1}^{N} M_j(t)\log(|x - x_j(t)|) + v_{reg}(x,t), \tag{3.9}$$

where

$$v_{reg}(x,t) = -\frac{1}{2\pi}\int_{\mathbb{R}^2} \log(|x-y|)u_{reg}(y,t)d^2y. \tag{3.10}$$

By assumption in the outer region, $|x - x_j(t)| \gg \sqrt{\varepsilon}$, $u \ll \varepsilon^{-1}$, and therefore we can use the following approximation (cf. (1.3)):

$$G_\varepsilon(u) \approx u \ \text{ as } \ \varepsilon \to 0^+.$$

Taking into account (1.1) and (3.9) we then obtain the following equation for u_{reg} in the outer region as $\varepsilon \to 0^+$:

$$\frac{\partial u_{reg}}{\partial t} = \Delta u_{reg} + \frac{1}{2\pi} \sum_{j=1}^{N} M_j(t) \frac{(x - x_j(t))}{|x - x_j(t)|^2} \cdot \nabla u_{reg} - \nabla(u_{reg} \nabla v_{reg}), \tag{3.11}$$

where we have approximated u by u_{reg} in the outer region and where v_{reg} is given by (3.10).

It is interesting to remark that near any of the points $x = x_i(t)$, (3.11) can be approximated locally as

$$\frac{\partial u_{reg}}{\partial t} = \Delta u_{reg} + \frac{M_i(t)}{r_i} \cdot \frac{\partial u_{reg}}{\partial r_i} + \text{higher order terms},$$

where $r_i = |x - x_i(t)| \ll 1$. In other words, problem (3.11) behaves locally, near each point $x_i(t)$ as the heat equation in "space dimension" $(2 + \frac{M_i(t)}{2\pi})$. This feature suggests that (3.10), (3.11) complemented with suitable initial data is a well-posed mathematical problem (at least locally in time). This fact will be shown rigorously in [48]. The only crucial information needed at this stage is that problem (3.10), (3.11) admits solutions with bounded u_{reg}.

In order to compute the motion of the singular points, and also to show that the "ansatz" (3.8) leads to reasonable solutions of the original problem (1.1)–(1.5), we need to study in a detailed manner the "inner region" near each point $x = x_i(t)$. To this end we introduce a new set of inner variables:

$$u(x,t) = \frac{1}{\varepsilon} U(\xi, \tau), \tag{3.12}$$

$$x = x_i(t) + \sqrt{\varepsilon}\xi, \ t = \varepsilon\tau. \tag{3.13}$$

Using these variables (1.1)–(1.3) becomes

$$\frac{\partial U}{\partial \tau} - \sqrt{\varepsilon} \dot{x}_i(t) \nabla_\xi U = \Delta_\xi U - \nabla_\xi (Q(U) \nabla_\xi v), \tag{3.14}$$

$$\Delta_\xi v + U = 0. \tag{3.15}$$

We need to complement (3.14), (3.15) with suitable matching conditions. Notice that for bounded u_{reg}, there exists $\lim_{x \to x_i(t)} \nabla_x v_{reg}(x,t) = \nabla_x v_{reg}(x_i(t), t)$ (cf. (3.10)). Using (3.9) and (3.13) we obtain the following outer approximation for ∇v:

$$\nabla_\xi v \approx -\frac{M_i(t)}{2\pi} \cdot \frac{\xi}{|\xi|^2} + A_i(t)\sqrt{\varepsilon} \ \text{ as } \ |\xi| \gg 1, \ |x - x_i(t)| \ll 1, \tag{3.16}$$

where $A_i(t)$ is the vector

$$A_i(t) = -\sum_{j=1,\, j \neq i}^{N} \frac{M_j(t)}{2\pi} \cdot \frac{(x_i(t) - x_j(t))}{|x_i(t) - x_j(t)|^2} + \nabla_x v_{reg}(x_i(t), t). \tag{3.17}$$

The intuitive meaning of (3.16), (3.17) is rather clear. Indeed, notice that $A_i(t)$ is basically the "gravitational field" (that in the context of this problem is a "chemical field") induced by the singular points different from $x_i(t)$ plus a contribution due to the regular part of the concentration of organisms $u_{reg}(x_i(t), t)$.

To further simplify the problem (3.14)–(3.16) we introduce an auxiliary function ψ such that

$$v = \sqrt{\varepsilon} A_i(t) \cdot \xi + \psi.$$

Using this new variable instead of v, problem (3.14)–(3.16) becomes

$$\frac{\partial U}{\partial \tau} = \Delta_\xi U - \nabla_\xi (Q(U) \nabla_\xi \psi) + \sqrt{\varepsilon}[\dot{x}_i(t) \nabla_\xi U - A_i(t) \nabla_\xi (Q(U))], \tag{3.18}$$

$$\Delta_\xi \psi + U = 0, \tag{3.19}$$

$$\nabla_\xi \psi \sim -\frac{M_i(t)}{2\pi} \cdot \frac{\xi}{|\xi|^2} \text{ as } |\xi| \gg 1. \tag{3.20}$$

Notice that in the inner scale the time scale for stabilization of solutions to steady states is of order $t \approx \varepsilon$ (cf. (3.13)). Since these times are much shorter than those expected for having important variations of $z_i(t)$, $A_i(t)$, $M_i(t)$, we will assume that $z_i(t)$, $A_i(t)$, $M_i(t)$ are frozen and U, ψ have reached their equilibrium state in the time scale τ. In other words, we will approximate (3.18)–(3.20) by means of the steady state problem

$$\Delta_\xi U - \nabla_\xi (Q(U) \nabla_\xi \psi) + \sqrt{\varepsilon}[\dot{x}_i(t) \nabla_\xi U - A_i(t) \nabla_\xi (Q(U))] = 0, \tag{3.21}$$

$$\Delta_\xi \psi + U = 0 \tag{3.22}$$

complemented with (3.20). Since, as it will be checked below, the solutions of (3.21), (3.22) are close to the solutions of (2.5), (2.6) it follows that the main assumption that is implicitly made here is that the steady states obtained in section 2 are stable. The energy arguments in section 5 of [47] indicate that this is so.

In the limit $\varepsilon \to 0^+$ we can approximate (3.20)–(3.22) by means of the equation for the steady states described in section 2 (cf. (2.5), (2.6)):

$$\Delta_\xi \bar{U} - \nabla_\xi (Q(\bar{U}) \nabla_\xi \bar{v}) = 0, \tag{3.23}$$

$$\Delta_\xi \bar{v} + \bar{U} = 0, \tag{3.24}$$

$$\nabla_\xi \bar{v} \sim -\frac{M_i(t)}{2\pi} \cdot \frac{\xi}{|\xi|^2} \text{ as } |\xi| \gg 1. \tag{3.25}$$

It has been shown in section 2 that under assumption (2.20) the problem (3.23)–(3.25) admits a unique radial solution for each $M_i(t) > 8\pi$. Therefore we then obtain to the leading order the following approximation for U:

$$U \approx \bar{U}(\xi; \lambda_i(t)), \tag{3.26}$$

where we recall that $\bar{U}(\xi; \lambda)$ is the unique solution of (2.9)–(2.11). Notice that we are assuming that the value of λ_i that uniquely characterizes $\bar{U}$ is not fixed but changes in time. Moreover, we are assuming that this change of λ_i takes place in the time scale t. The fact that this is the right time scale for λ_i will be checked later "a posteriori," verifying the consistency of the derived asymptotics.

Our next goal is to show by means of a perturbative argument that it is possible to obtain solutions close to $(\bar{U}, \bar{\psi})$ and ε small enough for each value of $A_i(t)$ if $\dot{x}_i(t)$ is computed using a suitable compatibility condition. To this end we look for solutions of (3.23)–(3.25) in the form

$$U = \bar{U} + \sqrt{\varepsilon} U_1 + \varepsilon U_2 + \dots , \tag{3.27}$$

$$\psi = \bar{v} + \sqrt{\varepsilon} v_1 + \varepsilon v_2 + \dots . \tag{3.28}$$

Using these expansions and (3.21), (3.22) we obtain

$$\begin{aligned} -\Delta_\xi U_1 + \nabla_\xi (Q'(\bar{U}) \nabla_\xi \bar{v} U_1) + \nabla_\xi (Q(\bar{U}) \nabla_\xi v_1) = \\ \left[\dot{x}_i(t) \nabla_\xi \bar{U} - A_i(t) \nabla_\xi (Q(\bar{U})) \right] \equiv h, \end{aligned} \tag{3.29}$$

$$\Delta_\xi v_1 + U_1 = 0, \tag{3.30}$$

$$\begin{aligned} &- \Delta_\xi U_2 + \nabla_\xi (Q'(\bar{U}) \nabla_\xi \bar{v} U_2) + \nabla_\xi (Q(\bar{U}) \nabla_\xi v_2) \\ &= -\frac{\partial \bar{U}}{\partial \lambda} \frac{d\lambda}{dt}(t) + \nabla_\xi \left(Q'(\bar{U}) \nabla_\xi v_1 + \frac{1}{2} Q''(\bar{U}) \nabla_\xi \bar{v} U_1^2 \right) \\ &\quad + (A_i(t) \nabla_\xi (Q'(\bar{U}) U_1) - \dot{x}_{i,1}(t) \nabla_\xi \bar{U} - \dot{x}_{i,0}(t) \nabla_\xi U_1), \end{aligned} \tag{3.31}$$

$$0 = U_2 + \Delta_\xi v_2. \tag{3.32}$$

We begin analyzing the system (3.29)–(3.30). To this end it is convenient to rewrite this system in a more convenient manner. We define

$$F = U_1 - Q(\bar{U}) v_1. \tag{3.33}$$

Notice that (2.7) (with $c = 0$) implies $\nabla_\xi \bar{U} = Q(\bar{U}) \nabla_\xi \bar{v}$. Taking this into account (3.29), (3.30) become

$$\Delta_\xi F - \nabla_\xi \left(\frac{Q'(\bar{U}) \nabla_\xi \bar{U}}{Q(\bar{U})} F \right) + h = 0. \tag{3.34}$$

On the other hand, (3.30) can be rewritten as

$$\Delta_\xi v_1 + Q(\bar{U}) v_1 + F = 0. \tag{3.35}$$

Differentiating (3.24) with respect to ξ_k, $k = 1, 2$, and using that $\nabla_\xi \bar{U} = Q(\bar{U}) \nabla_\xi \bar{v}$, we obtain that $z_k = \frac{\partial \bar{v}}{\partial \xi_k}$ solves the equation

$$\Delta_\xi z_k + Q(\bar{U}) z_k = 0$$

or, in an equivalent manner, the functions z_i are eigenfunctions associated to a zero eigenvalue of the homogeneous part of (3.35) considering F as a source there. Therefore in order to solve the problem (3.34), (3.35) we need to impose the following compatibility condition:

$$\int_{\mathbb{R}^2} F z_k d^2\xi = \int_{\mathbb{R}^2} F \frac{\partial \bar{v}}{\partial \xi_k} d^2\xi = 0, \ k = 1, 2. \tag{3.36}$$

We need to reformulate the compatibility condition (3.36) in terms of function h in (3.34). To this end we define functions g_k as the solutions of the following equation:

$$\Delta_\xi (g_k) + \frac{Q'(\bar{U}) \nabla_\xi \bar{U}}{Q(\bar{U})} \cdot \nabla_\xi (g_k) = \frac{\partial \bar{v}}{\partial \xi_k}, \ k = 1, 2, \tag{3.37}$$

satisfying

$$g_k(\xi) = O(|\xi|) \text{ as } |\xi| \to 0, \tag{3.38}$$

$$g_k(\xi) = o(1) \text{ as } |\xi| \to \infty. \tag{3.39}$$

The main properties of the functions $g_k(\xi)$ are contained in the result that follows in Lemma 3.4.

LEMMA 3.4. *Problem* (3.37)–(3.39) *is uniquely solvable. The corresponding solutions have the form* $g_k(\xi) = \bar{g}(r)\frac{\xi_k}{|\xi|}$, *where* $r = |\xi|$ *and* $\bar{g}(r) > 0$ *for* $r > 0$ *and satisfies*

$$\bar{g}(r) = O(r) \text{ as } r \to 0, \tag{3.40}$$

$$\bar{g}(r) = o(1) \text{ as } r \to \infty. \tag{3.41}$$

Proof. Taking into account (3.37) it follows that $\bar{g}(r)$ solves

$$\bar{g}''(r) + \frac{1}{r}\bar{g}'(r) - \frac{\bar{g}(r)}{r^2} + \frac{Q'(\bar{U}(r))\bar{U}'(r)}{Q(\bar{U}(r))} \cdot \bar{g}'(r) = \bar{v}'(r). \tag{3.42}$$

Taking into account that (3.42) is a first order equation for $\bar{g}'(r)$ it follows that there exists at least one solution of (3.42) such that $\bar{g}(r) = O(r)$ as $r \to 0$ and $\bar{g}(r) = o(1)$ as $r \to \infty$. Uniqueness follows by means of a standard maximum principle argument. Uniqueness for (3.37)–(3.39) follows also by the maximum principle. Moreover, since $\bar{v}'(r) < 0$ it follows that $\bar{g}(r)$ solution of (3.42) is positive, since otherwise, due to the asymptotics of $\bar{g}$ as $r \to 0$ and $r \to \infty$, it would follow that $\bar{g}$ would have a negative minimum, but this is impossible due to the maximum principle argument. This concludes the proof of the lemma. ∎

Taking into account that the decay of g_i and their smoothness at the origin provide integrability for the integrals appearing in (3.36), we obtain that

$$\begin{aligned}\int_{\mathbb{R}^2} h g_k d^2\xi &= \int_{\mathbb{R}^2} \left[\Delta_\xi F - \nabla_\xi \left(\frac{Q'(\bar{U})\nabla_\xi \bar{U}}{Q(\bar{U})} F\right)\right] g_k d^2\xi \\ &= \int_{\mathbb{R}^2} F\left[\Delta_\xi(g_k) + \frac{Q'(\bar{U})\nabla_\xi \bar{U}}{Q(\bar{U})} \cdot \nabla_\xi(g_k)\right] d^2\xi \\ &= \int_{\mathbb{R}^2} F \frac{\partial \bar{v}}{\partial \xi_i} d^2\xi.\end{aligned}$$

The compatibility condition (3.36) then implies

$$\int_{\mathbb{R}^2} h g_k d^2\xi = 0, \ k = 1, 2. \tag{3.43}$$

Using (3.29) and (3.43) we obtain the following formula for $\dot{x}_i(t)$ to the leading order as $\varepsilon \to 0^+$:

$$\dot{x}_i(t) \int_{\mathbb{R}^2} g_k \nabla_\xi \bar{U} d^2\xi = A_i(t) \int_{\mathbb{R}^2} g_k \nabla_\xi (Q(\bar{U})) d^2\xi, \ k = 1, 2.$$

Since the angular dependence of g_k and $\frac{\partial \bar{U}}{\partial \xi_k}$ is the same, it follows that

$$\dot{x}_i(t) = \Gamma(M_i) A_i(t), \tag{3.44}$$

where we have defined

$$\Gamma(M) \equiv \frac{\int_{\mathbb{R}^2} g_1(\xi,\lambda(M))\frac{\partial}{\partial\xi_1}(Q(\bar U(\xi,\lambda(M))))d^2\xi}{\int_{\mathbb{R}^2} g_1(\xi,\lambda(M))\frac{\partial\bar U}{\partial\xi_1}(\xi,\lambda(M))d^2\xi},\ M>8\pi, \tag{3.45}$$

where $\lambda=\lambda(M)$ is the only value of λ associated to a given value of the mass $M>8\pi$ (cf. (2.11) and Proposition 2.2).

Relation (3.44) provides the law of motion for the concentration regions placed at $x=x_i(t)$, $i=1,\dots,N$.

For further reference we derive some estimates for $\Gamma(M)$ that will be included in the following lemma.

LEMMA 3.5. *The function* $\Gamma(M)$ *defined by* (3.45) *has the following properties:*

$$0<\Gamma(M)<1, \tag{3.46}$$

$$\lim_{M\to 8\pi^+}\Gamma(M)=1, \tag{3.47}$$

$$\lim_{M\to\infty}\Gamma(M)=0. \tag{3.48}$$

Proof. Using (3.45) we obtain

$$\Gamma(M)=\frac{\int_{\mathbb{R}^2} g_1 Q'(\bar U)\frac{\partial\bar U}{\partial\xi_1}d^2\xi}{\int_{\mathbb{R}^2} g_1\frac{\partial\bar U}{\partial\xi_1}d^2\xi}.$$

As indicated before, g_1 and $\frac{\partial\bar U}{\partial\xi_1}$ have a similar angular dependence $\cos(\theta)$. Moreover, as indicated in Lemma 3.4, g_1 can be written as $g_1=\bar g_1(|\xi|)\frac{\xi_1}{|\xi|}$ with $\bar g_1$ having a constant sign. Moreover, $\frac{\partial\bar U}{\partial\xi_1}=\bar U'(|\xi|)\frac{\xi_1}{|\xi|}$. Since $\bar U$ is decreasing on $|\xi|$ we obtain that $\frac{g_1\frac{\partial\bar U}{\partial\xi_1}}{\int_{\mathbb{R}^2} g_1\frac{\partial\bar U}{\partial\xi_1}d^2\xi}>0$. Therefore (2.22), (2.23), and (3.45) yield (3.46). The asymptotics (3.47) and (3.48) are a consequence of the asymptotics of $\bar U$ as $\lambda\to 0$ and $\lambda\to\infty$ as well as (1.4), (1.5) and the regularity properties of $Q(s)$. ∎

It is interesting to notice that the monotonicity of $\Gamma(M)$ on M as well as (3.44) implies that for a given value of $A_i(t)$ its effect on the velocity of the singular point is smaller. In other words, it is harder to move larger singular points, as could be expected in an intuitive manner.

We remark that the expansion (3.27), (3.28) loses its validity for $|\xi|$ large. Indeed (2.17), (2.19) imply $\bar U\sim\frac{\bar K(M_i)}{|\xi|^{\frac{M_i}{2\pi}}}$ as $|\xi|\gg 1$. Therefore for large $|\xi|$ we can approximate (3.34) as

$$\Delta_\xi F+\frac{M_i}{2\pi}\cdot\frac{\xi}{|\xi|^2}\nabla_\xi F=\frac{(\dot x_i(t)-A_i(t))M_i}{2\pi}\cdot\frac{\xi}{|\xi|^{\frac{M_i}{2\pi}+2}}.$$

The asymptotics for the solutions of this problem is given by

$$F\sim\frac{(\dot x_i(t)-A_i(t))\bar K(M_i)}{|\xi|^{\frac{M_i}{2\pi}-1}}\cdot\frac{\xi}{|\xi|}\ \text{ as }\ |\xi|\gg 1. \tag{3.49}$$

On the other hand, the leading terms in (3.35) as $|\xi|\to\infty$ are $\Delta_\xi v_1+F=0$. Using (3.49), as well as the fact that the angular dependence of v_1 is the same as that of $\dot x_i(t)\cdot\frac{\xi}{|\xi|}$, it follows that $v_1=O(\frac{1}{|\xi|})$ as $|\xi|\to\infty$. Therefore $U_1\sim F$ as $|\xi|\to\infty$. Taking into account (2.17) it then follows that the asymptotics (3.27) breaks down at distances $|\xi|\sim\frac{1}{\sqrt{\varepsilon}}$ or, equivalently, for $|x-x_i(t)|\approx 1$.

3.3. Computation of the rate of change of the concentration regions. Notice that the set of equations (3.10), (3.11), (3.44), (3.45) describes a set of equations for the concentration regions coupled with the external fields u_{reg}, v_{reg} that has to be completed with the evolution law for the mass of the concentration regions $M_i(t)$. Computing the evolution law for $M_i(t)$ is equivalent to computing that of $\lambda_i(t)$ (cf. (3.26)). This can be done by means of a matching condition between the outer contribution due to U_2 (cf. (3.27)) and the inner contribution due to u_{reg}. Although this detailed computation is interesting, in order to check the consistency of the obtained asymptotics we will compute the rate of change of $M_i(t)$ by means of a simpler argument in Lemma 3.6 that consists of directly computing the derivative of $M_i(t)$ using the asymptotics already computed.

LEMMA 3.6. *For the formal solutions of* (1.1)–(1.5) *that behave asymptotically as in* (3.8) *the rate of change of the mass of the concentration regions is given by*

$$\frac{dM_i(t)}{dt} = u_{reg}(x_i(t),t)M_i(t),\ i=1,\dots,N. \tag{3.50}$$

The time scale for stabilization of u in regions close to the concentration region is, up to logarithmic terms, of order ε. This is much shorter than the time scale associated to the macroscopic evolution of the system that is of order one. Therefore, in order to compute the amount of mass lost by $u_{reg}(x,t)$ in a neighborhood of each concentration region we can assume that it can be described by a steady state. We then fix $\delta>0$ small, and integrate (3.11) in a ball $B_\delta(x_i(t))$. After some integrations by parts we obtain

$$\begin{aligned}\frac{d}{dt}&\left(\int_{\mathcal{U}\setminus B_\delta(x_i(t))} u_{reg}d^2x\right)\\ &= -\int_{\partial B_\delta(x_i(t))}\frac{\partial u_{reg}}{\partial n}dS_x - \frac{1}{2\pi}\sum_{j=1}^N M_j(t)\int_{\partial B_\delta(x_i(t))} u_{reg}\frac{(x-x_j(t))\cdot n}{|x-x_j(t)|^2}dS_x\\ &\quad+\int_{\partial B_\delta(x_i(t))} u_{reg}\frac{\partial v_{reg}}{\partial n}dS_x + \int_{\partial\mathcal{U}} f(x,t)dS_x,\end{aligned} \tag{3.51}$$

where $\mathcal{U}$ is a domain whose size is of order one containing only the concentration region $x_i(t)$ and $f(x,t)$ are the fluxes of u away from $\mathcal{U}$. Notice that the last term on the right-hand side of (3.51) does not contribute to the variation of the mass of the concentration region in $x=x_i(t)$. On the other hand, the regularity properties of u_{reg}, v_{reg} (cf. [48]), show that in the limit $\delta\to 0$, the contribution of the first and third term on the right-hand side of (3.51) approaches zero. The same occurs with the contributions in the second term, except for the one due to $j=i$. Then, taking into account that the flux of u lost for u_{reg} yields an increase in $M_i(t)$, we obtain

$$\frac{dM_i(t)}{dt} = \frac{M_i(t)}{2\pi}\lim_{\delta\to 0}\left(\int_{\partial B_\delta(x_i(t))} u_{reg}(x,t)\frac{(x-x_i(t))\cdot n}{|x-x_i(t)|^2}dS_x\right),$$

and taking into account that $\int_{\partial B_\delta(x_i(t))}\frac{(x-x_i(t))\cdot n}{|x-x_i(t)|^2}dS_x = 2\pi$, we arrive at (3.50). ∎

Notice that since $\frac{dM_i(t)}{dt}\geq 0$, the mass of each concentration region is an increasing function, as could be expected on account of the attractive character for the organisms.

We summarize here the main result concerning the motion of the concentration regions as a formal theorem, with the understanding that the proof has been obtained only at the level of formal asymptotics.

THEOREM 3.7 (formal). *Suppose that u solves* (1.1)–(1.5). *It is possible to obtain formal asymptotic expansions for the solutions of* (1.1)–(1.5) *that in the limit $\varepsilon \to 0$ behave asymptotically as in* (3.8), *where the dynamics of u_{reg}, $x_i(t)$, $M_i(t)$ is given by the following set of equations:*

$$\frac{\partial u_{reg}}{\partial t} = \Delta u_{reg} + \frac{1}{2\pi}\sum_{j=1}^{N} M_j(t)\frac{(x - x_j(t))}{|x - x_j(t)|^2}\cdot \nabla u_{reg} - \nabla(u_{reg}\nabla v_{reg}), \tag{3.52}$$

$$v_{reg}(x,t) = -\frac{1}{2\pi}\int_{\mathbb{R}^2} \log(|x-y|)u_{reg}(y,t)d^2y, \tag{3.53}$$

$$\dot{x}_i(t) = \Gamma(M_i(t))A_i(t),\ i = 1,\dots,N, \tag{3.54}$$

$$A_i(t) = -\sum_{j=1,\, j\neq i}^{N} \frac{M_j(t)}{2\pi}\cdot\frac{(x_i(t) - x_j(t))}{|x_i(t) - x_j(t)|^2} + \nabla_x v_{reg}(x_i(t),t),\ i = 1,\dots,N, \tag{3.55}$$

$$\frac{dM_i(t)}{dt} = u_{reg}(x_i(t),t)M_i(t),\ i = 1,\dots,N, \tag{3.56}$$

where function $\Gamma(M)$ is given in (3.45).

Remark 3.8. Notice that all the dependence of the limit problem in the particular shape of the nonlinearity $G_\varepsilon(u)$ is contained in $\Gamma(M)$.

Remark 3.9. As indicated below there are some analogies between (3.52)–(3.56) and the equations for the evolution of a set of point vortices (cf. [43]). In both cases a set of points moves with a velocity that can be obtained by adding the contributions of the other points. In both cases the interaction between the points decays according to the law $\frac{1}{r}$. A first difference is that in the case of the concentration regions obtained above the interaction between points follows a Newtonian gravitational law. On the contrary, in the case of vortices this interaction law should be replaced by the formula of the velocity induced by a fluid at a given point by a vortex line, which must be computed using Biot–Savart's law. In (3.52)–(3.56) there is also a background field u_{reg} interacting with the concentration regions. It is possible to also imagine situations in fluid mechanics where a background vorticity field could interact with a set of point vortices. There are, of course, several differences between both problems, but also enough analogies to suggest that perhaps some of the methods of analysis used in the theory of point vortices could have some usefulness for the study of (3.52)–(3.56).

Remark 3.10. The set of equations (3.52)–(3.56) is the main result of section 3. This problem defines an evolution problem for the points $x = x_i(t)$, whose dynamics is coupled with that of the fields u_{reg}, v_{reg}. Due to the fact that the domain where the problem has to be solved is not prescribed, but is part of the problem to be solved, the mathematical well-posedness of this system of equations is not standard. A rigorous analysis of the well-posedness of this problem can be found in [48]. This well-posedness result can be expected only locally in time because there are at least two different ways in which the solution of the problem (3.52)–(3.56) might develop singularities in a finite time. First, solutions could have additional blow-ups analogous to those considered in [17, 18, 19] at points where u, v are bounded for previous times. On the other hand, concentration regions could coalesce and merge in a finite time. Indeed, if two concentration regions at positions $x_k(t), x_\ell(t)$ are close enough, their

motion could be approximated by means of the system of ODEs (cf. (3.55))

$$\dot{x}_k(t) = -\frac{\Gamma(M_k(t))M_\ell(t)}{2\pi} \cdot \frac{(x_k(t) - x_\ell(t))}{|x_k(t) - x_\ell(t)|^2},$$
$$\dot{x}_\ell(t) = -\frac{\Gamma(M_\ell(t))M_k(t)}{2\pi} \cdot \frac{(x_\ell(t) - x_k(t))}{|x_\ell(t) - x_k(t)|^2}.$$

Therefore

$$\frac{d}{dt}(x_k(t) - x_\ell(t)) = -\frac{(\Gamma(M_\ell(t))M_k(t) + \Gamma(M_\ell(t))M_k(t))}{2\pi} \cdot \frac{(x_\ell(t) - x_k(t))}{|x_\ell(t) - x_k(t)|^2}.$$

Since the coalescence process is rather fast, the masses of the concentration regions can be considered to be constant during the process. If we denote the coalescence time as t^*, then it follows that

$$|x_\ell(t) - x_k(t)| \sim \sqrt{C(t^* - t)} \text{ as } t \to t^*, \tag{3.57}$$

where $C = \frac{(\Gamma(M_\ell(t^*))M_k(t^*) + \Gamma(M_\ell(t^*))M_k(t^*))}{4\pi}$. It then follows that the velocities of the concentration regions would become singular at $t = t^*$. One could also imagine coalescence of more than two concentration regions simultaneously, but such events should be "nongeneric." Clearly, in the original problem (1.1)–(1.5) functions u, v would have a complicated shape during this process that will not be considered here. In particular this analysis does not establish in a completely rigorous manner that the solutions of (1.1)–(1.5) would collapse as indicated in (3.57) if $\varepsilon > 0$. In a strict sense (3.57) holds only for the limit equations (3.52)–(3.56). In order to obtain rigorous coalescence results for the solutions (1.1)–(1.5) a careful analysis of the dynamics of two merging concentration regions should be made. It is unlikely that this could be achieved by means of explicit analytic formulae. In particular the quasi-steady approximation (3.21)–(3.22) would not hold during the whole coalescence process.

3.4. Analysis of a boundary layer in the inner region. In order to conclude this section we will verify that due to (3.56) (cf. (3.50)) it is possible to match the inner expansion (3.27) with the outer contribution of u as $x \to x_i(t)$ that is just given by $u_{reg}(x_i(t), t)$. This will be done to check the consistency of the computed asymptotics. In order to avoid tedious computations just a sketch of the main arguments will be given. We recall that (U_2, v_2) solves the system (3.31), (3.32). We introduce a new variable as

$$F_2 = U_2 - Q(\bar{U})v_2. \tag{3.58}$$

Then (3.31) becomes

$$\begin{aligned} &-\Delta_\xi F_2 + \nabla_\xi(Q'(\bar{U})\nabla_\xi \bar{v} F_2) \\ &= -\frac{\partial \bar{U}}{\partial \lambda}\frac{d\lambda}{dt}(t) + \nabla_\xi\left(Q'(\bar{U})\nabla_\xi v_1 + \frac{1}{2}Q''(\bar{U})\nabla_\xi \bar{v} U_1^2\right) \\ &-\dot{x}_{i,1}(t)\nabla_\xi \bar{U} + (A_i(t)\nabla_\xi(Q'(\bar{U})U_1) - \dot{x}_{i,0}(t)\nabla_\xi U_1). \end{aligned} \tag{3.59}$$

Let us denote as θ the angle that makes a given direction with the direction of $A_i(t)$. In another words, given a vector ξ, we define θ by means of

$$A_i(t) \cdot \frac{\xi}{|\xi|} = |A_i(t)| \cos(\theta). \tag{3.60}$$

Taking into account the invariance under rotations of (3.29) and (3.30), and also that the term h in (3.29) has the form $h(\xi,t)=\tilde{h}(r,t)\cos(\theta)$, where $r=|\xi|$, it follows, using separation of variables, that U_1, v_1 have the particular form

$$U_1=\tilde{U}_1(r)\cos(\theta),\ v_1=\tilde{v}_1(r)\cos(\theta), \tag{3.61}$$

where $\tilde{U}_1(r)$, $\tilde{v}_1(r)$ solve a system of ordinary differential equations whose precise form will not be needed here, and where from now on, by simplicity, we will not write explicitly the dependence of the different functions on t,λ. The only relevant information about $\tilde{U}_1(r)$, $\tilde{v}_1(r)$ that we will need is that both functions are bounded in compact sets, and that, due to (3.30), (3.33), and (3.49), they satisfy the inequalities

$$\tilde{U}_1=O\left(\frac{1}{r^{\frac{M_i}{2\pi}-1}}\right),\ \tilde{v}_1=O\left(\frac{1}{r}\right)\ \text{ as }\ r\to\infty. \tag{3.62}$$

We now proceed to simplify the different terms on the right-hand side of (3.59). First, notice that $\bar{U}=\bar{U}(r;\lambda)$ is radial for each λ. Therefore $\frac{\partial\bar{U}}{\partial\lambda}$ is a radial function. On the other hand, using the formula $\cos^2(\theta)=\frac{1+\cos(2\theta)}{2}$, as well as the formulae for the gradient and the divergence of a function in polar coordinates, it can be seen after some computations that

$$\nabla_\xi\left(Q'(\bar{U})\nabla_\xi v_1+\frac{1}{2}Q''(\bar{U})\nabla_\xi\bar{v}U_1^2\right)=\frac{1}{2r}\frac{\partial}{\partial r}(rf_0(r))+f_1(r)\cos(2\theta)\\ +f_2(r)\sin(2\theta),$$

where f_2, f_3 are bounded functions that decay fast enough as $r\to\infty$ and $f_0(r)$ is given by

$$f_0(r)=Q'(\bar{U}(r))\tilde{U}_1(r)\tilde{v}_1'(r)+\frac{1}{2}Q''(\bar{U}(r))\bar{v}'(r)(\tilde{U}_1(r))^2. \tag{3.63}$$

The term $-\dot{x}_{i,1}(t)\nabla_\xi\bar{U}$ gives a contribution with the angular dependence $\cos(\theta)$. The contribution of this term can be analyzed exactly as it was analyzed for the terms U_1, v_1. Finally, using (3.54) and (3.60), as well as the formulae $\cos^2(\theta)=\frac{1+\cos(2\theta)}{2}$, $\sin^2(\theta)=\frac{1-\cos(2\theta)}{2}$, it follows after some computations that the last term in (3.59) can be written as

$$(A_i(t)\nabla_\xi(Q'(\bar{U})U_1)-\dot{x}_{i,0}(t)\nabla_\xi U_1)=\frac{1}{r}\frac{\partial}{\partial r}(rg_0(r))+g_1(r)\cos(2\theta),$$

where $g_0(r)$, $g_1(r)$ are bounded functions, $g_1(r)$ decays fast enough as $r\to\infty$, and $g_0(r)$ satisfies

$$g_0(r)=\frac{|A_i(t)|}{2}[\Gamma(M_i)\tilde{U}_1(r)-Q'(\bar{U}(r))\tilde{U}_1(r)]=O\left(\frac{1}{r^{\frac{M_i}{2\pi}-1}}\right)\ \text{ as }\ r\to\infty. \tag{3.64}$$

Function F_2 can be decomposed as $F_2=\tilde{F}_{2,0}(r)+\tilde{F}_{2,1}(r)\cos(\theta)+\tilde{F}_{2,2,1}(r)\cos(2\theta)+\tilde{F}_{2,2,2}(r)\sin(2\theta)$. The main contribution to U_2 as $r\to\infty$ is due to the radial term $\tilde{F}_{2,0}(r)$ that solves the ODE

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\tilde{F}_{2,0}}{\partial r}\right)-\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{Q'(\bar{U})\bar{U}'}{Q(\bar{U})}\tilde{F}_{2,0}\right)\\ +\frac{1}{r}\frac{\partial}{\partial r}(r(g_0(r)+f_0(r)))=\frac{\partial\bar{U}}{\partial\lambda}\dot{\lambda}_i(t).$$

We can write $\tilde{F}_{2,0} = \tilde{F}_{2,0,1} + \tilde{F}_{2,0,2}$, where $\tilde{F}_{2,0,1}$, $\tilde{F}_{2,0,2}$ solve, respectively, the equations

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \tilde{F}_{2,0,1}}{\partial r}\right) - \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{Q'(\bar{U})\bar{U}'}{Q(\bar{U})}\tilde{F}_{2,0,1}\right) + \frac{1}{r}\frac{\partial}{\partial r}(r(g_0(r)+f_0(r))) = 0 \tag{3.65}$$

and

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \tilde{F}_{2,0,2}}{\partial r}\right) - \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{Q'(\bar{U})\bar{U}'}{Q(\bar{U})}\tilde{F}_{2,0,2}\right) = \frac{\partial \bar{U}}{\partial \lambda}\dot{\lambda}_i(t). \tag{3.66}$$

Integrating (3.65) once it is possible to find $\tilde{F}_{2,0,1}$ in an elementary manner. After some elementary computations we obtain

$$\tilde{F}_{2,0,1} = aQ(\bar{U}) - Q(\bar{U}(r)) \cdot \int_0^r \frac{[g_0(\eta)+f_0(\eta)]}{Q(\bar{U}(\eta))}d\eta,$$

where a is a real constant. Using (1.4), (2.17), (3.63), (3.64) and the fact that $\frac{M_i}{2\pi} > 4$, it follows that $\lim_{r\to\infty} \tilde{F}_{2,0,1} = 0$. It remains only to compute the contribution due to $\tilde{F}_{2,0,2}$. To this end, notice that after multiplying (3.66) by r, integrating, and after solving the resulting differential equation we arrive at

$$\tilde{F}_{2,0,2} = \dot{\lambda}_i(t)Q(\bar{U}(r))\int_0^r \left[\int_0^\eta \xi\frac{\partial \bar{U}}{\partial \lambda}(\xi)d\xi\right]\frac{d\eta}{\eta Q(\bar{U}(\eta))} + bQ(\bar{U}(r)),$$

where b is a real number. Using (1.4), (2.17) it follows that

$$\begin{aligned}\lim_{r\to\infty}\tilde{F}_{2,0,2} &= \left[\dot{\lambda}_i(t)\int_0^\infty s\frac{\partial \bar{U}}{\partial \lambda}(s)ds\right]\lim_{r\to\infty}\left[Q(\bar{U}(r))\int_0^r \frac{d\eta}{\eta Q(\bar{U}(\eta))}\right] \\ &= \frac{2\pi\dot{\lambda}_i(t)}{M_i(t)}\int_0^\infty s\frac{\partial \bar{U}}{\partial \lambda}(s)ds.\end{aligned}$$

To compute $\int_0^\infty s\frac{\partial \bar{U}}{\partial \lambda}(s)ds$ we differentiate (2.9) with respect to λ. Multiplying by r and integrating on $r > 0$, we then obtain

$$\begin{aligned}\int_0^\infty s\frac{\partial \bar{U}}{\partial \lambda}(s)ds &= \lim_{r\to\infty}\left[\frac{r}{\bar{U}^2}\frac{\partial \bar{U}}{\partial \lambda}\frac{\partial \bar{U}}{\partial r} - \frac{r}{\bar{U}}\frac{\partial}{\partial r}\left(\frac{\partial \bar{U}}{\partial \lambda}\right)\right] \\ &= -\lim_{r\to\infty}\left[r\frac{\partial}{\partial r}\left(\frac{\partial(\log(\bar{U}))}{\partial \lambda}\right)\right] = \frac{M_i'(\lambda)}{2\pi}.\end{aligned}$$

Therefore

$$\lim_{r\to\infty}\tilde{F}_{2,0,2} = \frac{1}{M_i(t)}\frac{dM_i(t)}{dt}.$$

Then using (1.4), (2.17), (3.32), (3.58) as well as the fact that the only term in F_2 that yields a nonzero contribution in the matching condition is $\tilde{F}_{2,0,2}$ we obtain the matching condition

$$U_2 \sim \frac{1}{M_i(t)}\frac{dM_i(t)}{dt}$$

as $|\xi| \gg 1$, $|x - x_i(t)| \ll 1$. Taking into account (2.17), (3.12), (3.27), and (3.49) we obtain the matching condition

$$u_{inner} \sim \frac{1}{M_i(t)} \frac{dM_i(t)}{dt},$$

as $|\xi| \gg 1$, $|x - x_i(t)| \ll 1$. Assuming, as we would expect, that u_{inner} coincides with $u_{reg}(x_i(t), t)$ that provides the outer contribution to u as $\varepsilon \to 0^+$, we then obtain the equation

$$u_{reg}(x_i(t), t) = \frac{1}{M_i(t)} \frac{dM_i(t)}{dt}$$

that is equivalent to (3.56). Therefore the consistency of the computed asymptotics follows.

4. Concluding remarks. In this paper a system of equations for the dynamics of a set of points whose evolution is coupled with a parabolic-elliptic system has been derived (cf. (3.52)–(3.56)). This system has been obtained by taking as the starting point a singular perturbation problem for a system of equations of the Keller–Segel type (cf. (1.6), (1.7)). The manner in which these singular points can develop in a finite time by means of a blow-up mechanism for the solutions of the limit problem has also been described. The significance of the derived results is that they provide a natural way of continuing the solutions of the limit problem (1.1), (1.2) after Dirac mass formation takes place for this model.

5. Appendix: Proof of Proposition 2.2. In this appendix we prove Proposition 2.2. Let us define $F(r) = \int_0^r \xi U(\xi) d\xi$. Using (2.9) we obtain

$$F_{rr} - \frac{F_r}{r} + Q\left(\frac{F_r}{r}\right) F = 0, \tag{5.1}$$

and (2.11) implies

$$F(r) \sim \frac{\lambda r^2}{2} \quad \text{as} \quad r \to 0^+. \tag{5.2}$$

We will denote as $F(r)$ the solution of (5.1), (5.2), although occasionally the notation $F(r, \lambda)$ will be used if the dependence on λ plays a crucial role.

Let us introduce the new variable

$$r = e^{\xi - \frac{1}{2}\log(\lambda)}. \tag{5.3}$$

Using ξ as the new variable, (5.1) becomes

$$F_{\xi\xi} - 2F_\xi + \frac{e^{2\xi}}{\lambda} Q(\lambda e^{-2\xi} F_\xi) F = 0, \tag{5.4}$$

or, in an equivalent manner,

$$F_\xi = V, \tag{5.5}$$

$$V_\xi = 2V - \frac{e^{2\xi}}{\lambda} Q(\lambda e^{-2\xi} V) F, \tag{5.6}$$

where (5.2) implies the condition

$$F \sim \frac{e^{2\xi}}{2},\ V \sim e^{2\xi} \ \text{ as } \ \xi \to -\infty. \tag{5.7}$$

Notice that assumption (2.20) implies

$$\frac{e^{2\xi}}{\lambda} Q(\lambda e^{-2\xi} V) < \lim_{\theta \to \infty} \left[\theta Q\left(\frac{V}{\theta}\right) \right] = V. \tag{5.8}$$

Let us denote as $(\bar{F}, \bar{V})$ the solution of

$$\bar{F}_\xi = \bar{V}, \tag{5.9}$$

$$\bar{V}_\xi = 2\bar{V} - \bar{V}\bar{F} \tag{5.10}$$

satisfying (5.7). Function $(\bar{F}, \bar{V})$ is easily computed and is given by

$$\bar{F} = \frac{4}{(1 + 8e^{-2\xi})},\ \bar{V} = \frac{64e^{-2\xi}}{(1 + 8e^{-2\xi})^2},\ \bar{V} = 2\bar{F} - \frac{(\bar{F})^2}{2}.$$

On the other hand, using (5.8) it follows, arguing by comparison, that (F, V) satisfies

$$V > 2F - \frac{F^2}{2},\ F > 0. \tag{5.11}$$

Actually, a similar argument can be applied to any two values of λ. Given $\lambda_1 < \lambda_2$, let us consider the corresponding solutions of (5.5)–(5.7) and let us denote them as (F_1, V_1), (F_2, V_2), respectively. Using (5.1), (5.2) it follows that

$$F_i(r) \sim \frac{\lambda_i r^2}{2} - \frac{Q(\lambda_i)\lambda_i}{16} r^4 + \dots \ \text{ as } \ r \to 0, \ \ i = 1, 2,$$

or, equivalently, using (5.3),

$$F_i \sim \frac{e^{2\xi}}{2} - \frac{Q(\lambda_i)}{16\lambda_i} e^{4\xi} \ \text{ as } \ \xi \to -\infty,\ i = 1, 2, \tag{5.12}$$

$$V_i \sim e^{2\xi} - \frac{Q(\lambda_i)}{4\lambda_i} e^{4\xi} \ \text{ as } \ \xi \to -\infty,\ i = 1, 2. \tag{5.13}$$

Therefore

$$V_i = 2F_i - \frac{Q(\lambda_i)}{2\lambda_i} (F_i)^2 + \dots \ \text{ as } \ \xi \to -\infty,\ i = 1, 2. \tag{5.14}$$

By (2.20), $\frac{Q(\lambda_1)}{\lambda_1} > \frac{Q(\lambda_2)}{\lambda_2}$. Then, for F small enough we have

$$V_2(F) > V_1(F). \tag{5.15}$$

We can deduce some general properties of F, the solution of (5.1), (5.2). Combining (5.5), (5.11) we easily obtain

$$\liminf_{r \to \infty} F(r) \geq 4 \tag{5.16}$$

for any $\lambda > 0$. Using (5.1) we obtain $F_{rr} \leq \frac{F_r}{r}$, whence $\frac{F_r}{r} \leq C$. Notice that (1.3), (1.4) imply $Q(\xi) \geq \beta\xi$ for $\xi \in (0, C)$. Using these estimates and (5.16) it then follows using (5.1) that

$$F_{rr} - \frac{\beta}{r} F_r \leq 0$$

for some $\beta < 1$ and r large enough. Then $F_r \leq C_1 + C_2 r$, whence $\lim_{r\to\infty}(\frac{F_r}{r}) = 0$. Therefore, it is possible to approximate (5.1) as $r \to \infty$, by means of the equation $F_{rr} - \frac{F_r}{r} + \frac{FF_r}{r} = 0$. Moreover, since $F \geq 4 - \delta_0$, as $r \to \infty$ for some $\delta_0 > 0$, it then follows that F approaches a constant value as $r \to \infty$:

$$\lim_{r\to\infty} F(r) = M(\lambda) < +\infty. \tag{5.17}$$

In order to show the monotonicity of $M(\lambda)$ we will show that (5.15) is satisfied for all the values of F, and not only for small values of F. Suppose that there exists $\tilde{F} > 0$ such that $V_2(\tilde{F}) = V_1(\tilde{F}) \equiv \tilde{V}$. Let us denote as $\xi_i(\tilde{F})$, $i = 1, 2$, the corresponding value of ξ where this identity is reached. Using (5.5) it follows that

$$\int_{F_i(\xi)}^{\tilde{F}} \left(\frac{1}{V_i(f)} - \frac{1}{2f} \right) df + \frac{1}{2} \log \left(\frac{\tilde{F}}{F_i(\xi)} \right) = \xi_i(\tilde{F}) - \xi, \ i = 1, 2.$$

Using (5.7) and taking the limit $\xi \to -\infty$ it then follows that

$$\int_0^{\tilde{F}} \left(\frac{1}{V_i(f)} - \frac{1}{2f} \right) df + \frac{1}{2} \log(\tilde{F}) = \xi_i(\tilde{F}), \ i = 1, 2. \tag{5.18}$$

Using (5.15) we then obtain that $\xi_2(\tilde{F}) < \xi_1(\tilde{F})$. Since $\lambda_2 > \lambda_1$ it follows from (2.20) that

$$\frac{e^{2\xi_2(\tilde{F})}}{\lambda_2} Q(\lambda_2 e^{-2\xi_2(\tilde{F})} \tilde{V}) < \frac{e^{2\xi_1(\tilde{F})}}{\lambda_1} Q(\lambda_1 e^{-2\xi_1(\tilde{F})} \tilde{V}),$$

and using (5.5), (5.6) we then obtain that at the intersection point $(\tilde{F}, \tilde{V})$ it holds that

$$\frac{dF_1}{d\xi}(\xi_1(\tilde{F})) = \frac{dF_2}{d\xi}(\xi_2(\tilde{F})),$$
$$\frac{dV_1}{d\xi}(\xi_1(\tilde{F})) < \frac{dV_2}{d\xi}(\xi_2(\tilde{F})),$$

or, equivalently, $\frac{dV_1}{dF} < \frac{dV_2}{dF}$. However, this is impossible due to (5.15) that was satisfied for $0 < F < \tilde{F}$. Hence, (5.15) holds during the whole evolution of the trajectories in the upper plane. Using this, it then follows that

$$\frac{M(\lambda_1)}{2\pi} = F_\infty \equiv \lim_{r\to\infty} F_1(r) \leq \lim_{r\to\infty} F_2(r) = \frac{M(\lambda_2)}{2\pi}.$$

Actually, this inequality is strict. In order to show that, we argue as follows. Notice that (5.15) and (5.18) imply

$$\xi_2(F) < \xi_1(F). \tag{5.19}$$

Combining (5.5) and (5.6) we obtain

$$\frac{dV_i}{dF} = 2 - \frac{e^{2\xi_i(F)}}{\lambda_i} Q(\lambda_i e^{-2\xi_i(F)} V_i) \frac{F}{V_i},\ i = 1, 2, \tag{5.20}$$

where $V_i = V_i(F)$. Using (5.19) and taking into account (2.20) as well as the fact that $\lambda_2 > \lambda_1$, it follows that

$$\frac{e^{2\xi_2(F)}}{\lambda_2} Q(\lambda_2 e^{-2\xi_2(F)} V_2) < \frac{e^{2\xi_1(F)}}{\lambda_1} Q(\lambda_1 e^{-2\xi_1(F)} V_2).$$

On the other hand, (2.20) and (5.15) imply

$$\frac{Q(aV_2)}{V_2} = \frac{V_1}{V_2} \frac{Q\left(a \frac{V_1}{V_1/V_2}\right)}{V_1} < \frac{Q(aV_1)}{V_1}.$$

Therefore

$$\frac{e^{2\xi_2(F)}}{\lambda_2} \frac{Q(\lambda_2 e^{-2\xi_2(F)} V_2)}{V_2} < \frac{e^{2\xi_1(F)}}{\lambda_1} \frac{Q(\lambda_1 e^{-2\xi_1(F)} V_1)}{V_1}. \tag{5.21}$$

Using (5.20) and (5.21) we obtain

$$\frac{d(V_2 - V_1)}{dF} = \left[\frac{e^{2\xi_1(F)}}{\lambda_1} \frac{Q(\lambda_1 e^{-2\xi_1(F)} V_1)}{V_1} - \frac{e^{2\xi_2(F)}}{\lambda_2} \frac{Q(\lambda_2 e^{-2\xi_2(F)} V_2)}{V_2} \right] > 0, \tag{5.22}$$

an inequality that remains valid as long as $V_i(F),\ i = 1, 2$. Combining (5.15), (5.22) we then obtain that since $V_1(F_\infty) = 0$, we have $V_2(F) > 0$. Henceforth

$$\lim_{r \to \infty} F_2(r) > F_\infty.$$

In order to conclude the proof of Proposition 2.2 it remains to show that $\lim_{\lambda \to 0^+} M(\lambda) = 8\pi$ and $\lim_{\lambda \to \infty} M(\lambda) = \infty$ (or, in an equivalent manner, $F(\infty, \lambda) \to 4$ as $\lambda \to 0^+$, $F(\infty, \lambda) \to \infty$ as $\lambda \to \infty$). In section 2 we have already obtained these asymptotics in a formal manner. We provide here a rigorous proof of them. Taking into account the monotonicity of $V(F)$ on λ it follows that $V(F) \to \bar{V}(F)$ as $\lambda \to 0^+$, where $(\bar{F}, \bar{V})$ is the solution of (5.9), (5.10) satisfying (5.7), whence the computation concerning the first limit follows. On the other hand, a comparison argument shows that $F \leq \frac{e^{2\xi}}{2}$, $V \leq e^{2\xi}$. For ξ of order one (perhaps large) and V positive of order one, we have (cf. (1.3)) $\frac{e^{2\xi}}{\lambda} Q(\lambda e^{-2\xi} V) \to 0$ as $\lambda \to \infty$. Taking into account (5.5), (5.6) it then follows that F becomes arbitrarily large if λ is large. Henceforth $\lim_{\lambda \to \infty} F(\infty, \lambda) = \infty$. □

REFERENCES

[1] F. Alcantara and M. Monk, *Signal propagation during aggregation in the slime mold Dictyostelium discoideum*, J. Gen. Microbiol., 85 (1974), pp. 321–334.

[2] G. Bellettini and G. Fusco, *Stable dynamics of spikes in solutions to a system of reaction-diffusion equations*, Asymptot. Anal., 26 (2001), pp. 307–357.

[3] P. Biler, *Local and global solvability of some parabolic systems modelling chemotaxis*, Adv. Math. Sci. Appl., 8 (1998), pp. 715–743.

[4] J. T. Bonner, *The Cellular Slime Mold*, Princeton University Press, Princeton, NJ, 1967.

[5] M. P. Brenner, P. Constantin, L. P. Kadanoff, A. Schenkel, and S. C. Venkataramani, *Diffusion, attraction and collapse*, Nonlinearity, 12 (1999), pp. 1071–1098.
[6] M. P. Brenner, L. S. Levitov, and E. O. Budrene, *Physical mechanisms for chemotactic pattern formation by bacteria*, Biophys. J., 74 (1998), pp. 1677–1693.
[7] T. Bretschneider, B. Vasiev, and C. J. Weijer, *A model for Dictyostelium slug movement*, J. Theoret. Biol., 199 (1999), pp. 125–136.
[8] E. O. Budrene and H. C. Berg, *Dynamics of formation of symmetrical patterns by chemotactic bacteria*, Nature, 376 (1995), pp. 49–53.
[9] S. Childress, *Chemotactic collapse in two dimensions*, in Modelling of Patterns in Space and Time, Lecture Notes in Biomath. 55, Springer-Verlag, Berlin, 1984, pp. 61–66.
[10] S. Childress and J. K. Percus, *Nonlinear aspects of chemotaxis*, Math. Biosci., 56 (1981), pp. 217–237.
[11] P. Devreotes, *Dictyostelium discoideum: A model system for cell-cell interactions in development*, Science, 245 (1989), pp. 1054–1058.
[12] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.
[13] H. Gajewski and K. Zacharias, *Global behaviour of a reaction-diffusion system modeling chemotaxis*, Math. Nachr., 195 (1998), pp. 77–114.
[14] D. Gilbart and N. S. Trudinger, Elliptic Equations of Second Order, Springer-Verlag, Berlin, 1977.
[15] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin-New York, 1981.
[16] M. A. Herrero, *Asymptotic properties of reaction-diffusion systems modeling chemotaxis*, Applied and Industrial Mathematics, Venice-2, 1998, R. Spigler, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 89–108.
[17] M. A. Herrero and J. J. L. Velázquez, *Singularity patterns in a chemotaxis model*, Math. Ann., 306 (1996), pp. 583–623.
[18] M. A. Herrero and J. J. L. Velázquez, *Chemotactic collapse for the Keller-Segel model*, J. Math. Biol., 35 (1996), pp. 177–196.
[19] M. A. Herrero and J. J. L. Velázquez, *A blow-up mechanism for a chemotaxis model*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 1739–1754.
[20] T. Hillen and H. G. Othmer, *The diffusion limit of transport equations derived from velocity-jump processes*, SIAM J. Appl. Math., 61 (2000), pp. 751–775.
[21] T. Hillen and K. Painter, *Global existence for a parabolic chemotaxis model with prevention of overcrowding*, Adv. in Appl. Math., 26 (2001), pp. 280–301.
[22] T. Höfer, J. A. Sherrat, and P. K. Maini, *Cellular pattern formation during Dictyostelium aggregation*, Physica D, 85 (1995), pp. 425–444.
[23] T. Höfer, J. A. Sherrat, and P. K. Maini, *Dictyostelium discoideum: Cellular self-organization in an excitable biological medium*, Proc. Roy. Soc. London B, 259 (1995), pp. 249–257.
[24] D. Horstmann, *The nonsymmetric case of the Keller-Segel model in chemotaxis: Some recent results*, NoDEA Nonlinear Differential Equations Appl., 8 (2001), pp. 399–423.
[25] D. Horstmann and G. Wang, *Blow-up in a chemotaxis model without symmetry assumptions*, European J. Appl. Math., 12 (2001), pp. 159–177.
[26] D. Horstmann, *On the existence of radially symmetric blow-up solutions for the Keller-Segel model*, J. Math. Biol., 44 (2002), pp. 463–478.
[27] W. Jäger and S. Luckhaus, *On explosions of solutions to a system of partial differential equations modelling chemotaxis*, Trans. Amer. Math. Soc., 329 (1992), pp. 819–824.
[28] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.
[29] H. Levine and W. Reynolds, *Streaming instability of aggregating slime mold amoebae*, Phys. Rev. Lett. 66, 18 (1991), pp. 2400–2403.
[30] A. Liñán and F. A. Williams, *Fundamental Aspects of Combustion*, Oxford University Press, Oxford, UK, 1993.
[31] J. L. Martiel and A. Golbeter, *A model based on receptor desensitization for cyclic AMP signaling in Dictyostelium cells*, Biophys. J., 52 (1987), pp. 807–828.
[32] T. Nagai, *Blow-up of radially symmetric solutions to a chemotaxis system*, Adv. Math. Sci. Appl., (1995), pp. 1–21.
[33] T. Nagai, T. Senba, and T. Suzuki, *Chemotactic collapse in a parabolic system of mathematical biology*, Hiroshima Math. J., 30 (2000), pp. 463–497.
[34] V. Nanjudiah, *Chemotaxis, signal relaying and aggregation morphology*, J. Theoret. Biol., 26 (1988), pp. 263–298.

[35] W. M. Ni, *Diffusion, cross-diffusion and their spike-layer steady states*, Notices of the AMS, 45 (1998), pp. 9–18.
[36] W. M. Ni and I. Takagi, *Locating the peaks of least-energy solutions to a semilinear Neumann problem*, Duke Math. J., 70 (1993), pp. 247–281.
[37] H. G. Othmer and T. Hillen, *The diffusion limit of transport equations* II*: Chemotaxis equations*, SIAM J. Appl. Math., 62 (2002), pp. 1222–1250.
[38] H. G. Othmer and P. Schaap, *Oscillatory cAMP signaling in the development of Dictyostelium discoideum*, Comments Theor. Biol., 5 (1998), pp. 175–282.
[39] H. G. Othmer and A. Stevens, *Aggregation, blowup, and collapse: The ABC's of taxis in reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 1044–1081.
[40] M. del Pino and P. L. Felmer, *Spike-layered solutions of singularly perturbed elliptic problems in a degenerate setting*, Indiana Univ. Math. J., 48 (1999), pp. 883–898.
[41] M. Primicerio and B. Zaltmann, *Global in Time Solution to the Keller–Segel Model of Chemotaxis*, preprint.
[42] J. Rietdorf, F. Sieger, and C. J. Weijer, *Induction of optical density waves and chemotactic cell movement in Dictyostelium discoideum by microinjection of cAMP pulses*, Dev. Biol., 204 (1998), pp. 525–536.
[43] P. G. Saffmann, *Vortex Dynamics*, Cambridge University Press, Cambridge, UK, 1992.
[44] F. Siegert and C. J. Weijer, *Three-dimensional scroll waves organize Dictyostelium slugs*, Proc. Natl. Acad. Sci. USA, 89 (1992), pp. 6433–6437.
[45] A. Stevens, *The derivation of chemotaxis equations as limit dynamics of moderately interacting stochastic many-particle systems*, SIAM J. Appl. Math., 61 (2000), pp. 183–212.
[46] Y. H. Tang and H. G. Othmer, *A G-protein based model of adaptation in Dictyostelium discoideum*, Math. Biosci., 120 (1994), pp. 25–76.
[47] J. J. L. Velázquez, *Stability of some mechanisms of chemotactic aggregation*, SIAM J. Appl. Math., 62 (2002), pp. 1581–1633.
[48] J. J. L. Velázquez, *Well Posedness of a Model of Point Dynamics for a Limit of the Keller–Segel System*, preprint, Universidad Complutense de Madrid, Madrid, 2003.
[49] J. J. L. Velázquez, *Point dynamics in a singular limit of the Keller–Segel model* 2*: Formation of the concentration regions*, SIAM J. Appl. Math., 64 (2004), pp. 1224–1248.
[50] M. Ward, *An asymptotic analysis of localized solutions for some reaction-diffusion models in multidimensional domains*, Stud. Appl. Math., 97 (1996), pp. 103–126.
[51] J. Wei, *On the interior spike layer solutions of the Gierer-Meinhardt system: Uniqueness and spectrum estimates*, European J. Appl. Math., 10 (1999), pp. 353–378.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1224–1248

POINT DYNAMICS IN A SINGULAR LIMIT OF THE KELLER–SEGEL MODEL 2: FORMATION OF THE CONCENTRATION REGIONS*

J. J. L. VELÁZQUEZ†

Abstract. This paper continues the analysis started in the first part of this article (cf. [J. J. L. Velázquez, *SIAM J. Appl. Math.*, 64 (2004), pp. 1198–1223]). It was seen there, using the method of matched asymptotics, that a regularized version of the Keller–Segel system admits, for a suitable asymptotic limit, solutions with some regions of high concentrations for the cell density. This paper considers the relation between the phenomenon of blow-up for the limit problem and the dynamics of the concentration regions described in [J. J. L. Velázquez, *SIAM J. Appl. Math.*, 64 (2004), pp. 1198–1223]. In particular, this paper analyzes the precise way in which the regularization introduced in the Keller–Segel system stops the aggregation process and yields the formation of concentration regions.

Key words. chemotaxis, singular perturbations, matched asymptotics

AMS subject classifications. 35K45, 35B25, 92B05

DOI. 10.1137/S003613990343389X

1. Introduction.

The purpose of this paper is to continue the analysis begun in [22]. That paper studied a regularized version of a Keller–Segel model whose solutions blow-up in finite time.

A characteristic feature of some of the models that have been introduced in the literature to describe chemotactic aggregation is that in some cases their solutions blow-up in a finite time. This is a well-known fact, for instance, for the Keller–Segel model (cf. [16]). In particular, the study of blow-up for this model has received a lot of attention in recent years (cf. [1], [2], [3], [5], [6], [7], [8], [9], [10], [12], [13], [14], [15], [17], [18], [19], [21]).

The presence or absence of blow-up for each particular form of the Keller–Segel model, as well as the asymptotics of the solutions near the blow-up time, depends very sensitively on the choice of the aggregation function that measures cell sensitivity to the gradient of chemicals. In particular, for chemotactic functions that are linear in the cell concentration, the solutions of the resulting Keller–Segel model yield in some general circumstances Dirac mass aggregation in finite time (cf. [8], [9]).

In [22] we addressed the problem of describing the behavior of the solutions of a Keller–Segel system whose chemotactic function behaves linearly for values of the cell concentration that are not too high but are saturating to a constant value for high cell concentrations.

In a more precise manner, the problem considered in this paper is the following:

$$\frac{\partial u}{\partial t} = \Delta u - \nabla(G_\varepsilon(u)\nabla v), \quad x \in \mathbb{R}^2,\ t > 0, \tag{1.1}$$

$$\Delta v + u = 0, \quad x \in \mathbb{R}^2,\ t > 0, \tag{1.2}$$

where u denotes the concentration of the organism and v is the concentration of the

*Received by the editors June 29, 2003; accepted for publication (in revised form) September 11, 2003; published electronically May 5, 2004.
http://www.siam.org/journals/siap/64-4/43389.html
†Departamento de Matemática Aplicada, Facultad de Matemáticas, Universidad Complutense, Madrid 28040, Spain (JJ_velazquez@mat.ucm.es).

chemical secreted by it. We will make the following choice of chemotactic function:

$$G_\varepsilon(u) = \frac{1}{\varepsilon} Q(\varepsilon u), \tag{1.3}$$

where $\varepsilon > 0$ is a small parameter, and the function $Q(\xi)$ is an increasing function satisfying

$$Q(s) = s - \alpha s^2 + \cdots \quad \text{as } s \to 0, \tag{1.4}$$

$$Q(s) \sim L \quad \text{as } s \to \infty, \tag{1.5}$$

where $L > 0$, $\alpha > 0$ are given numbers. A typical example would be $Q(s) = \frac{s}{1+s}$. In other words, instead of assuming that the chemotactic function $G_\varepsilon(u)$ increases without limit as the concentration of the organism becomes high, it will be assumed that the mobility of the organism saturates to a constant value. Other choices of $G_\varepsilon(u)$ would be certainly possible, but we will use this particular one just for the sake of simplicity.

System (1.1)–(1.5) is a particular case of the Keller–Segel model. Let us remark that for $\varepsilon = 0$ the system (1.1)–(1.5) formally becomes

$$\frac{\partial u}{\partial t} = \Delta u - \nabla(u \nabla v), \quad x \in \mathbb{R}^2,\ t > 0, \tag{1.6}$$

$$\Delta v + u = 0, \quad x \in \mathbb{R}^2,\ t > 0. \tag{1.7}$$

It is known that solutions of the problem (1.6), (1.7) might blow-up in a finite time $T > 0$ (cf. [4], [8], [15], [17]). As long as u, v remain bounded, the limit from (1.1)–(1.5) to (1.6), (1.7) does not pose any serious mathematical problem. However, the situation becomes mathematically more interesting for times $t > T$, because the solutions of (1.1)–(1.5) are globally defined in time, something that does not occur for the solutions of (1.6), (1.7). It is then natural to ask what happens to the solutions of (1.1)–(1.5) as $\varepsilon \to 0$ and $t > T$.

In [22] it was seen that it is possible to obtain using matched asymptotics a set of solutions of the regularized problem (1.1)–(1.5) that in the limit $\varepsilon \to 0$ can be decomposed in a regular part plus a set of "concentration regions" where the values of u are very high, containing an amount of mass of order one in a very localized region. Such solutions behave asymptotically as

$$u \approx \sum_{i=1}^{N} M_i(t) \delta(x - x_i(t)) + u_{reg}(x,t), \tag{1.8}$$

where $u_{reg}(x,t)$ is a bounded function.

The detailed dynamics of the points $x_i(t)$, as well as their masses $M_i(t)$ and the values of $u_{reg}(x,t)$, has been derived in [22].

The goal of this paper is to verify that the fact that the solutions of (1.6), (1.7) might blow-up in a finite time provides a mechanism of transition between bounded solutions of (1.1)–(1.5), i.e., solutions containing only the regular part $u_{reg}(x,t)$ and solutions containing concentration regions as in (1.8). More precisely, we will describe how the blow-up mechanism described in [8], [9] evolves in a short transition time to a region of high density as those described in [22].

The detailed manner in which this transition from a "blowing up" regime evolves to a "quasi-steady regime" characterized by high cell densities depends on some details

of the function $Q(\cdot)$ given above. Nevertheless, only a few characteristics of the function $Q(\cdot)$ play a crucial role, namely, the precise quadratic corrective behavior as $s \to 0$ given in (1.4). The analysis in this paper should be modified for nonlinearities $Q(\cdot)$ having corrective behaviors different from the one given in (1.4). In any case, the goal of this paper is not to obtain the most general description of the transition between the blowing up regime and the quasi-steady regime, but only to point out that this feature takes place and to describe methods for performing the analysis of such transitions.

In section 2 of this paper the main characteristics of the blow-up mechanism for (1.6), (1.7) obtained in [8], [9] will be recalled. Section 3 summarizes some of the basic properties of the stationary states of (1.1)–(1.5) that have been derived in [22]. Section 4 describes in a detailed manner the transition between the blowing up regime and the quasi-steady regime.

2. Formation of concentration regions: The limit $\varepsilon = 0$. In this section we recall the mechanism of Dirac mass formation described in [8], [9] for the system (1.6), (1.7).

The main idea of this section is the following. For u bounded we can approximate (1.1), (1.2) in the limit $\varepsilon \to 0$ by the simpler model (1.6), (1.7).

It is a well-known fact that solutions of (1.6), (1.7) might develop singularities in finite time (cf. [4], [8], [15], [17]). Moreover, such solutions develop Dirac masses containing exactly the amount of mass 8π (cf. [8], [9], [18]) in a finite time, which will be denoted henceforth as $t = T > 0$.

Let us state in Theorem 2.1 the main result that will be recalled in this section.

THEOREM 2.1. *There exists a solution of* (1.6), (1.7) *that yields concentration of an amount of mass* 8π *at the origin. The asymptotics of this solution near the origin is given by*

$$u(x,t) \sim \frac{8}{(T-t)(\varepsilon(|\log(T-t)|))^2} \frac{1}{\left(1 + \frac{|x|^2}{(T-t)(\varepsilon(|\log(T-t)|))^2}\right)^2}$$

for $|x| = O(\varepsilon(|\log(T-t)|)\sqrt{(T-t)})$ *and* $t \to T^-$, *where the asymptotics* $\varepsilon(\tau)$ *is given by*

$$\varepsilon(\tau) \sim 2e^{-\frac{2+\gamma}{2}} e^{-\sqrt{\tau/2}} \left(1 + O\left(\frac{\log(\tau)}{\sqrt{\tau}}\right)\right) \quad \text{as } \tau \to \infty \tag{2.1}$$

and where $\gamma = 0.5772156\ldots$ *is the classical Euler constant.*

Theorem 2.1 has been proved rigorously in [8]. In this section we will just provide a formal description of the considered solution. Two other different ways of computing asymptotic descriptions of the sought-for solutions can be found in [9], [21]. Nevertheless, the variables used in this section to compute the asymptotics of the desired solution are more convenient than the previous ones, keeping in mind the analysis in section 4.

Formal proof of Theorem 2.1. It is convenient to introduce a "mass variable" as is usually done in the study of radial problems. Let $M(r,t)$ be

$$M(r,t) = \int_0^r \xi u(\xi,t) d\xi. \tag{2.2}$$

The name "mass variable" is due to the fact that $M(r,t)$ is, except for a factor 2π, the amount of mass of u in a ball of radius $B_r(0)$.

Differentiating (2.2) and using (1.2) we obtain

$$u(r,t) = \frac{1}{r}\frac{\partial M}{\partial r}, \tag{2.3}$$

$$\frac{\partial v}{\partial r} = -\frac{M}{r}. \tag{2.4}$$

Integrating (1.1) in a ball $B_r(0)$ and using (2.3), (2.4) we obtain

$$\frac{\partial M}{\partial t} = r\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial M}{\partial r}\right) + G_\varepsilon\left(\frac{M_r}{r}\right)M. \tag{2.5}$$

In the limit $\varepsilon = 0$, (2.5) becomes

$$\frac{\partial M}{\partial t} = r\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial M}{\partial r}\right) + \frac{MM_r}{r}, \tag{2.6}$$

that is, the parabolic equation that we would obtain applying the transformation (2.2) to the system (1.6), (1.7).

A crucial role in the analysis of the considered singularity formation mechanism is played by the steady state

$$\bar{M}(r) = \frac{4r^2}{r^2+1}. \tag{2.7}$$

In order to study the sought-for singularity formation mechanism it is convenient to introduce self-similar variables

$$M(r,t) = \bar{\Phi}\left(\frac{r}{\sqrt{T-t}}, -\log(T-t)\right), \quad \bar{y} = \frac{r}{\sqrt{T-t}}, \quad \bar{\tau} = -\log(T-t), \tag{2.8}$$

where $t = T$ is the blow-up time. The reason for writing bars above all the variables is that a set of slightly different variables (without bars) will be used later in section 4 to analyze the problem in the case $\varepsilon > 0$. In this set of variables (2.6) becomes

$$\bar{\Phi}_{\bar{\tau}} = \bar{y}\frac{\partial}{\partial \bar{y}}\left(\frac{1}{\bar{y}}\frac{\partial \bar{\Phi}}{\partial \bar{y}}\right) - \frac{1}{2}\bar{y}\bar{\Phi}_{\bar{y}} + \frac{1}{\bar{y}}\bar{\Phi}\bar{\Phi}_{\bar{y}}. \tag{2.9}$$

The singular solution described in [8] behaves like the steady state (2.7) in the region $|\bar{y}| \to 0$ after a suitable rescaling. In order to compute in a precise manner the size of such a region we introduce a new variable by means of

$$\bar{\xi} = \frac{\bar{y}}{\bar{\delta}(\bar{\tau})}, \tag{2.10}$$

where the undetermined function $\bar{\delta}(\bar{\tau})$ approaches zero as $\bar{\tau} \to \infty$.

Using $\bar{\xi}$ as a new space variable (1.6) reads as

$$\bar{\Phi}_{\bar{\tau}} = \frac{1}{\bar{\delta}^2}\left[\bar{\xi}\frac{\partial}{\partial \bar{\xi}}\left(\frac{1}{\bar{\xi}}\frac{\partial \bar{\Phi}}{\partial \bar{\xi}}\right) + \frac{1}{\bar{\xi}}\bar{\Phi}\frac{\partial \bar{\Phi}}{\partial \bar{\xi}}\right] + \frac{\bar{\delta}_{\bar{\tau}}}{\bar{\delta}}\bar{\xi}\frac{\partial \bar{\Phi}}{\partial \bar{\xi}} - \frac{\bar{\xi}}{2}\frac{\partial \bar{\Phi}}{\partial \bar{\xi}}. \tag{2.11}$$

In order to compute $\bar{\delta}(\bar{\tau})$ we need to obtain two terms in the asymptotics of $\bar{\Phi}(\bar{\xi},\bar{\tau})$ as $\bar{\tau} \to \infty$. By assumption, to the leading order, we have $\bar{\Phi}(\bar{\xi},\bar{\tau}) \to \bar{M}(\bar{\xi})$ as $\bar{\tau} \to \infty$. It is then natural to try an expansion of the form

$$\bar{\Phi}(\bar{\xi},\bar{\tau}) = \bar{M}(\bar{\xi}) + \psi(\bar{\xi},\bar{\tau}) + \cdots, \tag{2.12}$$

where $|\psi| \ll 1$ as $\bar{\tau} \to \infty$. Linearizing and keeping only the leading terms in (1.6) we obtain

$$\bar{\xi}\frac{\partial}{\partial\bar{\xi}}\left(\frac{1}{\bar{\xi}}\frac{\partial\psi}{\partial\bar{\xi}}\right)+\frac{1}{\bar{\xi}}\bar{M}(\bar{\xi})\frac{\partial\psi}{\partial\bar{\xi}}+\frac{1}{\bar{\xi}}\psi\frac{\partial\bar{M}}{\partial\bar{\xi}}=\left(\frac{\bar{\delta}^2}{2}-\bar{\delta}\bar{\delta}_{\bar{\tau}}\right)\bar{\xi}\frac{\partial\bar{M}}{\partial\bar{\xi}}. \tag{2.13}$$

Since $\varphi(\bar{\xi})=\frac{\bar{\xi}^2}{(\bar{\xi}^2+1)^2}$ solves the homogeneous problem associated to (1.6), we can solve this problem by looking for solutions in the form

$$\psi(\bar{\xi},\bar{\tau})=\varphi(\bar{\xi})F(\bar{\xi},\bar{\tau}). \tag{2.14}$$

Therefore

$$\bar{\xi}\frac{\partial}{\partial\bar{\xi}}\left(\frac{1}{\bar{\xi}}\frac{\partial F}{\partial\bar{\xi}}\right)+\frac{4}{(\bar{\xi}^2+1)}\left(\frac{1}{\bar{\xi}}\frac{\partial F}{\partial\bar{\xi}}\right)=8\left(\frac{\bar{\delta}^2}{2}-\bar{\delta}\bar{\delta}_{\bar{\tau}}\right). \tag{2.15}$$

We can assume without loss of generality that $F(0,\bar{\tau})=0$. Otherwise ψ would contain a contribution of the form $F(0,\bar{\tau})\varphi(\bar{\xi})$, where by hypothesis, $F(0,\bar{\tau})\to 0$ as $\bar{\tau}\to\infty$. Since the homogeneous solution $\varphi(\bar{\xi})$ is associated to the infinitesimal changes of $\bar{M}(\bar{\xi})$ due to the application of the rescaling group of (2.6), it would be possible to eliminate $F(0,\bar{\tau})$ by means of a change of $\bar{\delta}(\bar{\tau})$. In an equivalent way, the choice $F(0,\bar{\tau})=0$ is just a way of prescribing $\bar{\delta}(\bar{\tau})$ in a precise manner. In particular, using (2.2) and (1.6), it follows that the choice $F(0,\bar{\tau})=0$ is equivalent to setting $\frac{8}{(\bar{\delta}(\bar{\tau}))^2}=u(0,t)$. The factor 8 has been introduced in order to obtain some simpler formulae later.

Solving (2.15) with the additional condition $F(0,\bar{\tau})=0$ we obtain

$$F(\bar{\xi},\bar{\tau})=4\left(\frac{\bar{\delta}^2}{2}-\bar{\delta}\bar{\delta}_{\bar{\tau}}\right)\int_0^{\bar{\xi}}\frac{(\eta^2+1)^2}{\eta^3}\left[\log(\eta^2+1)-\frac{\eta^2}{\eta^2+1}\right]d\eta. \tag{2.16}$$

Standard computations yield the following asymptotics for $F(\bar{\xi},\bar{\tau})$ as $\bar{\xi}\to\infty$:

$$F(\bar{\xi},\bar{\tau})\sim 4\left(\frac{\bar{\delta}^2}{2}-\bar{\delta}\bar{\delta}_{\bar{\tau}}\right)[\bar{\xi}^2\log(\bar{\xi})-\bar{\xi}^2+O((\log(\bar{\xi}))^2)]\quad\text{as }\bar{\xi}\to\infty. \tag{2.17}$$

Combining (2.8), (2.12) and the asymptotics (2.17) we obtain the following matching condition:

$$\bar{\Phi}(\bar{\xi},\bar{\tau})\sim\frac{4\bar{\xi}^2}{(\bar{\xi}^2+1)}+4\left(\frac{\bar{\delta}^2}{2}-\bar{\delta}\bar{\delta}_{\bar{\tau}}\right)\left[\log(\bar{\xi})-1+O\left(\frac{(\log(\bar{\xi}))^2}{\bar{\xi}^2}\right)\right] \tag{2.18}$$

for $\bar{\xi}\gg 1$, $\bar{y}\ll 1$, $\bar{\tau}\to\infty$.

We now describe the asymptotics of $\bar{\Phi}$ on the "outer region" $\bar{y}\sim 1$. To this end we linearize (1.6) around its limit value $\bar{\Phi}_\infty=4$. More precisely, let us write

$$\bar{\Phi}=4+\psi. \tag{2.19}$$

Then ψ solves

$$\psi_{\bar{\tau}}=\psi_{\bar{y}\bar{y}}+\frac{3}{\bar{y}}\psi_{\bar{y}}-\frac{\bar{y}\psi_{\bar{y}}}{2}+\frac{(\bar{\Phi}-4)}{\bar{y}}\bar{\Phi}_{\bar{y}}. \tag{2.20}$$

At first glance it could seem possible to neglect the term $\frac{(\bar\Phi-4)}{\bar y}\bar\Phi_{\bar y} = \frac{\psi\psi_{\bar y}}{\bar y}$ that is quadratic on ψ. It turns out, however, that this is not possible, because this term yields a relevant contribution in the region $\bar y \approx \bar\delta(\bar\tau)$. Using the approximation $\bar\Phi \sim \bar M(\bar\xi)$ in the region $\bar y \approx \bar\delta(\bar\tau)$ we obtain

$$\frac{(\bar\Phi-4)}{\bar y}\bar\Phi_{\bar y} \approx -\frac{1}{\bar\delta^2}\cdot\frac{32}{(\bar\xi^2+1)^3} \quad \text{as } \bar\tau\to\infty.$$

Notice that the operator $\partial^2_{\bar y}+\frac{3}{\bar y}\partial_{\bar y}$ that appears on the right-hand side of (1.7) is the Laplacian acting on radial functions in four spatial dimensions. Since $\frac{(\bar\Phi-4)}{\bar y}\bar\Phi_{\bar y}$ is concentrated in the region $\bar y \approx \bar\delta(\bar\tau)$ it would be natural to approximate this term by some kind of Dirac mass function, but due to the dimensionality of the Laplacian operator in (1.7) the mass of the singular part has to be computed in four dimensions. Using the fact that $\Gamma \equiv \int_{\mathbb{R}^4}\frac{d^4\bar\xi}{(\bar\xi^2+1)^3} = \frac{\pi^2}{2}$, it is natural to approximate (2.20) as

$$\psi_{\bar\tau} = \psi_{\bar y\bar y} + \frac{3}{\bar y}\psi_{\bar y} - \frac{\bar y\psi_{\bar y}}{2} - 32\Gamma(\bar\delta(\bar\tau))^2\delta^{(4)}(\bar y), \tag{2.21}$$

where $\delta^{(4)}(\bar y)$ is a four-dimensional Dirac mass. In order to study (1.7) it is convenient to decompose ψ as

$$\psi(\bar y,\bar\tau) = \bar a_0(\bar\tau) + Q(\bar y,\bar\tau), \tag{2.22}$$

where $\langle Q,1\rangle = 0$, and where from now on,

$$\langle f,g\rangle = \int_{\mathbb{R}^4} f(\bar y)g(\bar y)e^{-\frac{|\bar y|^2}{4}}d^4\bar y. \tag{2.23}$$

It then follows that

$$\dot a_0(\bar\tau) = -\frac{32\Gamma}{\langle 1,1\rangle}(\bar\delta(\bar\tau))^2 = -(\bar\delta(\bar\tau))^2, \tag{2.24}$$

where we have used the fact that $\langle 1,1\rangle = 32\Gamma$.

Notice that $\psi\to 0$ as $\bar\tau\to\infty$. Therefore

$$\bar a_0(\bar\tau) = \int_{\bar\tau}^{\infty}(\bar\delta(s))^2 ds. \tag{2.25}$$

On the other hand, $Q(\bar y,\bar\tau)$ solves

$$Q_{\bar\tau} = Q_{\bar y\bar y} + \frac{3}{\bar y}Q_{\bar y} - \frac{\bar yQ_{\bar y}}{2} - 32\Gamma(\bar\delta(\bar\tau))^2\left[\delta^{(4)}(\bar y) - \frac{1}{\langle 1,1\rangle}\right]. \tag{2.26}$$

The function $\bar\delta(\bar\tau)$ does not contain exponential factors in its asymptotics (cf. [8]), or in a more precise way, we can assume that $|\frac{\bar\delta_{\bar\tau}}{\bar\delta}| \ll 1$. It then follows that we can approximate $Q(\bar y,\bar\tau)$ as the unique solution of

$$0 = Q_{\bar y\bar y} + \frac{3}{\bar y}Q_{\bar y} - \frac{\bar yQ_{\bar y}}{2} - 32\Gamma(\bar\delta(\bar\tau))^2\left[\delta^{(4)}(\bar y) - \frac{1}{\langle 1,1\rangle}\right] \tag{2.27}$$

satisfying $\langle Q, 1\rangle = 0$. The general solution of (2.27) is

$$Q(\bar{y},\bar{\tau}) = (\bar{\delta}(\bar{\tau}))^2 \left[-\frac{4}{\bar{y}^2} + 2\log(\bar{y}) + B\right],$$

where $B \in \mathbb{R}$ has to be obtained using the orthogonality condition $\langle Q, 1\rangle = 0$. It then follows, after some computations, that $B = \gamma - \log(4)$, where $\gamma = 0.577215\ldots$, is the standard Euler constant. Therefore

$$Q(\bar{y},\bar{\tau}) = \Omega(\bar{y})(\bar{\delta}(\bar{\tau}))^2, \tag{2.28}$$

where $\Omega(\bar{y}) \equiv [-\frac{4}{\bar{y}^2} + 2\log(\bar{y}) + \gamma - \log(4)]$.

In order to obtain a matching condition between the inner and the outer expansion we need to compute an additional term in the asymptotics of $Q(\bar{y},\bar{\tau})$. We write

$$Q(\bar{y},\bar{\tau}) = \Omega(\bar{y})(\bar{\delta}(\bar{\tau}))^2 + R(\bar{y},\bar{\tau}), \tag{2.29}$$

where R satisfies

$$2\bar{\delta}\bar{\delta}_{\bar{\tau}}\Omega(\bar{y}) + R_{\bar{\tau}} = R_{\bar{y}\bar{y}} + \frac{3}{\bar{y}}R_{\bar{y}} - \frac{\bar{y}R_{\bar{y}}}{2}, \quad \langle R, 1\rangle = 0. \tag{2.30}$$

The "algebraic-like" behavior of R (i.e., the absence of exponential terms), suggests that approximating $R(\bar{y},\bar{\tau})$ to the leading order as $R(\bar{y},\bar{\tau}) = \bar{\delta}\bar{\delta}_{\bar{\tau}}W(\bar{y})$, where $W(\bar{y})$, solves

$$2\Omega(\bar{y}) = W_{\bar{y}\bar{y}} + \frac{3}{\bar{y}}W_{\bar{y}} - \frac{\bar{y}W_{\bar{y}}}{2}, \quad \langle W, 1\rangle = 0. \tag{2.31}$$

It is possible to obtain an explicit formula for $W(\bar{y})$, but since it will not be needed for our purposes we will not pursue that here. The only information that we will need about $W(\bar{y})$ is the asymptotics

$$W(\bar{y}) \sim -4\log(\bar{y}) + \zeta + o(1) \quad \text{as } \bar{y} \to 0, \tag{2.32}$$

where ζ is a real number whose precise numerical value will not be of relevance here.

Let us summarize. We have obtained the following asymptotics for $\bar{\Phi}$ in the region $\bar{y} \sim 1$ (cf. (2.19), (2.22), (2.25), (2.28)):

$$\bar{\Phi}(\bar{y},\bar{\tau}) = \int_{\bar{\tau}}^{\infty} (\bar{\delta}(s))^2 ds + \Omega(\bar{y})(\bar{\delta}(\bar{\tau}))^2 + \bar{\delta}(\bar{\tau})\bar{\delta}_{\bar{\tau}}(\bar{\tau})W(\bar{y}) + \cdots \tag{2.33}$$

as $\bar{\tau} \to \infty$.

Using the asymptotics of $\Omega(\bar{y})$, $W(\bar{y})$ as $\bar{y} \to 0^+$ we obtain the following outer matching condition:

$$\begin{aligned}\bar{\Phi}(\bar{y},\bar{\tau}) \sim 4 - \frac{4(\bar{\delta}(\bar{\tau}))^2}{\bar{y}^2} + \int_{\bar{\tau}}^{\infty} (\bar{\delta}(s))^2 ds + 2(\bar{\delta}(\bar{\tau}))^2\log(\bar{y}) \\ +(\gamma - \log(4))(\bar{\delta}(\bar{\tau}))^2 - 4\bar{\delta}(\bar{\tau})\bar{\delta}_{\bar{\tau}}(\bar{\tau})\log(\bar{y}) + \zeta\bar{\delta}(\bar{\tau})\bar{\delta}_{\bar{\tau}}(\bar{\tau}) + \cdots\end{aligned} \tag{2.34}$$

for $\bar{\delta}(\bar{\tau}) \ll |\bar{y}| \ll 1$ and $\bar{\tau} \to \infty$.

On the other hand, using that $\bar{\xi} = \frac{\bar{y}}{\bar{\delta}(\bar{\tau})}$ as well as (1.6) we obtain the following inner matching condition:

$$\begin{aligned}\bar{\Phi}(\bar{y},\bar{\tau}) \sim 4 - \frac{4(\bar{\delta}(\bar{\tau}))^2}{\bar{y}^2} + 4\left(\frac{\bar{\delta}^2}{2} - \bar{\delta}\bar{\delta}_{\bar{\tau}}\right)\log(\bar{y}) - 4\left(\frac{\bar{\delta}^2}{2} - \bar{\delta}\bar{\delta}_{\bar{\tau}}\right)\log(\bar{\delta}) \\ - 4\left(\frac{\bar{\delta}^2}{2} - \bar{\delta}\bar{\delta}_{\bar{\tau}}\right) + \cdots\end{aligned} \tag{2.35}$$

for $\bar{\delta}(\bar{\tau}) \ll |\bar{y}| \ll 1$ and $\bar{\tau} \to \infty$.

Matching (2.34), (2.35) we obtain the following integral equation:

$$\begin{aligned}\int_{\bar{\tau}}^{\infty} (\bar{\delta}(s))^2 ds + (\gamma - \log(4))(\bar{\delta}(\bar{\tau}))^2 + \zeta\bar{\delta}(\bar{\tau})\bar{\delta}_{\bar{\tau}}(\bar{\tau}) \\ = -2(\bar{\delta}^2 - 2\bar{\delta}\bar{\delta}_{\bar{\tau}})\log(\bar{\delta}) - 2(\bar{\delta}^2 - 2\bar{\delta}\bar{\delta}_{\bar{\tau}}).\end{aligned} \tag{2.36}$$

The solution of (2.36) that approaches zero as $\bar{\tau} \to \infty$ behaves asymptotically as

$$\bar{\delta}(\bar{\tau}) \sim 2e^{-\frac{(\gamma+2)}{2}} e^{-\sqrt{\frac{\bar{\tau}}{2}}}(1 + o(1)) \quad \text{as } \bar{\tau} \to \infty. \tag{2.37}$$

Using the set of original variables x, t (cf. (2.8)), (2.37) means that the mass is concentrated in a region of size

$$|x| \sim 2e^{-\frac{(\gamma+2)}{2}}\sqrt{T-t}\, e^{-\sqrt{\frac{|\log(T-t)|}{2}}}(1 + o(1)) \quad \text{as t} \to \text{T}^-, \tag{2.38}$$

that is, exactly the size computed in [21]. A similar formula, where some of the numerical constants have not been computed in so detailed a manner and containing some typographical mistakes, can be found in [9]. This concludes the description of the formal argument behind the proof of Theorem 2.1. □

Let us remark that the blow-up profile $M(r,T)$ can also be computed. To this end we argue as follows. We fix $\bar{y}_0 > 0$ large enough and t_0 close to T. Taking into account (2.8) and (2.34) it follows that

$$M(r,t_0) \sim 4 + \int_{\bar{\tau}_0}^{\infty} (\bar{\delta}(s))^2 ds$$

for $r \sim \bar{y}_0\sqrt{T-t_0}$, $\bar{\tau}_0 = -\log(T-t_0)$. Using (2.37) we obtain the approximation

$$M(r,t_0) \sim 4 + 4e^{-(\gamma+2)}\sqrt{2\bar{\tau}_0}\, e^{-\sqrt{2\bar{\tau}_0}}, \quad r \sim \bar{y}_0\sqrt{T-t_0}. \tag{2.39}$$

Notice that for $t_0 < t < T$, (2.6) implies that $M(r,t)$ remains nearly constant on a set of the form $r \in [(\bar{y}_0 - C)\sqrt{T-t_0}, (\bar{y}_0 + C)\sqrt{T-t_0}]$ if $C < \bar{y}_0$ and C, $\bar{y}_0$ are chosen large enough. This is due to the fact that for times $(T-t_0)$, perturbations propagate at most distances of order $\sqrt{T-t_0}$ for the solutions of the parabolic equation (2.6). Using this fact, as well as (2.39), it follows that

$$M(r,T) \sim 4 + 4e^{-(\gamma+2)}\sqrt{2\bar{\tau}_0}\, e^{-\sqrt{2\bar{\tau}_0}}, \quad r \sim \bar{y}_0\sqrt{T-t_0}.$$

Since $\bar{\tau}_0 = -\log(\frac{r^2}{\bar{y}_0}) \sim -\log(r^2)$ as $r \to 0^+$ for each fixed $\bar{y}_0$, we have

$$M(r,T) \sim 4 + 8e^{-(\gamma+2)}\sqrt{|\log(r)|}\, e^{-2\sqrt{|\log(r)|}} \tag{2.40}$$

as $r \to 0^+$.

To conclude this section, let us remark that it is also possible to compute the asymptotics of $M(r,t)$ for $t \to T^+$, something that will be needed in order to derive a description of the precise manner in which the concentration region begins to increase its mass. Notice that the asymptotics (2.40), as well as the definition of $M(r,t)$ (cf. (2.2)), indicates the presence of a Dirac mass placed at $r = 0$, with a total amount of mass 8π. Actually, it is possible to obtain the desired asymptotics as follows. Near the origin $M \sim 4$, whence (2.6) can be approximated by the linear equation $M_t = M_{rr} + \frac{3}{r} M_r$ that is just the usual heat equation in dimension four. More precisely, replacing M by the new variable

$$\varphi = M - 4 \tag{2.41}$$

and neglecting quadratic terms on φ in (2.6), we obtain the equation

$$\varphi_t = \varphi_{rr} + \frac{3}{r}\varphi_r. \tag{2.42}$$

This equation can be solved using as initial data the function φ in (2.40), (2.41) and caloric kernels. Therefore, to the leading order,

$$\varphi(r,t) = \frac{1}{(4\pi(t-T))^2} \int_{\mathbb{R}^4} e^{-\frac{(x-\bar{\xi})^2}{4(t-T)}} \varphi(|\bar{\xi}|, T) d^4\bar{\xi}, \tag{2.43}$$

where $r = |x|$ and $\varphi(r,T) \sim 8e^{-(\gamma+2)}\sqrt{|\log(r)|}e^{-2\sqrt{|\log(r)|}}$ as $r \to 0^+$. In order to compute the asymptotics of $\varphi(r,t)$ as $t \to T^+$ it is convenient to deal in a different manner with the regions $|x| \le C\sqrt{t-T}$ and $|x| \gg \sqrt{t-T}$. Let us define $\bar{y} = \frac{x}{\sqrt{t-T}}$. Notice that (2.43) implies

$$\varphi(r,t) = \frac{1}{(4\pi)^2} \int_{\mathbb{R}^4} e^{-\frac{(\bar{y}-\eta)^2}{4}} \varphi(|\eta|\sqrt{t-T}, T) d^4\eta.$$

Using this formula, we obtain for $|\bar{y}| \le C$ the asymptotics

$$\begin{aligned} \varphi(r,t) \sim{}& 4e^{-(\gamma+2)}\sqrt{2|\log(t-T)|}e^{-\sqrt{2|\log(t-T)|}} \cdot \\ &\cdot \left\{ 1 - \frac{1}{\sqrt{|\log(t-T)|}} \frac{1}{(4\pi)^2} \int_{\mathbb{R}^4} e^{-\frac{(\bar{y}-\eta)^2}{4}} \log(|\eta|) d^4\eta + \cdots \right\} \end{aligned} \tag{2.44}$$

as $t \to T^+$.

Assume on the contrary that $\sqrt{t-T} \ll |x| \ll 1$. We then define $\bar{y}$ by means of

$$|\bar{\xi}| = r_0 + \bar{y}\sqrt{t-T},$$

where $r_0 = |x|$, $\sqrt{t-T} \ll r_0 \ll 1$.

Using the asymptotics of $\varphi(r,T)$ it follows that

$$\varphi(|\bar{\xi}|, T) \sim \varphi(r_0, T)$$

as $r_0 \to 0$. Using this as well as (2.43) it then follows that

$$\varphi(r_0,t) \sim \varphi(r_0,T) \sim 8e^{-(\gamma+2)}\sqrt{|\log(r_0)|}e^{-2\sqrt{|\log(r_0)|}}. \tag{2.45}$$

It is not hard to check that (2.44) and (2.45) match for $|\bar{y}| \gg 1$.

For further reference we notice that we have found the following formula for the amount of mass concentrated at the concentration region (cf. (2.44)):

$$m(t) \sim 8\pi \left(1 + e^{-(\gamma+2)}\sqrt{2|\log(t-T)|}e^{-\sqrt{2|\log(t-T)|}} + \cdots\right) \quad \text{as t} \to \text{T}^+. \tag{2.46}$$

3. Steady states. In this section, we recall the basic properties of the steady states of (1.1), (1.2) that have been obtained in [22]. More precisely we consider the radial solutions of the system

$$\Delta\bar{u} - \nabla(G_\varepsilon(\bar{u})\nabla\bar{v}) = 0, \quad x \in \mathbb{R}^2, \tag{3.1}$$

$$\Delta\bar{v} + \bar{u} = 0, \quad x \in \mathbb{R}^2, \tag{3.2}$$

where $G_\varepsilon(u)$ is as in (1.3)–(1.5), and where from now on bars will be introduced above the functions that denote steady states. The scaling structure of $G_\varepsilon(u)$ in (1.3) suggests introducing the new set of variables

$$\bar{u} = \frac{1}{\varepsilon}\bar{U}, \tag{3.3}$$

$$x = \sqrt{\varepsilon}y \tag{3.4}$$

that transforms (3.1), (3.2) into

$$\Delta_y\bar{U} - \nabla_y(Q(\bar{U})\nabla_y\bar{v}) = 0, \quad y \in \mathbb{R}^2, \tag{3.5}$$

$$\Delta_y\bar{v} + \bar{U} = 0, \quad y \in \mathbb{R}^2. \tag{3.6}$$

Let us write $r = |y|$. The following result has been obtained in [22].

THEOREM 3.1. *Suppose that $Q(\cdot) \in C^1(\mathbb{R}^+)$ is an increasing function satisfying* (1.4). *Let us assume also that $\frac{Q(s)}{s}$ is a decreasing function. Then, for each $M > 8\pi$ there exists a unique radial solution (up to translations) such that*

$$\int_{\mathbb{R}^2} \bar{U}(y;M)d^2y = M.$$

The function $\bar{U}(y;M) = \bar{U}(r;M)$ is decreasing on r and its asymptotic behavior as $r \to \infty$ is given by

$$\bar{U}(r;M) \sim \frac{k(M)}{r^{\frac{M}{2\pi}}} \quad \text{as } r \to \infty$$

for some suitable $k(M) > 0$.

In some of the computations below it will be convenient to use $\bar{U}(0;M)$ instead of M in order to characterize the steady states. In Theorem 3.2 we summarize the results that will be used in forthcoming sections and that have been proved in [22].

THEOREM 3.2. *Suppose that $\bar{U}(0;M) = \lambda$. Then, in the limit $\lambda \to 0$ the following asymptotics holds:*

$$\bar{U}(0;M) \sim \frac{\lambda}{8}\left(\varphi_0(\xi) + \frac{\alpha\lambda}{8}\varphi_1(\xi) + \cdots\right), \quad \xi = \frac{\sqrt{\lambda}r}{\sqrt{8}}, \tag{3.7}$$

uniformly in bounded regions of ξ, where α is as in (1.4) *and*

$$\varphi_0(\xi) = \frac{8}{(1+\xi^2)^2}, \qquad \varphi_1(\xi) = \frac{32}{3}\left[\frac{\xi^4 + 10\xi^2 + 2\ln(1+\xi^2) - 2\xi^4\ln(1+\xi x^2)}{(1+\xi^2)^4}\right]. \tag{3.8}$$

Moreover,

$$M = 8\pi + \frac{4\pi\alpha\lambda}{3} + o(\lambda) \quad \text{as } \lambda \to 0^+. \tag{3.9}$$

4. Formation of concentration regions: The limit $\varepsilon \to 0^+$.

4.1. Sketch of the main argument and preliminary results. Standard continuity results on the initial data for PDEs show that as long as the solutions of (1.6), (1.7) are bounded, the solutions of (1.1)–(1.5) converge to the solutions of (1.6), (1.7) as $\varepsilon \to 0^+$. On the other hand, it has been shown in [22] that the solutions of (1.6), (1.7) are globally bounded in time for any $\varepsilon > 0$ fixed. It also has been seen in [22] that there exist formal solutions of (1.1)–(1.5) containing "Dirac masses" for the cell density in some particular regions. In this section it will be confirmed that as one approaches the blow-up time the solutions of (1.1)–(1.5) make a transition between the blowing up behavior obtained for $\varepsilon = 0$ recalled in section 2 of this paper and the "quasi-steady behavior" for regions of high density that is in the basis of the dynamics described in [22].

We will restrict our analysis to radial solutions. In the general nonradial case computations become much more cumbersome, but on the other hand, irrelevant changes arise (cf. [21]). Restricting the analysis to radially symmetric situations does not suppose an important loss of generality because the transition described here occurs very fast and during these times the displacements of the concentration regions would be very small.

In all the analysis made in this paper the only blow-up mechanism that will be considered is the one described in section 2. It is not known if this blow-up mechanism is the only one that can take place for this equation in the two-dimensional case. It is known that for two-dimensional solutions blow-up for the system (1.6), (1.7) can occur only by means of the aggregation of a finite amount of mass at a point (cf. [18]).

If the blow-up mechanism considered here had been unstable it would not be realistic to describe the formation of concentration regions by means of the methods considered in this paper. The question of the stability of this blow-up mechanism has been addressed using asymptotic analysis in [21].

The goal of this section is to study the precise manner in which the blow-up mechanism described in section 2 is regularized for $\varepsilon > 0$ small due to the boundedness of $G_\varepsilon(u)$ in (1.1). Global well-posedness for (1.1)–(1.5) has been proved in [22].

As a preliminary step, it is possible to obtain a rough estimate of the time scales for which the approximation (1.6), (1.7) stops being valid. The term $G_\varepsilon(\frac{M_r}{r})$ differs most from its limit value in the region $r \approx 0$. We will approximate M in that region using (2.8), (1.6). Therefore $M(r,t) \sim \bar{M}(\frac{r}{\sqrt{T-t}\bar{\delta}(\bar{\tau})})$, where $\bar{\delta}(\bar{\tau})$ is as in (2.37). Whence $\frac{M_r}{r} \sim \frac{1}{(T-t)(\bar{\delta}(\bar{\tau}))^2}\frac{\bar{M}_{\bar{\xi}}}{\bar{\xi}}$. Recalling that $G_\varepsilon(s) = \frac{Q(\varepsilon s)}{\varepsilon}$, it would follow that the approximation $G_\varepsilon(s) \sim s$ ceases being valid if $\frac{\varepsilon M_r}{r}$ becomes of order one, i.e., for $\frac{\varepsilon}{(T-t)(\bar{\delta}(\bar{\tau}))^2} \sim 1$, or in an equivalent manner $(T-t) \sim \varepsilon e^{\sqrt{2|\log(\varepsilon)|}}$. Using (2.38), it would follow that the width of the region occupied for the concentration region would be of order $|x| \sim \sqrt{\varepsilon}$, something that agrees with the size of a developed concentration region as computed in [22].

Actually, the precise computation of the size of the transition between the "blow-up regime" described in section 4 and the "quasi-steady regime" described in [22] is more involved due to the presence of many logarithmic terms in the singularity formation mechanism (cf. (2.37), (2.38)), albeit the main rationale behind the forthcoming computations is basically the elementary computation above.

There are some related mathematical results that have been studied by some authors. In [11] a model has been introduced that also avoids large values of the

chemotactic function for large values of u. On the other hand, in [20] the authors study, using rigorous methods, a possible way of extending the solutions of (1.6), (1.7) for radial solutions. Although the approach and methods used in these papers are different from the one here, some of the basic ideas there are rather close to the approach of this paper.

In order to describe in detail the unfolding of the concentration region starting from the blow-up mechanism described in (1.6), (1.7) we will use the asymptotics of the steady states of (2.5) whose total mass approaches 8π. Such asymptotics has been recalled in section 3, Theorem 3.2. Let us describe here those results using as the variable the mass function $M(r)$ instead of the concentration u. Let us denote as $M_\lambda(r)$ with $\lambda > 0$ the steady state of (2.5) uniquely defined by means of the condition $U(0; M) = \lambda$. By (2.2) we have

$$M_\lambda(r) \sim \frac{\lambda r^2}{2} \quad \text{as } r \to 0^+. \tag{4.1}$$

Combining (3.8) and (2.2) we obtain the following asymptotics for $M_\lambda(r)$ as $\lambda \to 0^+$:

$$M_\lambda(r) \sim \frac{4\bar{\xi}^2}{1+\bar{\xi}^2} + \frac{\alpha\lambda}{8} m(\bar{\xi}) + \cdots \quad \text{as } \lambda \to 0^+, \tag{4.2}$$

where $\bar{\xi} = \frac{\sqrt{\lambda}}{\sqrt{8}} r$ and α is as in (3.7). Therefore

$$M_\lambda(\infty) \sim 4 + \frac{2\alpha\lambda}{3} + \cdots \quad \text{as } \lambda \to 0^+, \tag{4.3}$$

which is just another way of writing (3.9).

4.2. Quadratic terms in the chemotactic function stop aggregation.

4.2.1. Derivation of a differential equation for the size of the concentration region. To describe the unfolding mentioned above we begin analyzing the effect of the boundedness of $Q(\cdot)$ in the computations of section 2. If we just keep the first corrective order, instead of approximating $Q(s)$ just by s, we could approximate $Q(s)$ by $s - \alpha s^2$ as long as $\frac{|\varepsilon M_r|}{r} \ll 1$. In another way, we approximate (2.5) as

$$M_t = M_{rr} - \frac{M_r}{r} + \frac{M M_r}{r} - \alpha\varepsilon M \left(\frac{M_r}{r}\right)^2. \tag{4.4}$$

Our first goal is to compute in a detailed manner the effect of the term $\alpha\varepsilon M(\frac{M_r}{r})^2$ in the aggregation process described in section 2. To this end, it is convenient to make a small change of variables, replacing T in (2.8) by some suitable T_ε to be described in detail later. Clearly, if $\varepsilon > 0$, T_ε does not have the meaning of a blow-up time because the solutions of (4.4) do not blow-up in that case. The variable T_ε will be that suitable time scale satisfying $T_\varepsilon \sim T$ that characterizes when the transition between the blow-up regime described in section 2 and the "quasi-steady regime" in [22] has already taken place.

We define a new set of variables as follows:

$$\begin{aligned} M(r,t) &= \Phi\left(\frac{r}{\sqrt{T_\varepsilon - t}}, -\log(T_\varepsilon - t)\right), \quad y = \frac{r}{\sqrt{T_\varepsilon - t}}, \\ \tau &= -\log(T_\varepsilon - t) + \log\left(\frac{T_\varepsilon}{T}\right), \quad \xi = \frac{r}{\delta(\tau)\sqrt{T_\varepsilon - t}}. \end{aligned} \tag{4.5}$$

The term $\log(\frac{T_\varepsilon}{T})$ in the definition of τ has been added just to obtain the normalization $\tau = \bar{\tau} = -\log(T)$ for $t = 0$.

The term $-\alpha\varepsilon M(\frac{M_r}{r})^2$ will be more relevant in the region $r \approx 0$. It will then be convenient to rewrite (4.4) using the variable ξ in (4.5). Using this variable, (4.4) becomes

$$\delta^2 \frac{\partial M}{\partial \tau} + \left(\frac{\delta^2}{2} - \delta\delta_\tau\right)\xi M_\xi = M_{\xi\xi} - \frac{M_\xi}{\xi} + \frac{M M_\xi}{\xi} - \frac{\alpha\varepsilon}{(T_\varepsilon - t)\delta^2} M \left(\frac{M_\xi}{\xi}\right)^2. \tag{4.6}$$

For $\varepsilon = 0$, M could be approximated to the leading order as $\bar{M}(\xi)$ for $\tau \to \infty$, where $\bar{M}(\xi)$ is as in (2.7). We can expect this approximation to remain valid to the leading order as long as $\frac{\alpha\varepsilon}{(T_\varepsilon - t)\delta^2} \ll 1$. Taking into account the structure of (2.7) it would be natural to look for approximations of M in the form

$$M(\xi, \tau) = \bar{M}(\xi) + \left(\frac{\delta^2}{2} - \delta\delta_\tau\right) W_1(\xi) + \frac{\alpha\varepsilon}{(T_\varepsilon - t)\delta^2} W_2(\xi) + \cdots, \tag{4.7}$$

where $W_1(\xi)$, $W_2(\xi)$ satisfy, respectively,

$$W_{1,\xi\xi} - \frac{1}{\xi} W_{1,\xi} + \frac{\bar{M}}{\xi} W_{1,\xi} + \frac{\bar{M}_\xi}{\xi} W_1 = \xi \bar{M}_\xi, \tag{4.8}$$

$$W_{2,\xi\xi} - \frac{1}{\xi} W_{2,\xi} + \frac{\bar{M}}{\xi} W_{2,\xi} + \frac{\bar{M}_\xi}{\xi} W_2 = \bar{M}\left(\frac{\bar{M}_\xi}{\xi}\right)^2, \tag{4.9}$$

$$W_i(\xi) = o(\xi^2), \quad \xi \to 0^+, \ i = 1, 2. \tag{4.10}$$

Arguing as in [22, section 2], it follows that

$$W_1(\xi) = \frac{4\xi^2}{(\xi^2+1)^2} \int_0^\xi \frac{(1+\eta^2)^2}{\eta^3} \left[\log(1+\eta^2) - \frac{\eta^2}{(1+\eta^2)}\right] d\eta, \tag{4.11}$$

$$W_2(\xi) = m(\xi) = \frac{16}{3} \frac{\xi^2}{(\xi^2+1)^3} (\xi^4 + 4\xi^2 + 2(\xi^2+1)\log(\xi^2+1)). \tag{4.12}$$

Using (4.7) as well as (2.7), (4.11), and (4.12) we obtain the following matching condition:

$$M(\xi, \tau) \sim 4 - \frac{4}{\xi^2} + 4\left(\frac{\delta^2}{2} - \delta\delta_\tau\right)[\log(\xi) - 1] + \frac{16\alpha\varepsilon}{3(T_\varepsilon - t)\delta^2} + \cdots \tag{4.13}$$

as $|\xi| \gg 1$, $|y| \ll 1$.

Expansion (4.7) is not valid if $|x| \sim \sqrt{T_\varepsilon - t}$ as it also happened in section 2. Notice that as long as $\frac{\alpha\varepsilon}{(T_\varepsilon - t)\delta^2} \ll 1$ we can argue as in that section and approximate M as 4. Using then the self-similar variables (2.8) we obtain the following equation for ψ defined as in (1.6):

$$\psi_\tau = \psi_{yy} + \frac{3}{y}\psi_y - \frac{y\psi_y}{2}, \quad y > 0. \tag{4.14}$$

On the other hand, since $\psi \sim -\frac{4\delta^2}{|y|^2}$ as $\delta \ll |y| \ll 1$ (cf. (4.13)), we would obtain that ψ satisfies approximately (2.27) if we want to compute ψ in the region $|y| \sim 1$.

We decompose $\psi(y, \tau)$ as in (2.22), where $a_0(\tau)$ solves (2.24) and $Q(y, \tau)$ satisfies (2.26). As long as $|\delta\delta_\tau| \ll \delta^2$ we can approximate, as in section 2, $Q(y, \tau)$ by means

of the solution of (2.30), (2.31). Therefore, arguing exactly as in section 2, we would arrive at an outer matching condition similar to (2.34) with $\int_\tau^\infty (\delta(s))^2 ds$ replaced by $a_0(\tau)$ there. Matching that expansion with (4.13) we arrive at the following equation:

$$\begin{aligned} a_0(\tau) + (\gamma - \log(4))(\delta(\tau))^2 + \zeta\delta(\tau)\delta_\tau(\tau) \\ = -2(\delta^2 - 2\delta\delta_\tau)\log(\delta) \\ -2(\delta^2 - 2\delta\delta_\tau) + \frac{16\alpha\varepsilon}{3(T_\varepsilon - t)(\delta(\tau))^2} + \cdots . \end{aligned} \tag{4.15}$$

It is interesting to compare (4.15) with (2.36). Notice that the additional term $\frac{16\alpha\varepsilon}{3(T_\varepsilon - t)(\delta(\tau))^2}$ appears due to the presence of corrective terms $Q(s)$ as $s \to 0^+$. Notice that (4.15) has to be complemented with the analogue of (1.7) that we rewrite here by convenience as

$$a_{0,\tau}(\tau) = -(\delta(\tau))^2. \tag{4.16}$$

System (4.15), (4.16) provides a description of the solutions of (2.5) as long as $|\delta_\tau| \ll \delta$. It is important to remark that in the case $\varepsilon > 0$, it is not natural to assume that $\lim_{\tau\to\infty} a_0(\tau) = 0$, since for $\varepsilon > 0$, ψ does not need to approach zero as $\tau \to \infty$. Therefore, in order to determine $a_0(\tau)$ uniquely we will then use its asymptotics for the range of times where the regularizing terms of $Q(s)$ did not begin to act yet. We recall that (2.37), (4.16) imply, for $\varepsilon = 0$, the following asymptotics of $a_0(\tau)$:

$$a_0(\tau) \sim 4e^{-(\gamma+2)}\sqrt{2\bar\tau}e^{-\sqrt{2\bar\tau}}(1 + o(1)) \quad \text{as } \bar\tau \to \infty. \tag{4.17}$$

In order to simplify system (4.15), (4.16) it is convenient to eliminate some numerical constants by means of the following change of variables,

$$\delta(\tau) = 2e^{-\frac{(\gamma+2)}{2}} b(\tau), \qquad a_0(\tau) = 4e^{-(\gamma+2)}\hat a_0(\tau), \tag{4.18}$$

that transforms (4.15), (4.16) into

$$\hat a_0 + \nu b b_\tau + O(b_\tau^2 + b b_{\tau\tau}) = -2(b^2 - 2bb_\tau)\log(b) + \frac{\alpha e^{2(\gamma+2)}\varepsilon e^\tau}{3} \cdot \frac{T}{T_\varepsilon} \cdot \frac{1}{b^2}, \tag{4.19}$$

$$(\hat a_0)_\tau = -b^2, \tag{4.20}$$

where $\nu = \zeta - 4 + 2(\gamma + 2 - \log(4))$.

We now study the precise manner in which the last term in (4.19) and, in general, replacing $G_\varepsilon(s)$ by s in (1.1) modify the dynamics of $\delta(\tau)$, $a_0(\tau)$ (equivalently, b, $\hat a_0(\tau)$). As a first step we need to compute the change of $\delta(\tau)$ due to the fact that instead of solving (1.1), (1.2) with $G_\varepsilon(s)$ we use s instead. To the leading order we approximate $G_\varepsilon(s)$ by $s - \alpha\varepsilon s^2$. By assumption we solve (1.1), (1.2) with the same initial data $u_0(x)$ but choose either $\varepsilon = 0$ or $\varepsilon > 0$. Using the set of self-similar variables (4.5), (4.4) becomes

$$\Phi_\tau = \Phi_{yy} - \frac{y\Phi_y}{2} - \frac{\Phi_y}{y} + \frac{\Phi\Phi_y}{y} - \alpha\varepsilon e^\tau \frac{T}{T_\varepsilon}\Phi\left(\frac{\Phi_y}{y}\right)^2 \tag{4.21}$$

with initial data $\Phi_0(\tau) = u_0(\frac{r}{\sqrt{T_\varepsilon}})$. Notice that in the case $\varepsilon = 0$, (4.21) would be equivalent to (1.6) (with $\bar y$ replaced by y and $\bar\tau$ by τ). The corresponding initial data for $\varepsilon = 0$ would be $\bar\Phi(\bar y, -\log(T)) = \bar\Phi_0(\bar y) = u_0(\frac{r}{\sqrt{T}})$.

Let us write $\psi(\bar{y},\bar{\tau}) = \Phi(\bar{y},\bar{\tau}) - \bar{\Phi}(\bar{y},\bar{\tau})$. Notice that we are not choosing the independent variables for Φ as (y,τ) but instead we are taking as variables for this function the ones associated to $\bar{\Phi}$. This will simplify some computations later. To the leading order ψ solves

$$\psi_{\bar{\tau}} = \psi_{\bar{y}\bar{y}} - \frac{\bar{y}\psi_{\bar{y}}}{2} - \frac{\psi_{\bar{y}}}{\bar{y}} + \bar{\Phi}\frac{\psi_{\bar{y}}}{\bar{y}} + \frac{\bar{\Phi}_{\bar{y}}}{\bar{y}}\psi - \frac{\alpha\varepsilon T e^{\bar{\tau}}}{T_\varepsilon}\bar{\Phi}\left(\frac{\bar{\Phi}_{\bar{y}}}{\bar{y}}\right)^2 \tag{4.22}$$

with initial data

$$\psi(\bar{y}, -\log(T)) = -\frac{1}{2T}\bar{y}u_0'(\bar{y})(T_\varepsilon - T). \tag{4.23}$$

The difference ψ cannot be computed using the approximation (4.22), (4.23) near the blow-up time. However, we intend to use this approximation only as long as ψ is small enough compared with $\bar{\Phi}$. For later times the use of the whole nonlinear problem will be needed.

The next set of arguments is reminiscent of those used in [21], although with a different set of functions. Using the inner variable $\bar{\xi} = \frac{y}{\bar{\delta}(\tau)}$, (4.22) becomes

$$\bar{\delta}^2\psi_{\bar{\tau}} + \left(\frac{\bar{\delta}^2}{2} - \bar{\delta}\bar{\delta}_{\bar{\tau}}\right)\bar{\xi}\psi_{\bar{\xi}} = \psi_{\bar{\xi}\bar{\xi}} - \frac{\psi_{\bar{\xi}}}{\bar{\xi}} + \bar{\Phi}\frac{\psi_{\bar{\xi}}}{\bar{\xi}} + \frac{\bar{\Phi}_{\bar{\xi}}}{\bar{\xi}}\psi - \frac{\alpha\varepsilon T e^{\bar{\tau}}}{T_\varepsilon\bar{\delta}^2}\bar{\Phi}\left(\frac{\bar{\Phi}_{\bar{\xi}}}{\bar{\xi}}\right)^2. \tag{4.24}$$

We expand ψ in a series of terms having different relative sizes. We write

$$\psi = \psi_0 + \psi_1 + \cdots, \tag{4.25}$$

where ψ_0, ψ_1 solve the problems

$$\psi_{0,\bar{\xi}\bar{\xi}} - \frac{\psi_{0,\bar{\xi}}}{\bar{\xi}} + \bar{\Phi}\frac{\psi_{0,\bar{\xi}}}{\bar{\xi}} + \frac{\bar{\Phi}_{\bar{\xi}}}{\bar{\xi}}\psi_0 = 0, \tag{4.26}$$

$$\psi_{1,\bar{\xi}\bar{\xi}} - \frac{\psi_{1,\bar{\xi}}}{\bar{\xi}} + \bar{\Phi}\frac{\psi_{1,\bar{\xi}}}{\bar{\xi}} + \frac{\bar{\Phi}_{\bar{\xi}}}{\bar{\xi}}\psi_1 = \psi_{0,\tau} + \left(\frac{\bar{\delta}^2}{2} - \bar{\delta}\bar{\delta}_{\bar{\tau}}\right)\bar{\xi}\psi_{0,\bar{\xi}} + \frac{\alpha\varepsilon T e^{\bar{\tau}}}{T_\varepsilon\bar{\delta}^2}\bar{\Phi}\left(\frac{\bar{\Phi}_{\bar{\xi}}}{\bar{\xi}}\right)^2. \tag{4.27}$$

We choose ψ_0 in such a way that $\lim_{\xi\to 0^+}\frac{\psi(\bar{\xi},\bar{\tau})}{\psi_0(\bar{\xi},\bar{\tau})} = 1$. Since ψ, ψ_0, ψ_1 behave quadratically in $\bar{\xi}$ as $\bar{\xi}\to 0^+$ this means that we are assuming that $\psi_k(\bar{\xi},\bar{\tau}) = o(\bar{\xi}^2)$ as $\bar{\xi}\to 0^+$, $k = 1, 2, \ldots$.

The required solution of (4.26) is then given by

$$\psi_0(\bar{\xi},\bar{\tau}) = A(\bar{\tau})\varphi(\bar{\xi}), \tag{4.28}$$

where

$$\varphi(\bar{\xi}) = \frac{\bar{\xi}^2}{(\bar{\xi}^2+1)^2} \tag{4.29}$$

and $A(\bar{\tau}) = \lim_{\bar{\xi}\to 0}(\frac{\psi(\bar{\xi},\bar{\tau})}{\bar{\xi}^2})$. We can then solve (4.27) using variation of constants. After some computations we finally obtain the following:

$$\psi_1(\bar{\xi},\bar{\tau}) = \{W_1(\bar{\xi}) + W_2(\bar{\xi}) + W_3(\bar{\xi})\}\varphi(\bar{\xi}), \tag{4.30}$$

where

$$W_1(\bar\xi) \equiv \frac{A_{\bar\tau}\bar\delta^2}{2}\int_0^{\bar\xi}\frac{(\eta^2+1)}{\eta^3}((\eta^2+1)\log(\eta^2+1)-\eta^2)d\eta,$$
$$W_2(\bar\xi) \equiv A\left(\frac{\bar\delta^2}{2}-\bar\delta\bar\delta_{\bar\tau}\right)\int_0^{\bar\xi}\left(\frac{1+2\eta^2}{\eta}-\frac{(\eta^2+1)^2}{\eta^3}\log(\eta^2+1)\right)d\eta,$$
$$W_3(\bar\xi) \equiv \frac{16\alpha\varepsilon e^{\bar\tau}\left(\frac{T}{T_\varepsilon}\right)}{3\bar\delta^2}\left(\frac{\bar\xi^4+4\bar\xi^2}{\bar\xi^2+1}+2\log(\bar\xi^2+1)\right).$$

Using (4.25), (4.28), (4.29), (4.30) we can compute the asymptotics of $\psi(\bar\xi,\bar\tau)$ as $\bar\xi\to\infty$. We then deduce the following matching condition for the inner solution $\psi(\bar\xi,\bar\tau)$:

$$\psi(\bar\xi,\bar\tau)\sim \frac{A(\bar\tau)}{\bar\xi^2}+\left[\frac{A_{\bar\tau}\bar\delta^2}{2}+A\left(\bar\delta\bar\delta_{\bar\tau}-\frac{\bar\delta^2}{2}\right)\right]\left(\log(\bar\xi)-\frac{3}{2}\right) + \frac{A_{\bar\tau}\bar\delta^2}{4}+\frac{16\alpha T\varepsilon e^{\bar\tau}}{3T_\varepsilon\bar\delta^2}+\cdots \tag{4.31}$$

for $1\ll\bar\xi\ll\frac{1}{\bar\delta^2}$.

We now proceed to obtain an outer expansion for ψ. Taking into account the outer expansion for $\bar\Phi$ obtained in section 2 (cf. (2.34)) and neglecting terms of the form $O(\bar\delta^2\psi)$ as well as the term $\frac{\alpha T\varepsilon e^{\bar\tau}}{T_\varepsilon}\bar\Phi(\frac{\bar\Phi_{\bar y}}{\bar y})^2$ that will be negligible compared with $\bar\delta^4$ in the range of times where we will use the approximation of ψ currently under computation, we obtain the following equation for ψ:

$$\psi_{\bar\tau}=\psi_{\bar y\bar y}+\frac{3}{\bar y}\psi_{\bar y}-\frac{\bar y\psi_{\bar y}}{2}. \tag{4.32}$$

The matching condition (4.31) suggests that the order of magnitude of ψ is $A\bar\delta^2$. It is then convenient to introduce a new function G by means of the formula $\psi=A\bar\delta^2G$. Function G satisfies to the leading order

$$\frac{(A\bar\delta^2)_{\bar\tau}}{A\bar\delta^2}G+G_{\bar\tau}=G_{\bar y\bar y}+\frac{3}{\bar y}G_{\bar y}-\frac{\bar yG_{\bar y}}{2}. \tag{4.33}$$

Notice that the exponential growth of the last term in (4.31) indicates that for $\bar\tau$ large, A grows like $e^{\bar\tau}$, with perhaps some algebraic-like corrections. On the other hand, one can expect that A should contain some contribution behaving like those of the homogeneous part of (4.24). The leading part of such contributions also grows like $e^{\bar\tau}$, due to the fact that those terms are associated to changes in the blow-up time for solutions of (2.6) (see the related analysis in [21]). In another manner, it should be $\lim_{\tau\to\infty}\frac{(A\bar\delta^2)_{\bar\tau}}{A\bar\delta^2}=1$. It is then natural to look for expansions for G in the form

$$G(\bar y,\bar\tau)=G_0(\bar y,\bar\tau)+G_1(\bar y,\bar\tau)+\cdots, \tag{4.34}$$

where (cf. (4.33))

$$G_0=G_{0,\bar y\bar y}+\frac{3}{\bar y}G_{0,\bar y}-\frac{\bar yG_{0,\bar y}}{2}, \tag{4.35}$$

$$\left[\frac{(A\bar\delta^2)_{\bar\tau}}{A\bar\delta^2}-1\right]G_0+(G_{0,\bar\tau})+G_1=G_{1,\bar y\bar y}+\frac{3}{\bar y}G_{1,\bar y}-\frac{\bar yG_{1,\bar y}}{2}. \tag{4.36}$$

Higher order corrections would provide contributions of order $O((\frac{(A\bar\delta^2)_{\bar\tau}}{A\bar\delta^2}-1)^2+(\frac{(A\bar\delta^2)_{\bar\tau}}{A\bar\delta^2})_{\bar\tau})$. Our goal is to include in G_0 the singular contributions of G as $\bar y\to 0^+$. We then select G_0 satisfying the matching condition $G_0\sim\frac{1}{\bar y^2}$ as $\bar y\to 0^+$. We also assume that the corresponding solutions of (4.35), (4.36) do grow exponentially as $\bar y\to\infty$ because such growths would rule out the possibility of getting matchings with bounded solutions in the region $x\sim 1$. It then follows that

$$G_0(\bar y,\bar\tau)=\frac{1}{\bar y^2}. \tag{4.37}$$

We can then easily solve (4.36). Taking into account (4.31) it follows that we have to choose G_1 bounded as $\bar y\to 0$. The corresponding solution of (4.36) is given by

$$G_1(\bar y,\bar\tau)=-\left(\frac{(A\bar\delta^2)_{\bar\tau}}{A\bar\delta^2}-1\right)\frac{1}{\bar y^2}\int_0^{\bar y}\upsilon e^{\frac{\upsilon^2}{4}}\left(\int_\upsilon^\infty e^{-\frac{\eta^2}{4}}\frac{d\eta}{\eta}\right)d\upsilon. \tag{4.38}$$

Computing the asymptotics of G_1 as $\bar y\to 0$, and also using (4.37) as well as the fact that $\psi=A\bar\delta^2G$, we finally obtain the following matching condition for $\psi(\bar y,\bar\tau)$:

$$\psi\sim\frac{A\bar\delta^2}{\bar y^2}-\frac{A\bar\delta^2}{2}\left(\frac{(A\bar\delta^2)_{\bar\tau}}{A\bar\delta^2}-1\right)\left[-\frac12\log(\bar y)+\frac14(1+\log(4)-\gamma)\right]+\cdots \tag{4.39}$$

as $\bar y\to 0^+$. We then use $\bar\xi=\frac{\bar y}{\bar\delta}$ in order to match (4.31) with (4.39) to derive the following differential equation for $A(\bar\tau)$:

$$\begin{aligned}&-\left(\frac32+\log(\bar\delta)\right)\left(\frac{A_{\bar\tau}\bar\delta^2}{2}+A\left(\bar\delta\bar\delta_{\bar\tau}-\frac{\bar\delta^2}{2}\right)\right)+\frac{A_{\bar\tau}\bar\delta^2}{4}+\frac{16\alpha\varepsilon e^{\bar\tau}\left(\frac{T}{T_\varepsilon}\right)}{3\bar\delta^2}\\ &\quad=-\frac12(1+\log(4)-\gamma)\left[\frac{A_{\bar\tau}\bar\delta^2}{2}+A\left(\bar\delta\bar\delta_{\bar\tau}-\frac{\bar\delta^2}{2}\right)\right]\\ &\qquad+O\left(\left[\left(\frac{A_{\bar\tau}}{A}-1\right)^2+\left(\frac{\bar\delta_{\bar\tau}}{\bar\delta}\right)^2+\left(\frac{A_{\bar\tau}}{A}\right)_{\bar\tau}+\left(\frac{\bar\delta_{\bar\tau}}{\bar\delta}\right)_{\bar\tau}\right]A\bar\delta^2\right).\end{aligned} \tag{4.40}$$

This is the desired equation that will be used to study the evolution of the size of the concentration region during the transition between the blow-up regime and the quasi-steady regime.

4.2.2. Study of the solutions of the differential equation describing the size of the transition region. In this subsection we study the consequences of (4.40). As a first step notice that we can simplify this expression using the change of variables $\bar\delta=2e^{-\frac{(\gamma+2)}{2}}\bar b$. Equation (4.40) then becomes

$$\begin{aligned}&-\log(\bar b)\left(\frac{A_{\bar\tau}\bar b^2}{2}+A\left(\bar b\bar b_{\bar\tau}-\frac{\bar b^2}{2}\right)\right)+\frac{A_{\bar\tau}\bar b^2}{4}+\frac{16\alpha\varepsilon e^{\bar\tau}\left(\frac{T}{T_\varepsilon}\right)}{3\bar\delta^2}\\ &\quad=O\left(\left[\left(\frac{A_{\bar\tau}}{A}-1\right)^2+\left(\frac{\bar b_{\bar\tau}}{\bar b}\right)^2+\left(\frac{A_{\bar\tau}}{A}\right)_{\bar\tau}+\left(\frac{\bar b_{\bar\tau}}{\bar b}\right)_{\bar\tau}\right]A\bar b^2\right).\end{aligned}$$

The asymptotics for the solutions of this equation can be obtained using the change of variables $A=e^{\bar\tau}h$, in order to eliminate the exponential growth of A. Using

(2.37) it follows that

$$A(\bar\tau) \sim -\frac{2e^{2(\gamma+2)}\alpha\varepsilon}{3}\left(\frac{T}{T_\varepsilon}\right)e^{\bar\tau}e^{2\sqrt{2\bar\tau}} + C_\varepsilon e^{\bar\tau} \quad \text{as } \bar\tau \to \infty, \tag{4.41}$$

where C_ε is a real constant to be determined later.

Let us compute now the precise manner in which $\delta(\tau)$ differs from $\bar\delta(\bar\tau)$ due to the change of Φ to $\bar\Phi$ previously computed. We recall that $\psi(\bar y,\bar\tau) = \Phi(\bar y,\bar\tau) - \bar\Phi(\bar y,\bar\tau)$. On the other hand, we have defined $\delta(\tau)$ by means of the normalization $\Phi(y,\tau) \sim 4\xi^2 = 4(\frac{y}{\delta(\tau)})^2$ for $y \ll \delta(\tau)$, with $\xi = \frac{r}{\sqrt{T_\varepsilon - t(\delta(\tau))}}$, (cf. (4.7), (4.10)). The definition of $\bar\delta(\bar\tau)$ is similar, namely, $\bar\Phi(\bar y,\bar\tau) \sim 4(\frac{\bar y}{\bar\delta(\bar\tau)})^2$ for $\bar y \ll \bar\delta(\bar\tau)$. Since $\bar y$, $\bar\tau$ above have been used as dummy variables in the difference $\Phi(\bar y,\bar\tau) - \bar\Phi(\bar y,\bar\tau)$, we obtain, using (4.28), that

$$\Phi(\bar y,\bar\tau) \sim 4\left(\frac{\bar y}{\bar\delta(\bar\tau)}\right)^2 + A(\bar\tau)\left(\frac{\bar y}{\bar\delta(\bar\tau)}\right)^2$$

for $\bar y \ll \bar\delta(\bar\tau)$. Then $\frac{4}{(\delta(\bar\tau))^2} \sim \frac{4+A(\bar\tau)}{(\bar\delta(\bar\tau))^2}$ or also, using Taylor's theorem, $\delta(\bar\tau) \sim \bar\delta(\bar\tau)(1 - \frac{A(\bar\tau)}{8})$ as $\bar\tau \to \infty$. Using the relations between δ, b and $\bar\delta$, $\bar b$, as well as (4.18), (4.41), we obtain the following asymptotics for $b(\bar\tau)$:

$$b(\bar\tau) \sim e^{-\sqrt{\frac{\bar\tau}{2}}}\left(1 + \frac{e^{2(\gamma+2)}\alpha\varepsilon}{12}\left(\frac{T}{T_\varepsilon}\right)e^{\bar\tau}e^{2\sqrt{2\bar\tau}} - \frac{C_\varepsilon}{8}e^{\bar\tau}\right) \quad \text{as } \bar\tau \to \infty. \tag{4.42}$$

At this point, it remains only to compute the value of C_ε. Notice that the linearity of (4.22), (4.23) on $\varepsilon(\frac{T}{T_\varepsilon})$, $(T - T_\varepsilon)$ implies for C_ε the functional dependence $C_\varepsilon = C_1\varepsilon(\frac{T}{T_\varepsilon}) + C_2(T - T_\varepsilon)$, where C_1, C_2 are fixed real constants. The precise value of C_1 will not be relevant here and therefore it will not be computed in detail. However, the value of C_2 will be needed later, and therefore we proceed to compute it. In order to do this we would need to solve (4.22), (4.23), dropping the last term in (4.22). Notice however, that the argument that yields to (4.22), (4.23) implies that this problem is equivalent to computing the change in $\bar\delta(\bar\tau)$ due to using the variables (4.5) instead of (2.8) in the solution of (1.6). In order to do this, notice that by definition of $\bar\delta(\bar\tau)$,

$$\bar M(r,t) \sim \frac{4r^2}{(T-t)(\bar\delta(\bar\tau))^2} \quad \text{for } \mathrm{r} \ll \sqrt{\mathrm{T} - \mathrm{t}},\ \mathrm{t} \to \mathrm{T}^-. \tag{4.43}$$

On the other hand, let us denote as $\hat\delta(\tau)$ the corresponding function, which would be obtained in a similar manner, but assuming that the set of variables (4.5) is used instead. In another way, we assume that

$$\bar M(r,t) \sim \frac{4r^2}{(T_\varepsilon - t)(\tilde\delta(\tau))^2} \quad \text{for } \mathrm{r} \ll \sqrt{\mathrm{T}_\varepsilon - \mathrm{t}}, \approx \to \mathrm{T}_\varepsilon^-. \tag{4.44}$$

In order to compute the value of C_2 we need to compute the difference between $\hat\delta$ and $\bar\delta$ to the linear order for times not too close to the blow-up time. More precisely, let us assume that $|T - t| \gg |T_\varepsilon - T|$. Formulae (4.43), (4.44) combined with Taylor's expansion imply for this range of times that

$$\hat\delta(\tau) = \bar\delta(\bar\tau) + \frac{(T - T_\varepsilon)}{2}e^{\bar\tau}\bar\delta(\bar\tau) + \cdots. \tag{4.45}$$

It remains to estimate the correction due to the change of variables from $\bar\tau$ to τ, in order to have both functions $\bar\delta$, $\hat\delta$ written in the same variables. Notice that $|\bar\delta_{\bar\tau}| = O(\frac{\bar\delta}{\sqrt{\bar\tau}})$ and, on the other hand, $\tau - \bar\tau = O((T-T_\varepsilon)e^{\bar\tau})$ for $|T-t| \gg |T_\varepsilon - T|$. Henceforth $\bar\delta(\bar\tau) - \bar\delta(\tau) = O(\frac{(T-T_\varepsilon)e^{\bar\tau}\bar\delta(\bar\tau)}{\sqrt{\bar\tau}})$, and since this term is negligible compared with $\frac{(T-T_\varepsilon)}{2}e^{\bar\tau}\bar\delta(\bar\tau)$, we can ignore the fact that in different sides of (4.45) we are using different variables. We then finally obtain

$$\hat\delta(\bar\tau) = \bar\delta(\bar\tau) + \frac{(T-T_\varepsilon)}{2}e^{\bar\tau}\bar\delta(\bar\tau) + \cdots .$$

Using then the analogue of (4.42) that would be obtained for this problem, it then follows that $C_2 = -4$. Due to the linearity of the problem, we can then rewrite (4.42) as

$$\begin{aligned} b(\bar\tau) \sim e^{-\sqrt{\frac{\bar\tau}{2}}} \Bigg(1 + \frac{e^{2(\gamma+2)}\alpha}{12}\left(\frac{T}{T_\varepsilon}\right)\varepsilon e^{\bar\tau}e^{2\sqrt{2\bar\tau}} \\ + \left(C_3\frac{T}{T_\varepsilon}\varepsilon + \frac{(T-T_\varepsilon)}{2}\right)e^{\bar\tau}\Bigg) \quad \text{as } \bar\tau \to \infty, \end{aligned} \tag{4.46}$$

where $C_3 = -\frac{C_1}{8}$ is a real constant.

Remark 4.1. It is rather natural to ask if it would not be possible to simplify the rather cumbersome argument that yields (4.46) using instead the ODEs (4.19), (4.20) that provide a description of the width of the region where the mass is concentrated. The idea would be to use the fact that $b = \bar b(\tau)$, $\hat a_0(\tau) = \int_\tau^\infty (\bar b(s))ds$ provide a solution of (4.19), (4.20), with $\bar b = \frac{\bar\delta}{2e^{-\frac{(\gamma+2)}{2}}}$ and $\varepsilon = 0$. Linearizing around this solution in (4.19), (4.20) it would be possible to derive a linear set of ODEs that in principle could be used to derive (4.46). Actually such a method can be used to derive the term $\frac{4\alpha\varepsilon(\frac{T}{T_\varepsilon})e^{\bar\tau}e^{2\sqrt{2\bar\tau}}}{3(2e^{-\frac{(\gamma+2)}{2}})^4}$ in (4.46). Unfortunately, the last terms of (4.46) cannot be derived in that manner. The reason is the following. In (4.19) there are corrective terms that are neglected in the computations. A typical representative term of this type of term is, for instance, $bb_{\tau\tau}$. Suppose that we write $b = \bar b + \kappa$ and formally linearize. The term $bb_{\tau\tau}$ would generate terms in the linearized problem having the form $\bar b\kappa_{\tau\tau}$, whose effect should be compared with some of the terms arising from the leading terms in (4.19); a typical one would have the form $\bar b\log(\bar b)\kappa$. It is not hard to see, linearizing in the term $-2(b^2 - 2bb_\tau)\log(b)$, that κ contains as its main factor a term growing exponentially. However, the term $\bar b\kappa_{\tau\tau}$, as well as the relative sizes of the terms $\bar b$ and $\bar b\log(\bar b)$ (whose order is $\frac{1}{\sqrt{\bar\tau}}$), would generate "algebraic-like" corrections that would modify the exponential growth of κ. In particular, it would not be possible to determine using such a linearization argument if the last term in (4.46) has exactly the functional dependence $e^{\bar\tau}$ or, say, $e^{\bar\tau}(\bar\tau)^a$ for some $a > 0$. Unfortunately, this information will be relevant later. On the other hand, this problem could not be solved by just computing one additional term in the matching conditions that yields (4.19), because the same problem would be produced by $bb_{\tau\tau\tau}$ and similar corrective terms. The real reason for the difficulty lies in the exponential growth of the correction κ. This difficulty does not arise in computing b by means of (4.19) because b is an "algebraic-like" (in contrast with "exponential") function, and therefore differentiating on τ we obtain smaller functions as $\tau \to \infty$. This makes it possible to neglect terms like $bb_{\tau\tau}$ and similar ones. This argument illustrates some

difficulties that might lie in a naive handling of equations that have been derived by means of formal arguments and explains why the tortuous path that led to (4.46) has been used.

4.3. Regularizing terms yield a fully developed concentration region. After deriving (4.46) we now study another problem that we must consider in order to describe the process of concentration region formation. A crucial argument in the derivation of (4.19), (4.20) is the assumption $|\delta_\tau| \ll \delta$ or, equivalently, $|b_\tau| \ll b$. However, the exponential growth of the corrective terms in (4.46) implies that this assumption fails if $\bar\tau$ is large enough. In order to study the precise manner in which this failure takes place we use a standard boundary layer argument. We introduce a new set of variables by means of

$$b = e^{-\sqrt{\frac{\tau_\varepsilon}{2}}} h, \quad \hat a_0 = \sqrt{2\tau_\varepsilon} e^{-\sqrt{2\tau_\varepsilon}} \varphi, \quad \bar\tau = \tau_\varepsilon + s, \tag{4.47}$$

where τ_ε is defined by the following formula:

$$\frac{\alpha e^{2(\gamma+2)} \varepsilon e^{\tau_\varepsilon} e^{2\sqrt{2\tau_\varepsilon}}}{3} = \sqrt{2\tau_\varepsilon}. \tag{4.48}$$

This change of variables transforms (4.19), (4.20) into

$$\begin{aligned} &\sqrt{2\tau_\varepsilon}\varphi + \nu h h_s + O(h_s^2 + h h_{ss}) \\ &\qquad = \sqrt{2\tau_\varepsilon}(h^2 - 2hh_s) - 2\log(h)(h^2 - 2hh_s) + \sqrt{2\tau_\varepsilon}\frac{e^s}{h^2}, \end{aligned} \tag{4.49}$$

$$\varphi_s = -\frac{h^2}{\sqrt{2\tau_\varepsilon}} \tag{4.50}$$

whence, since $\tau_\varepsilon \to \infty$, we obtain to the leading order the problem

$$\varphi = h^2 - 2hh_s + \frac{e^s}{h^2}, \tag{4.51}$$

$$\varphi_s = 0. \tag{4.52}$$

Using (4.17) and (4.48) we obtain the matching condition $\varphi \to 1$ as $s \to -\infty$. The system (4.51), (4.52) can then be reduced to the ODE

$$1 = h^2 - 2hh_s + \frac{e^s}{h^2}. \tag{4.53}$$

On the other hand, (4.46) provides, after some computations, the following matching condition:

$$h(s) \sim 1 + \frac{s}{2}e^s + \left(\frac{3}{2} \frac{(T - T_\varepsilon)\sqrt{2\tau_\varepsilon} e^{-2(\gamma+2)}}{\alpha \varepsilon e^{2\sqrt{2\tau_\varepsilon}}} + \frac{\sqrt{2\tau_\varepsilon}}{4} \right) e^s + \cdots \tag{4.54}$$

as $s \to -\infty$. Notice that in deriving (4.54) we are assuming that $(T - T_\varepsilon) \ll 1$, and therefore $\frac{T}{T_\varepsilon} \to 1$. Notice also that we are neglecting the contribution of the term $C_3 \frac{T}{T_\varepsilon} \varepsilon e^{\bar\tau}$ in (4.46) that is negligible if compared with the previous term there.

Before finding the explicit solution of (4.53), (4.54) we briefly discuss its range of validity. Notice that in deriving (4.19), (4.20) we repeatedly use the assumption $|\delta_\tau| \ll \delta$ or, equivalently, $|h_s| \ll h$. This requirement is valid as $s \to -\infty$, but it will

be lost as s becomes of order one. It turns out, however, that the validity of (4.53) can be extended until the range of times in which h becomes of order one. In order to check this, we examine the changes that are required in the previous arguments that led to (4.19), (4.20) if the condition $|\delta_\tau| \ll \delta$ is lost. First, we remark that expansion (4.13) remains valid even if $|\delta_\tau| \sim \delta$, as long as δ remains small. However, if $|\delta_\tau| \sim \delta$, it is not possible to approximate the solutions of (4.22) by means of those of (2.20). We are then forced to study the complete solution of (4.22) as soon as the requirement $|\delta_\tau| \ll \delta$ is lost. Also using the asymptotics of $Q(y,\tau)$ we are then led to study the following problem instead of (2.27):

$$Q_s = Q_{yy} + \frac{3}{y}Q_y - \frac{yQ_y}{2} - 32\Gamma(\delta(s))^2\left[\delta^{(4)}(y) - \frac{1}{\langle 1,1\rangle}\right], \tag{4.55}$$

$$Q(y,s) = (\delta(s))^2\Omega(y) \quad \text{as } s \to -\infty, \tag{4.56}$$

where (4.18), (4.47), (4.54) provide a matching condition for $\delta(s)$ as $s \to -\infty$.

It is possible to write an explicit formula for $Q(y,s)$ using caloric kernels, as well as the following asymptotics for $Q(y,s)$ as $y \to 0^+$:

$$Q(y,s) \sim -\frac{4(\delta(s))^2}{y^2} + 4\left(\frac{\delta^2}{2} - \delta\delta_s\right)\log(y) + \lambda(s) + o(1), \tag{4.57}$$

where $\lambda(s)$ depends on the values of $\delta(\tilde{s})$ for $\tilde{s} \le s$. It would be possible to write this dependence using an explicit integral formula but this will not be needed. The only relevant information that we need at this point is that as far as $h(\tilde{s})$ remains bounded for $\tilde{s} \le s$ (equivalently, $|\delta(\tilde{s})| = O(e^{-\sqrt{\frac{\tau_\varepsilon}{2}}})$ for $\tilde{s} \le s$), then $|\lambda(s)| = O(e^{-\sqrt{2\tau_\varepsilon}})$. If we then match the asymptotics (4.13) with (4.57) we derive the following equation that would generalize (4.15) for $|\delta_\tau| \sim \delta$:

$$a_0(s) + \lambda(s) = -2(\delta^2 - 2\delta\delta_s)\log(\delta) - 2(\delta^2 - 2\delta\delta_s) + \frac{16\alpha\varepsilon}{3(T_\varepsilon - t)\delta^2}. \tag{4.58}$$

Using (4.18), (4.47) as well as the above-mentioned fact that $|\lambda(s)| = O(e^{-\sqrt{2\tau_\varepsilon}})$, it follows that to the leading order (4.53) remains valid as long as $h = O(1)$.

Having asserted the validity of (4.53) in the desired region of times we proceed to solve it with the matching condition (4.54). To this end we introduce a new variable $w = h^2e^{-s}$. Writing (4.53) in terms of w instead of h we would obtain

$$e^s w_s = \frac{1}{w} - 1$$

that is a separable equation with solution

$$w + \log(w-1) = e^{-s} + C, \tag{4.59}$$

where C is a real number. Deriving (4.59) we are implicitly assuming that $w > 1$, a property that can be expected at least for $s \to -\infty$ due to (4.54).

Using (4.59) we obtain the following asymptotics for $h(s)$:

$$h(s) \sim 1 + \frac{s}{2}e^s + \frac{C}{2}e^s + \cdots \quad \text{as } s \to -\infty, \tag{4.60}$$

$$h(s) \sim Ke^{\frac{s}{2}} + \cdots \quad \text{as } s \to \infty, \tag{4.61}$$

where

$$e^{K^2}(K^2-1)=e^C. \tag{4.62}$$

Combining (4.54) and (4.60) we then obtain the following value of C:

$$C=\frac{3(T-T_\varepsilon)\sqrt{2\tau_\varepsilon}e^{-2(\gamma+2)}}{\alpha\varepsilon e^{2\sqrt{2\tau_\varepsilon}}}+\frac{\sqrt{2\tau_\varepsilon}}{2}. \tag{4.63}$$

Notice that for any $C>0$ the solution of (4.53) is defined globally for $-\infty<x<\infty$. Indeed, notice that h cannot approach zero for a finite value of s because (4.53) implies that for h small enough h_s is positive. On the other hand, (4.53) implies that h_s is linearly bounded on h for h large.

At this point we need to choose T_ε. We will choose T_ε in such a way that C in (4.63) approaches $-\infty$ as $\varepsilon\to 0$. In another way, we assume that

$$\varepsilon e^{2\sqrt{2\tau_\varepsilon}}\sim\varepsilon e^{2\sqrt{2\log(\frac{1}{\varepsilon})}}\ll(T_\varepsilon-T)\ll 1. \tag{4.64}$$

At first glance the choice (4.64) might look a bit artificial, but on the contrary it turns out that it is rather natural. Choosing T_ε in this manner we will obtain that K in (4.61), (4.62) approaches its minimum value $K=1$ as $\varepsilon\to 0$. As we will show briefly, this choice of K (i.e., $K\approx 1$) will imply that for $t\approx T_\varepsilon$, $M(r,t)$ behaves approximately near the origin as one of the steady states $M_\lambda(r)$ described at the beginning of this section (cf. (4.1), (2.37)), having an amount of mass $4+a_0(\tau_\varepsilon)$ (cf. (4.3)). If the choice $K\approx 1$ had not been made, $M(r,T_\varepsilon)$ would have a width much larger than the steady state associated to mass $4+a_0(\tau_\varepsilon)$ or, in a more precise manner, it would not be possible to describe $M(r,T_\varepsilon)$ using a steady state. The choice of (4.64) essentially means that T_ε has been chosen close to T but at the same time $|T_\varepsilon-T|$ is large enough to allow stabilization of the solutions to steady states. Of course, if T_ε is not chosen satisfying (4.64) the aspect of the solution would not change. However, at $t=T_\varepsilon$, $M(r,T_\varepsilon)$ would not yet be a steady state, and some additional analysis would be required to describe how such approximation to a steady state would take place. The choice (4.64) (or $K\approx 1$) just simplifies the description of $M(r,t)$ in an analytical manner. Some intuitive understanding of the choice of T_ε can be acquired if we assume instead in a provisional manner that $T_\varepsilon=T$. In that case (4.63) would imply that $C=\frac{\sqrt{2\tau_\varepsilon}}{2}\to\infty$, and in that case $K\to\infty$. Therefore the width of the region where the mass is distributed would be larger (cf. (4.61)). In other words, the term $-\alpha\varepsilon M(\frac{M_r}{r})^2$ in (4.4) slows down the process of chemotactic aggregation as it could be expected.

Taking into account (4.18), (4.47) as well as the asymptotics (4.61) we obtain

$$\delta(s)\sim Ke^{-\sqrt{\frac{\tau_\varepsilon}{2}}}e^{\frac{s}{2}}=Ke^{-\sqrt{\frac{\tau_\varepsilon}{2}}}e^{-\frac{\tau_\varepsilon}{2}}e^{\frac{\bar\tau}{2}}\quad\text{as } s=\bar\tau-\tau_\varepsilon\to\infty, \tag{4.65}$$

where due to our choice of T_ε, K has to be set as one.

In order to describe the asymptotics of M (or Φ) as $s\to\infty$ we remark that as long as $\delta(s)\ll 1$ we can use the approximation (4.14) with the matching condition $\psi\sim-\frac{4\delta^2}{y^2}$ as $y\to 0^+$. It is natural to decompose ψ as in (2.22). Functions $a_0(s)$, $Q(y,s)$ evolve according to the equations (2.24), (2.26), respectively. To the leading order, and taking into account (4.50) (cf. also (4.18), (4.47)), $a_0(s)$ remains approximately constant if s remains of order one. However, this ceases being so if $s\to\infty$, due to the

exponential growth of $\delta(s)$. Integrating (2.24) we obtain the following approximation for $a_0(s)$:

$$a_0(s) \approx a_0(\tau_\varepsilon) - K^2 e^{-\sqrt{2\tau_\varepsilon}} e^s \tag{4.66}$$

for s of order one, or larger. Function $Q(y,s)$ can be approximated in the form $e^s W(y)$ as $s \to \infty$. It is easier to compute this approximation in the following manner. Formula (4.66) suggests looking for approximations of ψ in the form

$$\psi = a_0(\tau_\varepsilon) + (\delta(s))^2 Z(y)$$

with Z satisfying

$$Z = Z_{yy} + \frac{3}{y} Z_y - \frac{y Z_y}{2} - 32\Gamma \delta^{(4)}(y). \tag{4.67}$$

The solution of (4.67) can be computed explicitly as

$$Z = -\frac{4}{y^2}.$$

Therefore, to the leading order

$$\psi \sim a_0(\tau_\varepsilon) - \frac{4(\delta(s))^2}{y^2} \quad \text{as } s \to \infty,\ y \gg \delta(s). \tag{4.68}$$

Summarizing, using (1.6) and (4.68) we obtain that in the range of times $\tau_\varepsilon \leq \tau$, and as long as $\delta(s) \ll 1$, we can approximate Φ as

$$\Phi \approx 4 + a_0(\tau_\varepsilon) - \frac{4(\delta(s))^2}{y^2} \quad \text{for } y \gg \delta(s), \tag{4.69}$$

where in (4.69) we will understand that only the larger of the terms $a_0(\tau_\varepsilon)$, $\frac{4(\delta(s))^2}{y^2}$ is meaningful. In a more precise manner (4.69) means $\Phi = 4 + a_0(\tau_\varepsilon) - \frac{4(\delta(s))^2}{y^2} + o(\max\{a_0(\tau_\varepsilon), \frac{(\delta(s))^2}{y^2}\})$, where by assumption y is bounded and satisfies $y \gg \delta(s)$. Due to the exponential growth of $\delta(s)$ (cf. (4.65)), the term $\frac{4(\delta(s))^2}{y^2}$, which initially is negligible if y is of order one, becomes the dominant one as s grows.

If $y \sim \delta(s)$, Φ might be approximated using (4.7), a formula that matches (4.69) in the intermediate region $\delta(s) \ll y \ll 1$. In particular, to the leading order Φ is approximately a steady state in the region $y \sim \delta(s)$.

Approximation (4.69) is valid as long as $e^{-\sqrt{2\tau_\varepsilon}} e^s$ remains small, because as soon as this quantity becomes of order one $\delta(s)$ becomes of order one (cf. (4.65)), and Φ cannot be approximated by means of (4.13) anymore. Let us define s_ε by means of $e^{-\sqrt{2\tau_\varepsilon}} e^{s_\varepsilon} = 1$. Formulae (4.65) and (4.69) provide an approximation of Φ as long as $s \ll s_\varepsilon$. In the original variables, such an approximation would be valid for $t < t_\varepsilon$, $T - t_\varepsilon = O(e^{-(\tau_\varepsilon + \sqrt{2\tau_\varepsilon})})$.

We finally proceed to compute $M(r, T_\varepsilon)$ in a way analogous to that in the case $\varepsilon = 0$. In the region $|x| \geq e^{-\frac{\tau_\varepsilon}{2}} \approx \sqrt{\varepsilon} \frac{e^{\sqrt{2\log(\frac{1}{\varepsilon})}}}{(\log(\frac{1}{\varepsilon}))^{\frac{1}{4}}}$, we can argue exactly as in section 2 since $a_0(\tau)$ remains almost constant for t between t_ε and T_ε. On the other hand, for $|x| \leq e^{-\frac{\tau_\varepsilon}{2}}$ we can use the fact that Φ is already close to a steady state (cf. (4.65)),

and the solution continues being close to a steady state for $t \le t_\varepsilon$. Actually, it is interesting to note that the steady state that has been reached is the one associated to the amount of mass $4 + a_0(\tau_\varepsilon)$ (cf. (4.3) and (4.69)). Indeed, the asymptotics (4.3) gives a width for the steady state $(\delta(s))^2 \sim 4K^2 e^{-(\gamma+2)} e^{-\sqrt{2\tau_\varepsilon}+s}$, or, in the original variables, $|x|^2 \sim 4K^2 e^{-(\gamma+2)} e^{-\sqrt{2\tau_\varepsilon}-\tau_\varepsilon}$, where $K \approx 1$. Taking into account the form in which the parameter ε appears in (2.5), it turns out that the value of the parameter λ (cf. (4.1)) for a steady state with width $(\delta(s))^2$ is

$$\lambda \sim \frac{8\varepsilon e^{\bar\tau}}{(\delta(s))^2} \sim \frac{6e^{-(\gamma+2)}\sqrt{2\tau_\varepsilon}}{\alpha K^2} e^{-\sqrt{2\tau_\varepsilon}} \sim \frac{6e^{-\gamma}}{\alpha K^2} e^{-\sqrt{2\log(\frac{1}{\varepsilon})}} \sqrt{2\log\left(\frac{1}{\varepsilon}\right)}. \tag{4.70}$$

The mass associated to this value of λ is (cf. (4.3))

$$\begin{aligned} M_\lambda(\infty) &\sim 4 + \frac{2\alpha}{3}\lambda \sim 4 + 4e^{-(\gamma+2)}\sqrt{2\tau_\varepsilon} e^{-\sqrt{2\tau_\varepsilon}} \\ &\sim 4 + 4e^{-\gamma} e^{-\sqrt{2\log(\frac{1}{\varepsilon})}} \sqrt{2\log\left(\frac{1}{\varepsilon}\right)}, \end{aligned} \tag{4.71}$$

where we have used the fact that $K \approx 1$, which coincides exactly with the asymptotics of M in the outer region $|y| \gg \delta$. In a more precise manner for times $(T_\varepsilon - t) \le e^{-\tau_\varepsilon - s_\varepsilon}$ in a region close to the origin, M becomes close to a steady state of (2.5) having the mass concentrated in a region of size $|x| \sim \sqrt{\varepsilon}\frac{e^{\sqrt{\frac{1}{2}\log(\frac{1}{\varepsilon})}}}{(\log(\frac{1}{\varepsilon}))^{\frac{1}{4}}} \equiv \chi(\varepsilon)$. Notice that $\chi(\varepsilon) \gg \sqrt{\varepsilon}$; therefore the size of this steady state is still much larger than the developed concentration region described in [22]. This is due to the fact that this concentration region has a mass that is still very close to 8π.

Finally we describe the evolution of $M(r,t)$ for $\varepsilon\frac{e^{\sqrt{2\log(\frac{1}{\varepsilon})}}}{(\log(\frac{1}{\varepsilon}))^{\frac{1}{2}}} \ll t - T_\varepsilon \ll 1$. Notice that in a region with size $|x| \approx O(\sqrt{\varepsilon}\frac{e^{\sqrt{\frac{1}{2}\log(\frac{1}{\varepsilon})}}}{(\log(\frac{1}{\varepsilon}))^{\frac{1}{4}}})$, solutions stabilize to equilibrium solutions in times of order $\varepsilon\frac{e^{\sqrt{2\log(\frac{1}{\varepsilon})}}}{(\log(\frac{1}{\varepsilon}))^{\frac{1}{2}}}$. For longer times we can then assume that the inner region has already stabilized to the value of the steady state associated to the value of the mass M in the corresponding outer region. Therefore, if $t > T_\varepsilon$, the inner region just follows the steady state having the mass given by the "outer behavior" computed in (2.46). Since in the outer region $G_\varepsilon(\frac{M_r}{r})$ can be approximated as $\frac{M_r}{r}$, it then follows that the corresponding solution can be approximated by the solution of (2.6). The inner region has a size of order $|x| \le \chi(\varepsilon)$ for times $t \le \bar t_\varepsilon$, where $\sqrt{|\log(\bar t_\varepsilon - T_\varepsilon)|}e^{-\sqrt{2|\log(\bar t_\varepsilon - T\varepsilon)|}} \sim e^{-\sqrt{2\log(\frac{1}{\varepsilon})}}\sqrt{\log(\frac{1}{\varepsilon})}$ or, in an equivalent manner, $\bar t_\varepsilon - T_\varepsilon \sim \varepsilon$. If $t \ge \bar t_\varepsilon$ the asymptotics (2.44) cannot be assumed to be constant anymore and the steady state that describes the inner layer begins to be modified. The width of the region occupied for the concentration regions can then be computed using (4.3) as well as (2.44). Such width turns out to be given by

$$\lambda \sim \frac{6}{\alpha} e^{-(\gamma+2)}\sqrt{2|\log(t - T_\varepsilon)|}e^{-\sqrt{2|\log(t-T_\varepsilon)|}} \quad \text{as } \varepsilon \ll t - T_\varepsilon \ll 1,$$

where λ is as in (4.1).

Notice that, as it could be expected, λ becomes of order one for $|t - T_\varepsilon|$ of order one.

5. Concluding remarks. This paper continues the analysis begun in [22]. The question considered in both papers is to study the effect of cutting off the chemotactic function for large values of cell concentration. For linear chemotactic functions it is known that the Keller–Segel model yields formation of Dirac masses in finite time. If the chemotactic function is assumed to saturate, solutions of the Keller–Segel system are global in time. It was seen in [22], using matched asymptotics, that solutions of this last system can exhibit some concentration regions that interact among themselves and with the surrounding cells. The dynamics of such regions has been studied in [22] in detail. This paper describes how the transition between the blowing up solutions of limit system (1.6), (1.7) and the type of solutions described in [22] takes place.

REFERENCES

[1] P. Biler, *Local and global solvability of some parabolic systems modelling chemotaxis*, Adv. Math. Sci. Appl., 8 (1998), pp. 715–743.

[2] M. P. Brenner, P. Constantin, L. P. Kadanoff, A. Schenkel, and S. C. Venkataramani, *Diffusion, attraction and collapse*, Nonlinearity, 12 (1999), pp. 1071–1098.

[3] M. P. Brenner, L. S. Levitov, and E. O. Budrene, *Physical mechanisms for chemotactic pattern formation by bacteria*, Biophys. J., 74 (1998), pp. 1677–1693.

[4] S. Childress, *Chemotactic collapse in two dimensions*, in Modelling of Patterns in Space and Time, Lecture Notes in Biomath. 55, Springer-Verlag, Berlin, 1984, pp. 61–66.

[5] S. Childress and J. K. Percus, *Nonlinear aspects of chemotaxis*, Math. Biosci., 56 (1981), pp. 217–237.

[6] H. Gajewski and K. Zacharias, *Global behaviour of a reaction-diffusion system modeling chemotaxis*, Math. Nachr., 195 (1998), pp. 77–114.

[7] M. A. Herrero, *Asymptotic properties of reaction-diffusion systems modeling chemotaxis*, Applied and Industrial Mathematics, Venice-2, 1998, R. Spigler, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[8] M. A. Herrero and J. J. L. Velázquez, *Singularity patterns in a chemotaxis model*, Math. Ann., 306 (1996), pp. 583–623.

[9] M. A. Herrero and J. J. L. Velázquez, *Chemotactic collapse for the Keller–Segel model*, J. Math. Biol., 35 (1996), pp. 177–196.

[10] M. A. Herrero and J. J. L. Velázquez, *A blow-up mechanism for a chemotaxis model*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 1739–1754.

[11] T. Hillen and K. Painter, *Global existence for a parabolic chemotaxis model with prevention of overcrowding*, Adv. in Appl. Math., 26 (2001), pp. 280–301.

[12] D. Horstmann, *The nonsymmetric case of the Keller–Segel model in chemotaxis*: *Some recent results*, NoDEA Nonlinear Differential Equations Appl., 8 (2001), pp. 399–423.

[13] D. Horstmann and G. Wang, *Blow-up in a chemotaxis model without symmetry assumptions*, European J. Appl. Math., 12 (2001), pp. 159–177.

[14] D. Horstmann, *On the existence of radially symmetric blow-up solutions for the Keller–Segel model*, J. Math. Biol., 44 (2002), pp. 463–478.

[15] W. Jäger and S. Luckhaus, *On explosions of solutions to a system of partial differential equations modelling chemotaxis*, Trans. Amer. Math. Soc., 329 (1992), pp. 819–824.

[16] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.

[17] T. Nagai, *Blow-up of radially symmetric solutions to a chemotaxis system*, Adv. Math. Sci. Appl., (1995), pp. 1–21.

[18] T. Nagai, T. Senba, and T. Suzuki, *Chemotactic collapse in a parabolic system of mathematical biology*, Hiroshima Math. J., 30 (2000), pp. 463–497.

[19] H. G. Othmer and A. Stevens, *Aggregation, blowup, and collapse: The ABC's of taxis in reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 1044–1081.

[20] M. Primicerio and B. Zaltmann, *Global in Time Solution to the Keller–Segel Model of Chemotaxis*, preprint.

[21] J. J. L. Velázquez, *Stability of some mechanisms of chemotactic aggregation*, SIAM J. Appl. Math., 62 (2002), pp. 1581–1633.

[22] J. J. L. Velázquez, *Point dynamics in a singular limit of the Keller–Segel model* 1*: Motion of the concentration regions*, SIAM J. Appl. Math., 64 (2004), pp. 1198–1223.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1249–1267

PARAMETER ESTIMATION OF THE HODGKIN–HUXLEY GATING MODEL: AN INVERSION PROCEDURE*

GUAN JUN WANG† AND JACQUES BEAUMONT†

Abstract. The Hodgkin–Huxley (HH) gating model has been extensively employed over the last half century to describe bioelectricity phenomena related to normal and impaired electrophysiological functions. Since the HH gating model is relatively empirical, the associated modelling methodology requires estimating model parameters (including functions of membrane voltage) from experimental data. Until now, as is the case for most nonlinear models, parameter estimation has been carried out through nonlinear least square fitting, which presents important limitations for the modelling methodology.

Here we pursue a different approach to the estimation problem, which allows us to overcome all the limitations inherent to nonlinear fitting. As initially introduced by Beaumont, Roberge, and Leon [*Math. Biosci.*, 115 (1993), pp. 65–101], instead of fitting we invert the solution. Specifically, model parameters (including functions of membrane voltage) are obtained from multiple transformations (or modals) applied to the solution, or equivalently, from an experimental data set. Such transformations enable one to deduce, for a given data point, disjoint ranges of parameter values which allow the model to exactly reproduce the solution. Using sufficiently large data sets and continuity criteria, it is possible to narrow down estimates to a specific value.

Our main results are (i) a more accurate estimation procedure; (ii) the ability to determine whether a data set sufficiently constrains the model, i.e., whether it is complete; (iii) if it is not, the possibility to identify a model family capable of reproducing the entire data set; and (iv) stimulation protocols which can produce complete data sets.

Key words. Hodgkin–Huxley model, nonlinear parameter estimation, membrane current kinetics, inverse problem

AMS subject classifications. 92C08, 92B08, 41A27, 65D04

DOI. 10.1137/S0036139902419826

1. Introduction.

The elucidation of dynamic changes in membrane conductance at the origin of cell excitability as well as its modelling by Hodgkin and Huxley [18], and the introduction of patch clamp techniques by Neher and Sackmann (see [15] for a detailed description), are among the most important discoveries of the last century in bioelectricity. In their classical study [18], Hodgkin and Huxley introduced a nonlinear model (the HH model) which describes current kinetics as a function of membrane channel states. While relatively empirical, the HH model has provided many insights into various bioelectric phenomena. In fact, both patch clamp techniques and gating models like the HH model are still the basic tools employed in investigations of bioelectric phenomena, e.g., the initiation and perpetuation of cardiac arrhythmias [4, 5, 29, 27, 28, 20, 16, 13, 10], the effect of electrical shocks on the evolution of cardiac arrhythmias [7, 33, 25, 21], neuromuscular control [12], coding of visual [19, 14, 31] as well as auditory signals [8, 9], cognitive functions like memory [1] and learning [24, 32, 23, 30], and finally brain diseases [22].

In practice, there are two modes for membrane current recording: one for analysis of channel populations, and another one for isolated channels. Here, we are concerned

*Received by the editors December 16, 2002; accepted for publication (in revised form) October 8, 2003; published electronically May 5, 2004. This work was supported by the Whitaker Foundation, NIH National Heart Lung and Blood Institute PO1-HL39707, and NIH National Center of Research Resources 1S10RR12917.

http://www.siam.org/journals/siap/64-4/41982.html

†Department of Pharmacology, Upstate Medical University of SUNY at Syracuse, Syracuse, NY 13210 (gjwang@sundance.pharm.upstate.edu, beaumont@sundance.pharm.upstate.edu).

only with the electrical activity of a channel population. To date, the HH gating model has been the most popular way to represent such activity. An alternative to this model is the deterministic Markovian model [26]. The Markovian model describes the biophysics of the gating process by treating membrane channels as entities exhibiting multiple voltage-dependent transitions between discrete states, and thus it is in better agreement with experimentation [26]. Also, several details of the current kinetics cannot be reproduced with an HH model [26]. However the link between changes in channel conformation and current kinetics is far from being obvious, and arguments based on current kinetics were not supported by an extensive mathematical analysis.

Because the HH model is relatively empirical, its parameters can only be estimated from experimental data. In mathematical terms we are facing an inverse problem, i.e, given only the solution of a nonlinear ODE (HH model), estimate its initial conditions and parameters. Until now, parameter estimation has followed the practice introduced by Hodgkin and Huxley [18], where parameters are obtained from currents recorded during voltage clamp stimulations (voltage clamp data). In such protocols, the membrane potential is held constant for an interval of time sufficiently long for the current flow to reach a steady state. Then time-dependent currents are recorded following the application of a test potential. Sets of currents are generated by varying the holding or test potentials. In experimental studies, it has become a common practice to approximate the parameters associated with the steady state functions of the model from currents obtained by varying the holding potential, and the time constants by fitting exponentials to the rising or decaying phase of the currents. A mathematical analysis has shown that this approach may provide a good approximation to several model parameters if specific conditions are respected [2]. If they are not, the error may be very large [2]. The weakness of this approach was also reported by Willms and colleagues [34, 35] (it is referred to as the disjoint method in their paper).

If appropriately performed, nonlinear least square fitting of voltage clamp data can produce good estimates, even in the presence of Gaussian noise [34, 35]. Nevertheless, nonlinear fitting has important drawbacks. The existence of local minima in the merit function to be optimized is particularly problematic. While the parameters are estimated with voltage clamp data, in most applications, the model is employed in other conditions to simulate the generation and transmission of electrical impulses in biological tissues. Although parameter values associated with various local minima may allow good reproduction of the voltage clamp data at the origin of the fit, they may produce different predictions when inserted in a more macroscopic model. This becomes a serious limitation when, for example, studying how channel mutations may render the myocardium more susceptible to the initiation of various arrhythmias. Considering two limitations of voltage clamp data sets, the range of inverse solutions can be large. First, voltage clamp data are relatively imprecise due to imperfect voltage control, space clamp, and seal between the mouth of the pipette and the cell membrane [15]. These errors are difficult to correct, and taking them into consideration by accepting a variation in the solution, even if small, may lead to large variations in the model parameters. Second, we do not know whether a voltage clamp data set produced by current practice constrains the HH gating model sufficiently. In fact, from the theory developed here it will become evident that it does not in most cases.

Following earlier studies [2, 3], we present here an inversion procedure which overcomes all the difficulties inherent to nonlinear fitting. The estimation problem requires evaluation of one constant and, for each state variable included in the model,

two functions of membrane voltage, namely the steady state and the time constant. We introduce transformations which, when applied to voltage clamp data sets obtained by keeping the test potential constant, allow estimation of steady states. Once these functions are evaluated, we bound the inverse solution by applying other transformations to complementary data sets. Interestingly, the bounds delineate disjoint ranges of valid parameter values. For any point taken within these bounds, we can associate to it a time constant. Due to the existence of disjoint ranges, inversion of the time constants produces a structure with multiple branches. When the data set is complete, it is possible to deduce a function from such a structure. Important outcomes of the study are the ability to determine whether a voltage clamp data set is complete or not, and to identify a set of protocols which can produce a complete data set. Both elements are, we believe, important for experimental design and for the development of a systematic modelling methodology.

The paper is organized as follows. For the sake of completeness we introduce the model and various definitions in section 2. In section 3 we introduce transformations that allow estimation of the steady state functions. Finally, transformations to bound the inverse solution and to estimate the time constants are derived in section 4.

2. Preliminaries. In the HH gating model, a given channel is composed of several molecular components (gates). These change state under the influence of the electrostatic potential existing across the cell membrane (V). Each molecular component has two states: closed (C) or open (O). The channel is open when all its molecular components are in the open state. Otherwise it is closed. Such behavior is represented by the following state diagram:

$$C_i \underset{\beta_i(V)}{\overset{\alpha_i(V)}{\rightleftharpoons}} O_i, \tag{2.1}$$

where V is the membrane potential and $\alpha(V)$ and $\beta(V)$ are, respectively, the forward and backward rates of transition between the open (O) and closed (C) states of the molecular gate i. The law of mass action governs the dynamical changes between the two states. Precisely, the variable y represents the fraction of a population of gates in the open state, the kinetics of which obeys

$$\frac{dy_i(V,t)}{dt} = \alpha_i(V)\,(1 - y_i(V,t)) - \beta_i(V) y_i(V,t). \tag{2.2}$$

We refer to the variables y_i as the gating variables. When in the open state, the channel is seen as a resistive barrier to the passage of ions and has a fixed maximal conductance denoted by $\bar{g}$. We assume $\bar{g}$ constant, but, as will be clear later, the method applies as well when $\bar{g}$ is a function of voltage, as long as this function is known. The driving force for the passage of ions across the channel is the electrochemical gradient, which is given by $V - e_k$, where e_k is the Nernst potential related to an ion species "k" [11, Chap. 2]. The membrane current for an ion species "k," denoted by $I_k(V,t)$, is

$$I_k(V,t) = \bar{g}_k \left(\prod_{n=0}^{\bar{n}(k)} y_{k,n}(V,t)^{\lambda_{k,n}} \right) (V - e_k), \tag{2.3}$$

where $\bar{n}(k)$ is the number of distinct molecular gates of a channel, and $\lambda_{k,n}$ the number of similar components of a molecular gate n in a channel k.

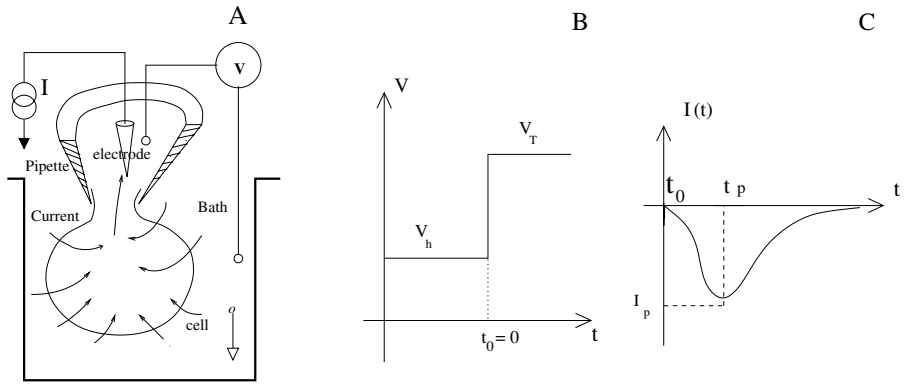


FIG. 2.1. A: *Setup for voltage clamp experiment.* B: *Stimulation protocol.* C: *Typical current wave shape.*

Data at the basis of the parameter estimation are membrane currents recorded in isolated cells during voltage clamp stimulation. The experimental setup is illustrated in Figure 2.1 and described in detail in [15]. Briefly, cells are isolated, then a pipette is apposed against the membrane surface, and suction is applied to break the membrane under the pipette tip. The interior of the pipette is then in continuum with the intracellular space. An electronic feedback circuit allows one to simultaneously impose potential and record current on the electrode placed inside the pipette. The voltage clamp stimulation (Figure 2.1B) consists of clamping the membrane potential at a holding potential V_H for an interval of time sufficiently long for ionic fluxes to reach steady state. Then at t_0 a step to a test potential V_T is applied. The membrane current is recorded just after the application of the step. In the following we assume that only one type of membrane channel produces the current. Thus we use only one subscript to refer to the gating variables. In these conditions, the time course of a gating variable y_i is given by

$$y_i(t; V_H, V_T) = s_i(V_T) + (s_i(V_H) - s_i(V_T))\, e^{-t/\tau_i(V_T)}, \tag{2.4}$$

where $s_i(V)$ and $\tau_i(V)$ are, respectively, the steady state and time constant associated to the gating variables y_i. They are related to the forward and backward rates of transition by

$$s(V) = \frac{\alpha(V)}{\alpha(V) + \beta(V)} \quad \text{and} \quad \tau(V) = \frac{1}{\alpha(V) + \beta(V)}. \tag{2.5}$$

Because $s(V)$ and $\tau(V)$ are parameters more directly measurable, the format

$$\frac{dy(V,t)}{dt} = \frac{s(V) - y(V,t)}{\tau(V)} \tag{2.6}$$

for the expression governing the dynamical changes in y_i is preferred. It is an experimental fact that, for the great majority of membrane channels, $s_i(V)$ are monotonic functions exhibiting sigmoidal shape. They are commonly parameterized by

$$s(V) \approx \frac{1}{1 + e^{(V - V_0)/s_f}}, \tag{2.7}$$

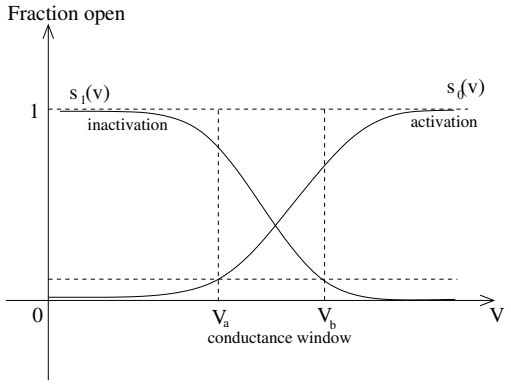


FIG. 2.2. *Steady state functions $s_i(V)$ of a two-component gating model.*

which for reasons unknown to these authors is called a Boltzmann function. Although there exist some arguments justifying parameterizing $s_i(V)$ with the above function [6], [17, Chap. 2], here we assume only that $s_i(V)$ is monotonic.

When $s_i(V)$ is monotonically increasing/decreasing with the membrane potential, we refer to the related gating variable (y_i) as an activation/inactivation gate (Figure 2.2). In this paper we deal with gating models incorporating only two gates, one activation and one inactivation gate, which is a situation frequently encountered in bioelectricity. We term the voltage range where $\prod_{n=0}^{\bar{n}} s_n^{\lambda_n}(V) > 1\%$ the conductance window. To simplify the writing, we denote $s_i(V_a) \equiv s_a^i$ and similarly for the time constants $\tau_i(V_a) \equiv \tau_a^i$, where V_a refers to a specific membrane voltage. For an HH model with one activation (y_0) and one inactivation gate (y_1), the current waveform during a voltage clamp stimulation has at most two extrema. The time coordinates of the peak currents are given by the zeros of

$$\delta(t) = \frac{\tau_T^0 \lambda_1}{\tau_T^1 \lambda_0} \frac{s_T^0}{(s_T^0 - s_H^0)} e^{t/\tau_T^0} + \frac{s_T^1}{(s_T^1 - s_H^1)} e^{t/\tau_T^1} - \left(1 + \frac{\tau_T^0 \lambda_1}{\tau_T^1 \lambda_0}\right), \tag{2.8}$$

which has a critical point at

$$t_c = \frac{\tau_T^0 \tau_T^1}{\tau_T^0 - \tau_T^1} \ln\left[\frac{\lambda_0 s_T^1 (s_T^0 - s_H^0)}{\lambda_1 s_T^0 (s_H^1 - s_T^1)}\right]. \tag{2.9}$$

At the critical point,

$$\left.\frac{d^2\delta(t)}{dt^2}\right|_{t=t_c} = \left(\frac{1}{\tau_T^0} - \frac{1}{\tau_T^1}\right) \left[\frac{\lambda_0 (s_T^0 - s_H^0)}{s_T^0}\right]^{\frac{\tau_T^0}{(\tau_T^1 - \tau_T^0)}} \left[\frac{s_T^1}{\lambda_1 (s_H^1 - s_T^1)}\right]^{\frac{\tau_T^1}{(\tau_T^0 - \tau_T^1)}}. \tag{2.10}$$

The reader should consult [2] for more details about (2.8)–(2.10).

Below, we denote time derivatives with a dot above the differentiated variables. For an arbitrary couple V_H, V_T, we denote the membrane current by $I(t)$, $J(t) \equiv \dot{I}(t)/I(t)$, and $I_N(t; t_r) \equiv I(t)/I(t_r)$, where t is a continuous variable and t_r a given reference time within the recording interval $[t_a, t_b]$. Sometimes it is necessary to specify the potentials of the voltage clamp stimulation associated with a current. In this case we denote the current by $I(t; V_H, V_T)$, where the potentials on the right of the semicolon are in the order in which they are applied in time during stimulation. A similar notation applies to any function derived from such a current, e.g.,

$J(t) \rightarrow J(t; V_H, V_T)$, $I_N(t; t_r) \rightarrow I_N(t, t_r; V_H, V_T)$. Transformations derived below are applied to sets of functions (or currents) which share similar properties. Examples of such sets are $I(t)$ acquired through repetitive application of voltage clamp steps where only one potential of the step is varied at each stimulation. We denote sets by indicating the voltage partition associated with the stimulation sequence in the arguments. For example, we denote by $I(t; [V_{H0} \cdots V_{Hn}], V_T)$, $[V_{H0} \cdots V_{Hn}]$, a partition of $[V_{H0}, V_{Hn}]$, and similarly when V_T is varied. Sets of functions derived from a set of currents are denoted analogously, e.g., $J(t; [V_{H0} \cdots V_{Hn}], V_T)$, $I_N(t, t_r; [V_{H0} \cdots V_{Hn}], V_T)$. We denote by $\Omega_{t,V}$ the space $t \in [t_a, t_b]$, $V \in [V_{H0}, V_{Hn}]$ generated during the acquisition of $I(t; [V_{H0} \cdots V_{Hn}], V_T)$ or $I(t; V_H, [V_{T0 \ldots} V_{Tn}])$. The context always makes clear what is intended as the V axis of $\Omega_{t,V}$. Finally, since λ_i are integers, and based on experimental observations they are small (i.e., $\lambda_i < 6$), in our estimation procedure we assume these constants to be known.

3. Estimation of the steady states, $s_i(V)$. Since the expression of any $I(t; V, V_T) \in I(t; [V_{H0} \cdots V_{Hn}], V_T)$, $V \in [V_{H0} \cdots V_{Hn}]$, differs only in s_H^i (see (2.4)), the data set $I(t; [V_{H0} \cdots V_{Hn}], V_T)$ poses considerable constraints on the model parameters. Here we exploit this property to evaluate $s_i(V)$.

THEOREM 3.1. *For any $I(t)$ and associated $J(t)$, $t \in [t_a, t_b]$, obtained from a single step stimulation,*

$$\frac{s_H^i}{s_T^i} = 1 - \frac{J(t_r) + \lambda_{\bar{\imath}}/\tau_T^{\bar{\imath}}(1 - \varepsilon_{\bar{\imath}}(t_r))}{J(t_r) + \lambda_i/\tau_T^i + \lambda_{\bar{\imath}}/\tau_T^{\bar{\imath}}(1 - \varepsilon_{\bar{\imath}}(t_r))}\, e^{t_r/\tau_T^i}, \qquad t_r \in [t_a, t_b],\ i \in [0, 1],\ \bar{\imath} = 1 - i, \tag{3.1}$$

where

$$\varepsilon_j(t_r) = \frac{s_T^j}{s_T^j + (s_H^j - s_T^j)e^{-t_r/\tau_T^j}}, \qquad t_r \in [t_a, t_b],\ j \in [0, 1]. \tag{3.2}$$

Proof. Take the derivative with respect to time of (2.3), divide the resulting expression by (2.3), and replace in the latter $dy_i(V,t)/dt$ by (2.6). It follows that

$$J(t) = \frac{\lambda_0}{y_0}\frac{dy_0}{dt} + \frac{\lambda_1}{y_1}\frac{dy_1}{dt} = -\frac{\lambda_0}{\tau_T^0}(1 - \varepsilon_0(t)) - \frac{\lambda_1}{\tau_T^1}(1 - \varepsilon_1(t)). \tag{3.3}$$

Solve $\varepsilon_i(t)$ for s_H^i/s_T^i; (3.1) follows immediately. ∎

Interesting simplifications occur when the following holds.

CONDITION 3.2. $\mathrm{Max}\{\varepsilon_1(t; V, V_T)\} \leq \varepsilon_{th} \ll 1$ *over a set* $I(t; V, V_T)$, $t, V \in \Omega_{t,V}$.

Here dependence on $s_1(V)$ is eliminated in (3.3). The counterpart to this condition is the following.

CONDITION 3.3. $\mathrm{Max}\{\varepsilon_0(t; V, V_T)\} \leq \varepsilon_{th} \ll 1$ *over a set* $I(t; V, V_T)$ $t, V \in \Omega_{t,V}$.

Fortunately, stimulation sequences producing currents satisfying these conditions exist.

PROPOSITION 3.4. *Condition* 3.2 *is satisfied if*

$$\left[\frac{I(t_a; V, V_T)}{I(t_{max}; V, V_T) + I(t_b; V, V_T)}\right]^{1/\lambda_1} < \epsilon_{th}, \qquad t_{max}, V \in \Omega_{t,V},\ V < V_T, \tag{3.4}$$

and 3.3 *if*

$$\left[\frac{I(t_a; V, V_T)}{I(t_{max}; V, V_T) + I(t_b; V, V_T)}\right]^{1/\lambda_0} < \epsilon_{th}, \qquad t_{max}, V \in \Omega_{t,V},\ V > V_T, \tag{3.5}$$

where $t \in [t_a, t_b]$, $t_a < t_{max} < t_b$, $I(t_{max}) = \mathrm{Max}\{I(t)\}$, $I_{t_a} > I_{t_b}$.

Proof. If in a single step stimulation sequence the currents are acquired in the space $\Omega_{t,V}$, which also satisfies $V_H < V_T$, then $d\epsilon_0(t)/dt < 0$, $d\epsilon_1(t)/dt > 0$, $\epsilon_0(t) \in (1, \infty)$, $\epsilon_1(t) \in (0, 1)$. Remarking that $\epsilon_j(t) = s_T^j / y_j(t)$, then

$$\frac{I(t_a)}{I(t_{max}) + I(t_b)} = \frac{\epsilon_0^{\lambda_0}(t_{max})\epsilon_1^{\lambda_1}(t_{max}) + \epsilon_0^{\lambda_0}(t_b)\epsilon_1^{\lambda_1}(t_b)}{\epsilon_0^{\lambda_0}(t_a)\epsilon_1^{\lambda_1}(t_a)}. \tag{3.6}$$

Using the monotonicity of $\epsilon_j(t)$, the fact that $\epsilon_1(t)$ is bounded, and that the left member of 3.6 as well as $\epsilon_j^{\lambda_j}(t)$ are positive functions,

$$\begin{aligned} \frac{I(t_a)}{I(t_{max}) + I(t_b)} &> \frac{\epsilon_0^{\lambda_0}(t_{max})}{\epsilon_0^{\lambda_0}(t_a)} + \frac{\epsilon_0^{\lambda_0}(t_b)\epsilon_1^{\lambda_1}(t_b)}{\epsilon_0^{\lambda_0}(t_a)\epsilon_1^{\lambda_1}(t_a)} \\ &> \epsilon_1^{\lambda_1}(t_a)\left[\frac{\epsilon_0^{\lambda_0}(t_{max})}{\epsilon_0^{\lambda_0}(t_a)}\frac{\epsilon_0^{\lambda_0}(t_a)}{\epsilon_0^{\lambda_0}(t_b)} + \frac{\epsilon_1^{\lambda_1}(t_b)}{\epsilon_1^{\lambda_1}(t_a)}\right] > \epsilon_1^{\lambda_1}(t_b) \\ &= [\mathrm{Max}\{\epsilon_1(t)\}]^{\lambda_1}, \quad t \in \{t_a, t_b\}. \end{aligned} \tag{3.7}$$

The same applies to Condition 3.3 when currents are acquired in a space $\Omega_{t,V}$ satisfying $V > V_T$. The proof easily follows from the above and is left to the reader. ∎

Interestingly, due to the structure of the HH gating model, most of the relations developed for the estimation of $s_i(V)$ are symmetric with respect to Conditions 3.2 and 3.3.

DEFINITION 3.5 (symmetric expression). *Take two arbitrary sets of currents* $\mathcal{I}_1$ *and* $\mathcal{I}_2$ *respectively satisfying some conditions* $\mathcal{C}_1$ *and* $\mathcal{C}_2$. *Assume expressions* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *are obtained from the application of the same transformations respectively to* $\mathcal{I}_1$ *and* $\mathcal{I}_2$. *We say that* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *have a symmetric structure or simply are symmetric with respect to the conditions* $\mathcal{C}_1$ *and* $\mathcal{C}_2$ *if* $\mathcal{E}_1$ *can be obtained from* $\mathcal{E}_2$, *and vice versa, simply by interchanging the indices* 0 *and* 1 *of each parameter and function.*

Obviously (3.1) is symmetric with respect to Conditions 3.2 and 3.3.

We now deduce modals which allow us to evaluate τ_T^i in (3.1).

LEMMA 3.6. *Consider a set* $I(t; V, V_T)$, $t, V \in \Omega_{t,V}$, *which satisfies Condition* 3.2. *Then for any* $J(t, V, V_T)$ *and* $I_N(t, t_r; V, V_T)$, $t, V \in \Omega_{t,V}$, *deduced from this set,*

$$J(t; t_r; V, V_T) = \frac{\frac{\lambda_0}{\tau_T^0}\left(J(t_r; V, V_T) + \frac{\lambda_1}{\tau_T^1}\right)e^{-(t-t_r)/\tau_T^0}}{\left(J(t_r; V, V_T) + \frac{\lambda_0}{\tau_T^0} + \frac{\lambda_1}{\tau_T^1}\right) - \left(J(t_r; V, V_T) + \frac{\lambda_1}{\tau_T^1}\right)e^{\frac{-(t-t_r)}{\tau_T^0}}} - \frac{\lambda_1}{\tau_T^1}, \tag{3.8}$$

$$\begin{aligned} I_N(t; t_r; V, V_T)^{\frac{1}{\lambda_0}} = \frac{\tau_T^0}{\lambda_0}\Bigg[&\left(J(t_r; V, V_T) + \frac{\lambda_0}{\tau_T^0} + \frac{\lambda_1}{\tau_T^1}\right) \\ &- \left(J(t_r; V, V_T) + \frac{\lambda_1}{\tau_T^1}\right)e^{\frac{-(t-t_r)}{\tau_T^0}}\Bigg]\, e^{\frac{-\lambda_1(t-t_r)}{\lambda_0\tau_T^1}}. \end{aligned} \tag{3.9}$$

Furthermore, both (3.8) *and* (3.9) *are symmetric with respect to Conditions* 3.2 *and* 3.3.

Proof. Equation (3.8) follows from substituting (3.1) into (3.3). To obtain (3.9), integrate (3.8) in the interval $[t_r, t]$, take the exponential of the resulting expression, and elevate each member to the power $1/\lambda_0$. Symmetry of (3.8) and (3.9) with respect to Conditions 3.2 and 3.3 follows from symmetry of (3.1) to the same conditions. ∎

Expressions of Lemma 3.6 provide two independent relations dependent on two unknowns: τ_T^0 and τ_T^1. Thus we search for combinations of (3.8) and (3.9) which allow one to factor these unknowns.

THEOREM 3.7. *For any $I_N(t;t_r;V,V_T)$ and $J(t_r;V,V_T)$ deduced from $I(t;V,V_T)$, $t,V\in\Omega_{t,V}$, $t_r\in[t_a,t_b]$, which satisfies Condition* 3.2*,*

$$I_N^{\frac{1}{\lambda_0}}(t,t_r;V,V_T)-1=\theta_0\left(-\lambda_0\frac{dI_N(t,t_r;V,V_T)}{dt}+J(t_r;V,V_T)\right)-\theta_1\int_{t_r}^{t}I_N(t,t_r;V,V_T)dt, \tag{3.10}$$

with

$$\theta_0=\frac{\tau_T^0\tau_T^1}{(2\lambda_1\tau_T^0+\lambda_0\tau_T^1)}\quad\text{and}\quad\theta_1=\frac{\lambda_1}{\lambda_0\tau_T^1}\left(\frac{\lambda_1\tau_T^0+\lambda_0\tau_T^1}{2\lambda_1\tau_T^0+\lambda_0\tau_T^1}\right), \tag{3.11}$$

from (3.11) *we have*

$$\tau_T^0=\frac{\theta_0\lambda_0}{(1-4\lambda_0\theta_0\theta_1)^{1/2}},\qquad\tau_T^1=\frac{2\lambda_1\theta_0}{1-(1-4\lambda_0\theta_0\theta_1)^{1/2}}, \tag{3.12}$$

where (3.10)–(3.12) *are symmetric with respect to Conditions* 3.2 *and* 3.3.

Proof. Consider a set $I(t,V,V_T)$, $t,V\in\Omega_{t,V}$, satisfying Condition 3.2. Solve (3.9) for $e^{-(1/\tau_T^0+\lambda_1/(\lambda_0\tau_T^1))(t-t_r)}$. Then substitute the resulting expression into the derivative of (3.9). This leads to

$$\frac{dI_N(t,t_r;V,V_T)^{\frac{1}{\lambda_0}}}{dt}=\frac{\left(1+\frac{\tau_T^0\lambda_1}{\tau_T^1\lambda_0}\right)}{\tau_T^0}\left[\frac{\left(\frac{\tau_T^0}{\lambda_0}J(t_r)+1+\frac{\tau_T^0\lambda_1}{\tau_T^1\lambda_0}\right)}{\left(1+\frac{\tau_T^0\lambda_1}{\tau_T^1\lambda_0}\right)}e^{\frac{-\lambda_1}{\lambda_0\tau_T^1}(t-t_r)}-I_N(t,t_r;V,V_T)^{\frac{1}{\lambda_0}}\right]. \tag{3.13}$$

From the integration of (3.13) in the interval $[t_r,t]$ we have

$$I_N(t,t_r;V,V_T)^{\frac{1}{\lambda_0}}-1=-\frac{\left(\frac{\tau_T^0}{\lambda_0}J(t_r;V,V_T)+1+\frac{\tau_T^0\lambda_1}{\tau_T^1\lambda_0}\right)}{\left(\frac{\tau_T^0\lambda_1}{\tau_T^1\lambda_0}\right)}\left(e^{\frac{\lambda_1}{\lambda_0\tau_T^1}(t-t_r)}-1\right)$$
$$-\frac{\left(1+\frac{\tau_T^0\lambda_1}{\tau_T^1\lambda_0}\right)}{\tau_T^0}\int_{t_r}^{t}I_N(t,t_r;V,V_T)^{1/\lambda_0}dt. \tag{3.14}$$

Solve for $e^{(-\lambda_1/(\lambda_0\tau_T^1))(t-t_r)}$ using (3.13) and substitute the resulting expression into (3.14); formula (3.10) follows. Finally, symmetry of (3.13) and (3.14) with respect to Conditions 3.2 and 3.3 follows from the symmetry of (3.9) with respect to the same conditions. ∎

In practice the above procedure does not provide satisfying results when the conductance window is very narrow, a situation that is commonly encountered in bioelectricity (e.g., sodium current of many tissues from human and most animal species). In such an instance, the range of holding potentials for which it is possible to record a current and to satisfy Conditions 3.2 or 3.3 allows one to estimate only

the very foot of $s_i(V)$. To overcome this problem, we further develop a procedure introduced by Beaumont, Roberge, and Lemieux [3], which is based on an analytical expression for the normalized currents and their derivative.

THEOREM 3.8. *Consider a set $I(t;V,V_T)$, satisfying Condition* 3.2*, and $J(t;V,V_T)$, $t,V\in\Omega_{t,V}$, deduced from this set. Then*

$$\frac{I(t;V,V_T)}{I(t;V_R,V_T)}=\left[\frac{s_1(V)}{s_R^1}\right]^{\lambda_1}\left[\frac{J(t;V_R,V_T)+\frac{\lambda_0}{\tau_T^0}+\frac{\lambda_1}{\tau_T^1}}{J(t;V,V_T)+\frac{\lambda_0}{\tau_T^0}+\frac{\lambda_1}{\tau_T^1}}\right]^{\lambda_0},\ V_R\in[V_{H0}\cdots V_{Hn}], \tag{3.15}$$

where (3.15) *is symmetric with respect to Conditions* 3.2 *and* 3.3.

Proof. Consider any $I(t;V_H,V_T)\in I(t;V,V_T)$, $t,V\in\Omega_{t,V}$, satisfying Condition 3.2. Solve (3.3) for e^{-t/τ_i} and substitute the resulting expression into (2.4). Then at $t_r\in[t_a,t_b]$,

$$y_0(t_r;V,V_T)=s_T^0\frac{\frac{\lambda_0}{\tau_T^0}}{J(t_r;V,V_T)+\frac{\lambda_0}{\tau_T^0}+\frac{\lambda_1}{\tau_T^1}}, \tag{3.16}$$

$$y_1(t_r;V,V_T)=s_H^1\left[\frac{J(t_r;V,V_T)+\frac{\lambda_1}{\tau_T^1}}{J(t_r;V,V_T)+\frac{\lambda_0}{\tau_T^0}+\frac{\lambda_1}{\tau_T^1}}\,\frac{s_T^0}{s_T^0-s_H^0}\right]^{\tau_T^0/\tau_T^1}. \tag{3.17}$$

Substitute (3.16) and (3.17) into (2.3). Evaluate the resulting expression at two different potentials, and divide one of them by the other; then (3.15) follows. Symmetry of (3.15) with respect to Conditions 3.2 and 3.3 follows from symmetry of (3.16) and (3.17), which in turn follows from symmetry of (3.3) with respect to the same conditions. ∎

Theorem 3.8 is complementary to Theorem 3.1 in the sense that it provides, for the same set, information about $s_{\bar{\imath}}(V)$, $\bar{\imath}=1-i$. Fortunately, the conditions of application for each theorem are the same.

There remains another problem with the practical application of Theorems 3.1 and 3.8. Under restrictions imposed by Conditions 3.2 and 3.3, some models may not produce currents detectable by available instrumentation. This occurs with Condition 3.3 in a step simulation when $\tau_T^0<\tau_T^1$, $V_T<V_a$, $V_H>V_b$. In other words, during a step stimulation, the gate which is closing reacts more rapidly to changes in membrane potential than the one which is opening. This is the case for the sodium channel of cardiac cells of most animal species. To remedy this problem, we propose modifying the stimulation protocol by inserting a conditioning pulse prior to the application of the test pulse. The protocol is illustrated in Figure 3.1.

In this double step stimulation we consider ν to be an independent variable, and ν_r an arbitrary value along the ν axis. To be consistent with the notation introduced here, we denote an arbitrary current by $I(t;\nu_r)$, a current produced with specific potentials by $I(t;\nu_r;V_H,V_C,V_T)$, and sets of currents produced by varying either V_H or ν in a stimulation sequence, respectively, by $I(t;\nu_r;[V_{H0}\cdots V_{Hn}],V_C,V_T)$ and $I(t;[\nu_0\cdots\nu_n];V_H,V_C,V_T)$, where $[\nu_0\cdots\nu_n]$ is a partition of $[\nu_0,\nu_n]$, $\nu_n>\nu_0$. In this last case the acquisition space is denoted by $\Omega_{t,\nu}$. As with the single step stimulation, a similar notation applies to functions or sets deduced from these sets. Finally, we still denote the recording interval during the application of the test pulse by $[t_a,t_b]$.

Although the addition of a conditioning pulse complicates the analytical expression for the membrane current, it is still possible to express $s_i(V)$ as a function of various modals of the current.

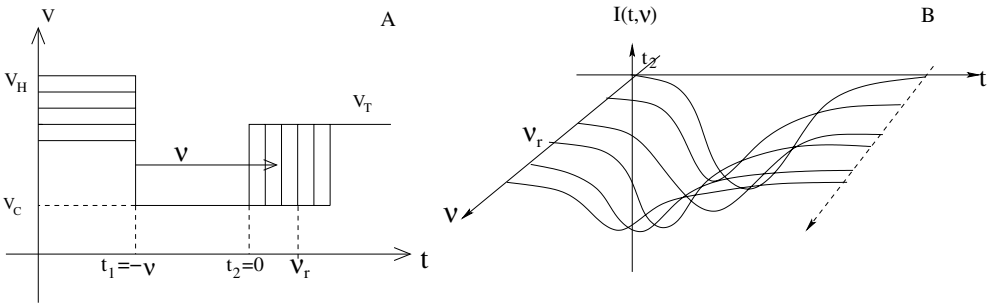


FIG. 3.1. A: *Voltage clamp protocol, which includes the insertion of a conditioning pulse (potential V_C and duration ν) prior to the application of a test pulse.* B: *Typical current traces.*

THEOREM 3.9. *For any $I(t,\nu_r) \in I(t;\nu_r,V,V_C,V_T)$ and $J(t,\nu_r) \in J(t;\nu_r,V,V_C,V_T)$, $t,V \in \Omega_{t,V}$, deduced from the same set,*

$$\frac{s_H^i}{s_T^i} = \frac{s_C^i}{s_T^i} - \left(\frac{s_C^i}{s_T^i} - 1\right) e^{\frac{\nu_r}{\tau_c^i}} - \frac{\left(J(t_r;\nu_r) + \frac{\lambda_{\bar\imath}}{\tau^{\bar\imath}}(1-\psi_{\bar\imath}(t_r;\nu_r))\right)}{\left(J(t_r;\nu_r) + \frac{\lambda_i}{\tau_T^i} + \frac{\lambda_{\bar\imath}}{\tau_T^{\bar\imath}}(1-\psi_{\bar\imath}(t_r;\nu_r))\right)} e^{\left(\frac{t_r}{\tau_T^i}+\frac{\nu_r}{\tau_C^i}\right)},$$

$$t_r \in [t_a,t_b],\ i \in [0,1],\ \bar\imath = i-1, \tag{3.18}$$

with

$$\psi_j(t_r;\nu_r) = \frac{s_T^j}{s_T^j + (s_C^j - s_T^j)e^{\frac{-t_r}{\tau_T^j}} + (s_H^j - s_C^j)e^{-\left(\frac{t_r}{\tau_T^j}+\frac{\nu_r}{\tau_c^j}\right)}}, \qquad t_r \in [t_a,t_b]. \tag{3.19}$$

Proof. The time course of $y_i(t)$, $t \in [t_a,t_b]$, in a double step voltage clamp stimulation is given by

$$y_i(t,\nu_r) = s_T^i + (s_C^i - s_T^i)e^{\frac{-t}{\tau_T^i}} + (s_H^i - s_C^i)e^{-\left(\frac{t}{\tau_T^i}+\frac{\nu_r}{\tau_C^i}\right)}. \tag{3.20}$$

Inserting (3.20) and its derivative into (3.3), we get

$$J(t,\nu_r) = -\frac{\lambda_0}{\tau_T^0}(1-\psi_0(t;\nu_r)) - \frac{\lambda_1}{\tau_T^1}(1-\psi_1(t;\nu_r)). \tag{3.21}$$

Solving $\psi_i(t;\nu_r)$ in (3.21) for $(s_H^i/s_T^i - s_C^i/s_T^i)$, (3.18) follows immediately. ∎

As with the single step stimulation, interesting simplifications occur when the following holds.

CONDITION 3.10. $\mathrm{Max}\{\psi_1(t;\nu;V,V_C,V_T)\} \le \psi_{th} \ll 1$ *over a set* $I(t;\nu;V,V_C,V_T)$, $t,V \in \Omega_{t,V}$ *or* $t,\nu \in \Omega_{t,\nu}$.

Here $s_1(V)$ is eliminated in (3.18). The counterpart to this condition is the following.

CONDITION 3.11. $\mathrm{Max}\{\psi_0(t;\nu;V,V_C,V_T)\} \le \psi_{th} \ll 1$ *over a set* $I(t;\nu;V,V_C,V_T)$, $t,V \in \Omega_{t,V}$ *or* $t,\nu \in \Omega_{t,\nu}$.

Fortunately, double step stimulation sequences producing currents satisfying these conditions exist.

PROPOSITION 3.12. *Condition* 3.10 *is satisfied in* $t,V \in \Omega_{t,V}$ *or* $t,\nu \in \Omega_{t,\nu}$ *when*

$$\left[\frac{I(t_a;\nu;V,V_C,V_T)}{I(t_{max};\nu;V,V_C,V_T) + I(t_b;\nu;V,V_C,V_T)}\right]^{1/\lambda_1} < \psi_{th} \quad \text{if } V,V_C \le V_T, \tag{3.22}$$

and Condition 3.11 *is satisfied when*

$$\left[\frac{I(t_a;\nu;V,V_C,V_T)}{I(t_{max};\nu;V,V_C,V_T)+I(t_b;\nu;V,V_C,V_T)}\right]^{1/\lambda_0} < \psi_{th} \quad \text{if } V,V_C \geq V_T. \tag{3.23}$$

Proof. The proof follows from the proof of the existence of Conditions 3.2 and 3.3. If in a double step stimulation sequence currents are acquired in the space $\Omega_{t,V}$ or $\Omega_{t,\nu}$ which also satisfy $V,V_C \leq V_T$, then $d\psi_0(t)/dt \leq 0$, $d\psi_1(t)/dt \geq 0$, $\psi_0(t) \in (\infty,1)$, $\psi_1(t) \in (0,1)$. Remarking that $\psi_j(t) = s_T^j/y_j(t)$, it becomes obvious that inequalities derived in the proof of the existence of Conditions 3.2 and 3.3 apply as well here, when replacing $\epsilon_j(t)$ by $\psi_j(t)$.

Similarly a condition exists, (2.3), for currents acquired in the space $\Omega_{t,V}$ or $\Omega_{t,\nu}$, if for each current $V,V_C \geq V_T$. □

Finally, relation (3.18) is symmetric with respect to Conditions 3.10 and 3.11, a property that follows from symmetry of (3.21) with respect to the same conditions.

The ability to extract information for estimation of $s_i(V)$ from currents acquired in double step stimulations allows experimentalists to design experiments where acquisition of bioelectric signals is more amenable to currently available instrumentation. However, since utilization of Theorem 3.9 requires knowing s_C^i/s_T^i, V_C should be set to a value where it is possible to record currents satisfying either Condition 3.2 or 3.3 during single step stimulation. For some models such a condition may be impossible to meet. To circumvent this problem, we look for new stimulation protocols which can provide more freedom as to the values at which V_C can be set. First, we remark that the expression of several modals of membrane currents acquired with double step stimulation can be simplified when $V_H = V_T$. Specifically, we claim the following.

LEMMA 3.13. *Take a set* $I(t;\nu;V_T,V_C,V_T)$ *satisfying Condition* 3.10. *Then for any* $J(t;\nu_r;V_T,V_C,V_T) \in J(t;\nu;V_T,V_C,V_T)$, $t,\nu \in \Omega_{t,\nu}$,

$$\frac{s_C^0}{s_T^0} = 1 - \left(\frac{J(t_r,\nu_r)+\frac{\lambda_1}{\tau_T^1}}{J(t_r,\nu_r)+\frac{\lambda_0}{\tau_T^0}+\frac{\lambda_1}{\tau_T^1}}\right)\left(\frac{e^{\frac{t_r}{\tau_T^0}}}{1-e^{\frac{-\nu_r}{\tau_c^0}}}\right), \tag{3.24}$$

$$J(t,\nu) = \frac{\left(\frac{\lambda_0}{\tau_T^0}\right)\left(J(t_r,\nu_r)+\frac{\lambda_1}{\tau_T^1}\right)h(\nu;\nu_r)e^{\frac{-(t-t_r)}{\tau_T^0}}}{\left(J(t_r,\nu_r)+\frac{\lambda_0}{\tau_T^0}+\frac{\lambda_1}{\tau_T^1}\right)-\left(J(t_r,\nu_r)+\frac{\lambda_1}{\tau_T^1}\right)h(\nu;\nu_r)e^{\frac{-(t-t_r)}{\tau_T^0}}} - \frac{\lambda_1}{\tau_T^1}, \tag{3.25}$$

$$h(\nu;\nu_r) = \frac{1-e^{\frac{-\nu}{\tau_C^0}}}{1-e^{\frac{-\nu_r}{\tau_C^0}}}, \tag{3.26}$$

where (3.24)–(3.26) *are symmetric with respect to Conditions* 3.10 *and* 3.11. *Furthermore* (3.25) *is satisfied for at most one value of* τ_C^0.

Proof. Equation (3.24) is obtained by setting $s_H^0 = s_T^0$ in (3.18). Substituting (3.24) into (3.21) in which we impose $s_H^0 = s_T^0$, and $\psi_1(t;\nu) = 0$, we get (3.21). Symmetry of (3.24) and (3.25) with respect to Conditions 3.10 and 3.11 follows from the symmetry of (3.18) and (3.21) to the same conditions. To demonstrate the uniqueness of τ_C^0, solve (3.25) for $h(\nu)$, which leads to

$$\rho(t,\nu)e^{\frac{-\nu_r}{\tau_c^0}} - e^{\frac{-\nu}{\tau_c^0}} = \rho(t,\nu) - 1, \tag{3.27}$$

$$\rho(t,\nu) = \frac{\left(J(t_r,\nu_r)+\frac{\lambda_0}{\tau_T^0}+\frac{\lambda_1}{\tau_T^1}\right)\left(J(t,\nu)+\frac{\lambda_1}{\tau_T^1}\right)}{\left(J(t,\nu)+\frac{\lambda_0}{\tau_T^0}+\frac{\lambda_1}{\tau_T^1}\right)\left(J(t_r,\nu_r)+\frac{\lambda_1}{\tau_T^1}\right)}. \tag{3.28}$$

For a specific time $t_0 \in [t_a, t_b]$ we search for a value of τ_C^0 satisfying the above relation. Since the left member of (3.27) is equal to $\rho(t,\nu) - 1$ as $1/\tau_C^0 \to 0$, approaches 0 as $1/\tau_C^0 \to \infty$, and has an extrema at

$$\tau_C^0 = \frac{(\nu - \nu_r)}{\ln\left(\nu/(\rho(t,\nu)\nu_r)\right)}, \tag{3.29}$$

(3.27) admits at most one solution for τ_C^0 at any given time. □

In summary, the estimation of $s_i(V)$ proceeds as follows. Take sets $I(t; V, V_T)$, $t, V \in \Omega_{t,V}$ with $V < V_T$ and $t, V \in \Omega_{t,V}$ with $V > V_T$. Use Theorems 3.7, 3.1, and 3.8 to estimate $s^i(V)/s_T^i$. If the range over which $s_i(V)/s_T^i$ can be evaluated is too narrow, augment the data suite at the origin of the estimation with sets $I(t; \nu_r; V, V_C, V_T)$, $t, V \in \Omega_{t,V}$, and $I(t; \nu; V_T, V_C, V_T)$, $t, \nu \in \Omega_{t,\nu}$, for which either $V, V_C > V_T$ or $V, V_C < V_T$. Use Lemma 3.13 and Theorem 3.9 to evaluate $s_i(V)/s_T^i$. Repeat the same procedure with different values of V_C until the range over which $s_i(V)/s_T^i$ can be evaluated is sufficiently large to appropriately define $s_i(V)$.

4. Estimation of the time constants, $\tau_i(V)$. Transformations of the previous section enable us to estimate all model parameters. However, all of them require the test potential of the stimuli (single or double step) to be outside the conductance window. Since estimation of $\tau_i(V)$ requires analyzing currents generated at any test potential, we need to develop other transformations which do not present any constraints as to where V_T can be set. From this point we assume $s_i(V)$ known and take advantage of this situation to develop an inversion procedure for $\tau_i(V)$. The basic idea consists of determining ranges of possible values of the model parameter $R = 1/\bar{g}$. We show that there exist disjoint ranges for R which enable the model to reproduce a data point. To any value of R picked within the allowed ranges, we associate a finite number of time constants which enable the model to reproduce a given data point. From these time constants and continuity criteria, we determine functions of voltage which complete the parameter estimation.

The cornerstone of the inversion is provided below.

THEOREM 4.1. *For an arbitrary data point taken on $I(t_r)$, $t_r \in [t_a, t_b]$,*

$$\prod_{i \in \bar{\mathcal{A}}} \gamma_i(y_i; V_H, V_T) \;=\; e^{t_r J(t_r)} \prod_{i \in \mathcal{A}} \gamma_i(y_i; V_H, V_T) \tag{4.1}$$

with

$$\gamma_i \;=\; \begin{cases} \left[\frac{s_T^i - s_H^i}{s_T^i - y_i}\right]^{\frac{\lambda_i (s_T^i - y_i)}{y_i}} & \text{if } i \in \mathcal{A}, \\ \left[\frac{s_T^i - s_H^i}{y_i - s_T^i}\right]^{\frac{\lambda_i (y_i - s_T^i)}{y_i}} & \text{if } i \in \bar{\mathcal{A}}, \end{cases} \tag{4.2}$$

where $\mathcal{A}$ is the set of indices for which sign $ds_i(V)/dV =$ sign $(V_T - V_H)$, *and $\bar{\mathcal{A}}$ its complement. The functions γ_i are defined in the range $y_i \in [s_H^i, s_T^i]$, they are equal to unity at each end of their domain, and they have one and only one extremum.*

Proof. Substitute (2.6) into (3.3) to get

$$J(t) = -\sum_{i=0}^{\bar{n}} \left(\frac{\lambda_i}{\tau_i} \frac{y_i - s_T^i}{y_i} \right). \tag{4.3}$$

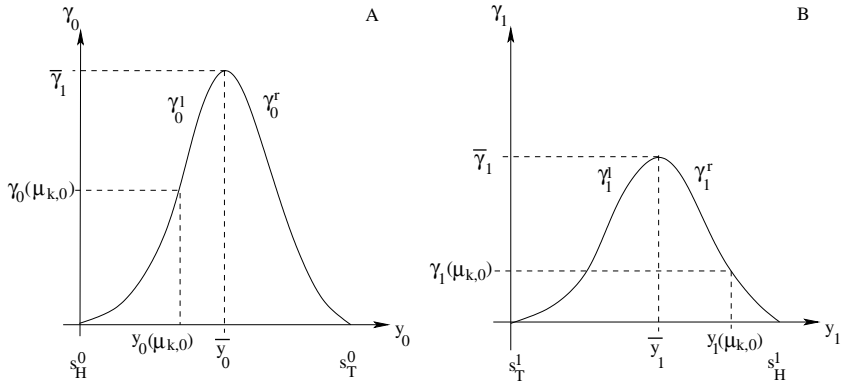


FIG. 4.1. *Functions* $\gamma_0(y_0)$ (A) *and* $\gamma_1(y_1)$ (B). *Segments on the left and right sides of the extrema are respectively denoted* γ_i^l, γ_i^r, $i \in [0,1]$.

Solve (2.4) for $1/\tau_T^i$ and substitute the resulting expression into (4.3); (4.1) follows immediately. One may easily verify that $\gamma_i(y_i)$ has one and only one extremum at

$$\bar{y}_i = s_T^i \ln\left[\frac{s_T^i - s_H^i}{s_T^i - \bar{y}_i}\right], \tag{4.4}$$

$\gamma_i = 1$ at $y_i = s_H^i, s_T^i$, and that

$$\gamma(\bar{y}_i) = \begin{cases} e^{\frac{\lambda_i(s_T^i - \bar{y}_i)}{s_T^i}} & \text{if } i \in \mathcal{A}, \\ e^{\frac{\lambda_i(\bar{y}_i - s_T^i)}{s_T^i}} & \text{if } i \in \bar{\mathcal{A}}. \end{cases} \tag{4.5}$$

The reader is referred to [2] for more details about the above derivations. □

Typical functions γ are shown in Figure 4.1. Theorem 4.1 provides a restriction on y_i with respect to a data point. Note that the latter is a generalization of a relation derived in Beaumont, Roberge, and Leon [2]. We use a parameter $\mu \in [0,1]$ to refer to a specific couple $(\gamma_0(\mu), \gamma_1(\mu))$ satisfying (4.1). Specifically, $[\gamma_a^i, \gamma_b^i]$ is the range of γ_i over which (4.1) can be satisfied. It is defined by

$$\left[\gamma_a^0, \gamma_b^0\right] = \left[\mathrm{Max}(1, e^{-tJ(t)}), \mathrm{Min}(\bar{\gamma}_0, e^{-tJ(t)}\bar{\gamma}_1)\right], \tag{4.6}$$

$$\left[\gamma_a^1, \gamma_b^1\right] = \left[\mathrm{Max}(1, e^{tJ(t)}), \mathrm{Min}(\bar{\gamma}_1, e^{tJ(t)}\bar{\gamma}_0)\right]. \tag{4.7}$$

The parameter μ of the parametric form of (4.1),

$$\gamma_i(\mu) = \gamma_a^i + (\gamma_b^i - \gamma_a^i)\mu, \qquad \mu \in [0,1], \tag{4.8}$$

allows us to systematically sweep this range (see Figure 4.2).

For the sake of the description of the inversion procedure, we introduce functions $\gamma_i^l(y_i)$ and $\gamma_i^r(y_i)$, which are the left and right (with respect to the extremum) branches of $\gamma_i(y_i)$ (Figure 4.1). Inversion of the left and right branches of these functions are respectively denoted by $y_i = \gamma_i^{-l}(\gamma_i)$ and $y_i = \gamma_i^{-r}(\gamma_i)$. Now we use Theorem 4.1 to bound ranges of $R = 1/\bar{g}$ allowing the model to reproduce a given data point. First we define

$$R_{m,n}(\mu) = \frac{V - e_k}{I(t_r; V_H, V_T)} \left[\gamma_0^{-m}(\gamma_0(\mu))\right]^{\lambda_0} \left[\gamma_1^{-n}(\gamma_1(\mu))\right]^{\lambda_1}, \tag{4.9}$$

$$m, n \in [(l,l),(r,r),(l,r),(r,l)],\ \mu \in [0,1].$$

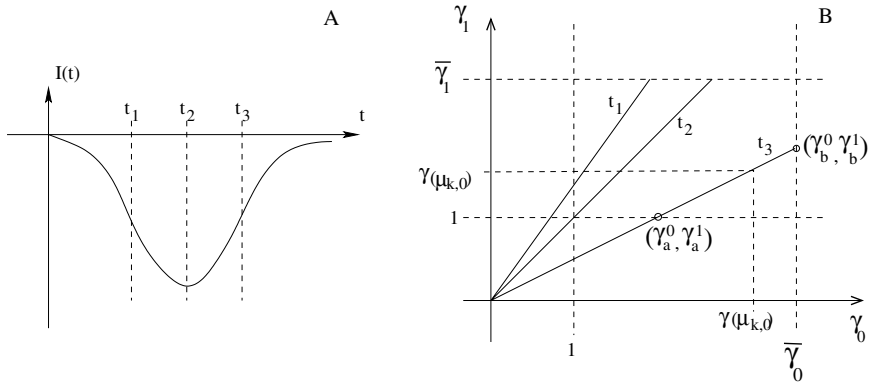


FIG. 4.2. *Implication of Theorem* 4.1. A: *Current trace in time.* B: *Geometrical construction associated with* (4.2) *for the three specific data points shown in* A.

Note that the order of the indices in $R_{mn}(\mu)$ matters. The first index is related to gating variable 0 and the second to gating variable 1.

From (4.10) it is obvious that $R_{m,m}(\mu)$, $m \in [l, r]$, are monotonic, and $R_{m,n}(\mu)$, $m, n \in [l, r]$, $m \neq n$, can be, and in fact are most of the time, nonmonotonic. Note that the line (4.8) crosses either the axis $\gamma_0 = \gamma_b^0$ or $\gamma_1 = \gamma_b^1$ of the (γ_0, γ_1) plane (Figure 4.2). Therefore each function $R_{m,n}(\mu = 1)$, $m, n \in [l, r]$, $m \neq n$, intersects either $R_{ll}(\mu = 1)$ or $R_{rr}(\mu = 1)$ at $\mu = 1$.

The functions $R_{m,n}(\mu)$ $m, n \in [(l, l), (r, r), (l, r), (r, l)]$ allow us to bound the inverse solution in $\mu \in [0, 1]$.

DEFINITION 4.2 (Bounds$\{R(\mu)\}$). *Ranges of* R *for which it is possible to invert the solution are denoted by* Bounds$\{R(\mu)\}$. *Such bounds are evaluated from the functions* $R_{m,n}(\mu)$, $m, n \in [(l, l), (r, r), (l, r), (r, l)]$. *Formally, when* $R_{l,r(\mu)}$ *and* $R_{r,l}(\mu)$, $\mu \in [0, 1]$, *intersect,*

$$\text{Bounds}\{R(\mu)\} = [\text{Min}\{R_{ll}(\mu)\}, \text{Max}\{R_{rr}(\mu)\}], \quad \mu \in [0, 1]. \tag{4.10}$$

However, when they don't intersect, the bounds are constituted by two disjoint ranges. Specifically, if $\gamma_0(\bar{y}_0) > \gamma_1(\bar{y}_1)$, *then* $R_{l,l}(\mu = 1) = R_{l,r}(\mu = 1)$, $R_{r,r}(\mu = 1) = R(r, l)(\mu = 1)$, *and*

$$\begin{aligned}\text{Bounds}\{R(\mu)\} = {} & [\text{Min}\{R_{ll}(\mu)\}, \text{Max}\{R_{l,r}(\mu)\}] \\ & \cup [\text{Min}\{R_{r,l}(\mu)\}, \text{Max}\{R_{,rr}(\mu)\}], \ \ \mu \in [0, 1];\end{aligned} \tag{4.11}$$

if $\gamma_0(\bar{y}_0) < \gamma_1(\bar{y}_1)$, *then* $R_{l,l}(\mu = 1) = R_{r,l}(\mu = 1)$, $R_{r,r}(\mu = 1) = R(l, r)(\mu = 1)$, *and*

$$\begin{aligned}\text{Bounds}\{R(\mu)\} = {} & [\text{Min}\{R_{ll}(\mu)\}, \text{Max}\{R_{r,l}(\mu)\}] \\ & \cup [\text{Min}\{R_{l,r}(\mu)\}, \text{Max}\{R_{r,r}(\mu)\}], \ \ \mu \in [0, 1].\end{aligned} \tag{4.12}$$

Due to the nature of $R_{m,n}(\mu)$, $m, n \in [l, r]$, $m \neq n$, in general their inversion at a specific $R_k \in \text{Bounds}\{R(\mu)\}$ leads to a set of coordinates μ, which we denote by $[\mu_{k0} \cdots \mu_{kn}] = R^{-1}(R_k)$ (Figure 4.3). To this set correspond couples $\gamma_0(\mu_{k,i}), \gamma_1(\mu_{k,i})$ (Figure 4.2). For each of these couples we can obtain a specific value of y_i by inverting the appropriate branch of $\gamma_i(y_i)$ (Figure 4.1). Since the steady states are known, the time constants $\tau_i(V)$ follow from (2.4).

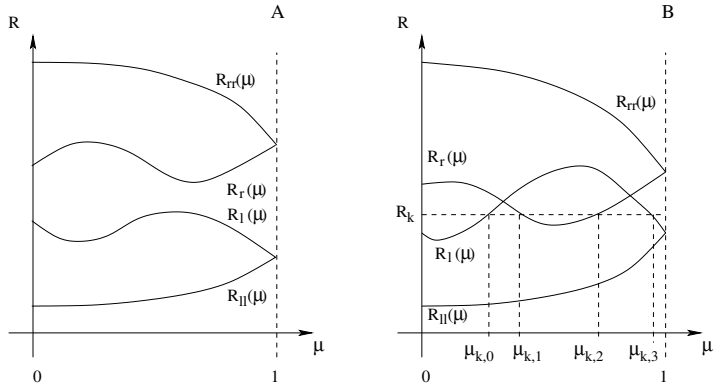


FIG. 4.3. *Functions $R_{ij}(\mu)$ built with the* 4 *different branches of the functions γ. See text for more details about the construction. Functions $R_r(\mu)$ and $R_l(\mu)$ may* (A) *or may not* (B) *intersect.*

From the above, the general inversion procedure should be obvious to the reader. Take sets $I(t; V_H, V)$, $t, V \in \Omega_{t,V}$. Find Bounds$\{R(\mu)\}$ over $\Omega_{t,V}$. The inverse solution ($R_k = 1/\bar{g}$) lies within the global bounds, which are the intersection of Bounds$\{R(\mu)\}$ evaluated in $\Omega_{t,V}$. We refer to these global bounds by Bounds$\{R(\mu)\} \cap \Omega_{t,V}$. If the data sufficiently constrain the model, Bounds$\{R(\mu)\} \cap \Omega_{t,V}$ should be very narrow. Then for each $R_k \in$ Bounds$\{R(\mu)\} \cap \Omega_{t,V}$ (ideally one value), evaluate the time constants, i.e., invert $\gamma_i(\mu)$, then $\gamma_i(y_i)$, and solve (2.4) for $\tau_i(V)$. Even if the bounds are very narrow, the inversion of $\tau_i(V)$ produces a structure with multiple branches along the V axis since inversion $R^{-1}(R_k)$ is not unique. Because the inverse solution exists, at least one root of such structure should traverse $[V_{T0}, V_{Tn}]$. If the data set does not sufficiently constrain the model, then several branches may traverse $[V_{T0}, V_{Tn}]$.

Inversion of $R_{m,n}(\mu)$ $m, n \in [l, r]$, $m \neq n$, requires determining the location of all extrema of each function. In Beaumont, Roberge, and Leon [2] it was shown that when $e^{tJ(t)} = 1$ (i.e., for peak currents), these functions have at most two extrema. At this time, it has not been possible to generalize the proof for the case where $e^{tJ(t)} \neq 1$. However, below we devise a numerical algorithm which allows us to determine the number of extrema and locate all of them. First, we state the following.

THEOREM 4.3. *For any data point taken on $I(t_r; V_H; V_T)$, $t_r \in [t_a, t_b]$, at the extrema of $R_{mn}(\mu)$, $m \neq n$, $m, n \in [l, r]$, $\mu \in [0, 1]$,*

$$\sigma_0^m(\mu) = \sigma_1^n(\mu), \qquad m \neq n,\ m, n \in [l, r], \tag{4.13}$$

with

$$\sigma_i(y_i) = \left[\frac{s_T^i - y_i}{s_T^i - s_H^i}\right]^{s_T^i/y_i}, \qquad \sigma_i^m(\mu) = \left[\frac{s_T^i - y_i^m(\mu)}{s_T^i - s_H^i}\right]^{s_T^i/y_i^m(\mu)}, \tag{4.14}$$

and

$$y_i^m(\mu) = \gamma_i^{-m}(\gamma_i(\mu)), \quad m \in [l, r]\ \ i \in [0, 1]. \tag{4.15}$$

Functions $\sigma_i^m(\mu)$ are monotonic with sign $d\sigma_0^m(\mu)/d\mu$ = sign $d\sigma_1^n(\mu)/d\mu$ if $m \neq n$, and are bounded between in $\mu[0, 1]$.

Proof. From the derivative chain rule we have

$$\frac{\partial R_{mn}(\mu)}{\partial \mu} = 0 \quad \rightarrow \quad \frac{y_0^m(\mu)}{\lambda_0\left(\gamma_b^0 - \gamma_a^0\right)} \frac{d\gamma_0^m}{dy_0} = \frac{-y_1^n(\mu)}{\lambda_1\left(\gamma_b^1 - \gamma_a^1\right)} \frac{d\gamma_1^n}{dy_1}. \tag{4.16}$$

Since $\partial \gamma_i^j(y_i)/\partial y_i \neq 0$, $j \in [l, r]$, $i \in [0, 1]$,

$$\frac{d\gamma_i^{-j}(\gamma_i)}{d\gamma_i} = \frac{dy_i}{d\gamma_i^j} = \frac{1}{d\gamma_i^j/dy_i}, \qquad i, j \in [l, r]. \tag{4.17}$$

Inserting $d\gamma_i^j(y_i)/dy_i$, $j \in [l, r]$, into (4.16), we get (4.13).

We now examine the shape of $\sigma_i(\mu)$ in $\mu \in [0, 1]$. From $d\sigma_i(y_i)/dy_i = 0$ we have

$$e^{\frac{-y_i}{(s_T^i - y_i)}} = \frac{(s_T^i - y_i)}{(s_T^i - s_H^i)}. \tag{4.18}$$

The left and right terms of (4.18) are monotonic. The left term is bounded between $[0, 1]$ in the range $y_i \in [s_H^i, s_T^i]$, and is 0 at $y_i = s_T^i$. Since its second derivative is positive in this range, (4.18) cannot be satisfied, and therefore $\sigma_i(y_i)$ is monotonic. Since

$$\frac{d\sigma_i^m(\mu)}{d\mu} = \frac{d\gamma(\mu)}{d\mu} \frac{\frac{d\sigma_1^m(y_i)}{dy_i}}{\frac{d\gamma_i^m(y_i)}{dy_i}}, \tag{4.19}$$

$\operatorname{sign} d\sigma_0(y_0)/dy_0 \neq \operatorname{sign} d\sigma_1(y_1)/dy_1$, and $\operatorname{sign} d\gamma_0^m(y_0)/dy_0 \neq \operatorname{sign} d\gamma_1^n(y_1)/dy_1$, then $\operatorname{sign} d\sigma_0^m(\mu)/d\mu = \operatorname{sign} d\sigma_1^n(\mu)/d\mu$, $m, n \in [l, r]$, $m \neq n$, $\mu \in [0, 1]$. □

We are now ready to expose the algorithm for the determination of the extrema of $R_{mn}(\mu)$, $m, n \in [l, r]$, $m \neq n$. In this algorithm we need to invert $\sigma_i^m(\mu)$, $m \in [l, r]$, $i \in [0, 1]$. We denote the inversion by $\mu_k = \sigma_i^{-m}(\sigma_k)$, σ_k being a specific value of σ.

ALGORITHM 4.4.

1. *Take a function $R_{mn}(\mu)$, $m, n \in [l, r]$, $m \neq n$, $\mu \in [0, 1]$.*
2. *Start at an arbitrary point $\mu_k \in [0, 1]$.*
3. *If $\sigma_0^m(\mu_k) > \sigma_1^n(\mu_k)$, then $p = m$, $q = n$, $j = 0$, and $k = 1$. Else $p = n$, $q = m$, $j = 1$, $k = 0$.*
4. *If functions $\sigma_0^m(\mu)$ and $\sigma_1^n(\mu)$ have multiple intersections, they form loops within which one can move until crossing points are reached. Specifically, to move right along the μ axis, i.e., $\mu_k \rightarrow \mu_{k+1}$ with $\mu_{k+1} > \mu_k$, proceed as follows: $\mu_{k+1} = \sigma_j^{-p}\left(\sigma_k^q(\mu_k)\right)$.*
5. *Move left along the μ axis, i.e., $\mu_k \rightarrow \mu_{k-1}$ with $\mu_{k-1} < \mu_k$ as follows: $\mu_{k-1} = \sigma_k^{-q}\left(\sigma_j^p(\mu_k)\right)$.*
6. *Repeat step 4 to obtain $\mu_{k+1} \rightarrow \mu_{k+2}$, and step 5 to obtain $\mu_{k-1} \rightarrow \mu_{k-2}$, until an intersection between $\sigma_0^m(\mu)$ and $\sigma_1^n(\mu)$ is found or until an end point of the search interval is reached.*
7. *Repeat these operations starting from step 4 until the excursion performed over the μ-axis covers the search interval, i.e., $[0, 1]$.*

As mentioned in the previous section, for some models it is not possible to obtain measurable currents when V_T of a step voltage clamp simulation is outside the conductance window. In this case, we can proceed as for the estimation of $s_i(V)$ by interpolating a conditioning pulse prior to the application of the test pulse. Fortunately, the theory developed in this section is still applicable. It requires only

$$s_H^i \rightarrow s_C^i + (s_H^i - s_C^i)e^{-\nu/\tau_C^i}, \tag{4.20}$$

which operation can be performed before starting the inversion of the time constants.

5. Conclusions. Procedures currently used to estimate the parameters of the HH gating model are all based on nonlinear least square fitting. It is well known that this approach has several drawbacks. Namely, the parameters obtained in this manner may correspond to a local minimum of the merit function undergoing minimization. Thus not only may estimates lack accuracy, but there is no way to know whether other minima exist. Knowing the existence of such minima is important for the modelling methodology because, considering the experimental error, several of them may correspond to plausible inverse solutions. In this case gating model predictions related to each minimum need to be examined. If model predictions are too wide-ranging, then additional experiments need to be performed to obtain more specific current kinetics. While this is a highly desirable approach to the modelling of bioelectric phenomena, it is not possible to pursue in practice due to the inability to locate, in a rigorous manner, minima of the objective function.

Here, we have proposed an inversion methodology which overcomes all these limitations. The procedure is accurate because it guarantees that the model will exactly reproduce the data set. If the solution does not sufficiently constrain the model, bounds for the inverse solution are large, and it is possible to identify a model family capable of accurately reproducing the data set at the origin of the estimation. In addition, since the inversion relates the functions of the model ($s_i(V)$ and $\tau_i(V)$) to the experimental data, the estimation procedure specifies in which potential range current kinetics need to be better characterized. In other words the estimation method may help to design experiments.

Is the inverse solution of the HH gating model with respect to a given voltage clamp data set unique? Answering this question is not trivial, but is crucial for the modelling methodology. As we have stressed above, due to the existence of multiple inverse solutions, conclusions drawn from simulation results can be entirely erroneous. At this point it should be clear to the reader that if the data set does not probe the steady states and time constants over a potential range sufficiently large to define these functions appropriately, the model is under-determined. Unfortunately this seems to be the case with most data sets employed in common practice.

Furthermore, incomplete data sets result in an inverse solution which is characterized by disjoint ranges of valid values. Consequently in this case very different models can reproduce a given experimental data set well. This warrants great care in drawing conclusions from simulations carried out with macroscopic models. This said, in many cases the lack of data is due to the technical difficulties discussed in section 4. Here we have provided means to analyze data generated with more elaborate protocols (double step) than the conventional step stimulation. We believe this will provide experimentalists the flexibility they need to overcome the difficulties commonly encountered.

Acknowledgment. We are grateful to Dr. Lawrence J. Lardy from the Department of Mathematics at Syracuse University for fruitful discussions and for assisting with the preparation of this manuscript.

REFERENCES

[1] D.A. Baxter, C.C. Canavier, J.W. Clark, Jr., and J.H. Byrne, *Numerical model of the serotonergic modulation of sensory neurons in Aplysia*, J. Neurophysiol., 82 (1999), 2914–2935.

[2] J. Beaumont, F.A. Roberge, and L.J. Leon, *On the interpretation of voltage-clamp data using the Hodgkin–Huxley model*, Math. Biosci., 115 (1993), pp. 65–101.

[3] J. Beaumont, F.A. Roberge, and D.R. Lemieux, *Estimation of the steady-state characteristics of the Hodgkin–Huxley model from voltage clamp data*, Math. Biosci., 11 (1993), pp. 145–186.
[4] J. Beaumont, N. Davidenko, J.M. Davidenko, and J. Jalife, *Spiral waves in two-dimensional models of ventricular muscle: Formation of a stationary core*, Biophys. J., 75 (1998), pp. 1–14.
[5] J. Beaumont, N. Davidenko, and A. Goodwin, *Vortices of electrical waves in the heart muscle. Mechanisms of stabilization at high frequencies*, Biophys. J., submitted.
[6] L. Beccuci, M.R. Moncelli, and R. Guidelli, *Pore formation by 6-ketocholestanol in phospholipid monolayers and its interpretation by a general nucleation-and-growth model accounting for the sigmoidal shape of voltage-clamp curves of ion channels*, J. Amer. Chem. Soc., 125 (2003), pp. 3784–3792.
[7] J.A. Benett and B.J. Roth, *Time dependence of anodal and cathodal refractory periods in cardiac tissue*, Pacing & Clinical Electrophysiol., 22 (1999), pp. 1031–1038.
[8] L.A. Cartee, C. van den Honert, C.C. Finley, and R.L. Miller, *Evaluation of a model of the cochlear neural membrane.* I. *Physiological measurement of membrane characteristics in response to intrameatal electrical stimulation*, Hearing Res., 146 (2000), pp. 143–152.
[9] L.A. Cartee, *Evaluation of a model of the cochlear neural membrane.* II: *Comparison of model and physiological measures of membrane properties measured in response to intrameatal electrical stimulation*, Hearing Res., 146 (2000), pp. 153–166.
[10] P. Comtois and A. Vinet, *Curvature effects on activation speed and repolarization in an ionic model of cardiac myocytes*, Phys. Rev. E., 60 (1999), pp. 4619–4628.
[11] J. Cronin, *Mathematical Aspects of Hodgkin–Huxley Neural Theory*, Cambridge University Press, London, Cambridge, 1987.
[12] O. Ekeberg and S. Grillner, *Simulations of neuromuscular control in lamprey swimming*, Phil. Trans. R. Soc. London B, 354 (1999), pp. 895–902.
[13] F. Fenton and A. Karma, *Fiber-rotation-induced vortex turbulence in thick myocardium*, Phys. Rev. Lett., 81 (1998), pp. 481–484.
[14] R.J. Greenberg, T.J. Velte, M.S. Humayun, G.N. Scarlatis, and E. de Juan, Jr., *A computational model of electrical stimulation of the retinal ganglion cell*, IEEE Trans. Biomed. Engrg., 46 (1999), pp. 505–514.
[15] O.P. Hamill, A. Marty, E. Neher, B. Sackmann, and F.J. Sigworth, Plügers Arch., 391 (1981), pp. 85–100.
[16] D.M. Harrild and C.S. Henriquez, *A Computer Model of Normal Conduction in the Human Atria*, Circ. Res., 87 (2000), pp. e25–e36.
[17] B. Hille, *Ionic Channels of Excitable Membranes*, Sinauer, Sunderland, MA, 1992.
[18] A.L. Hodgkin and A.F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol., 117 (1952), pp. 500–544.
[19] Y. Kamiyama, T. Ogura, and S. Usui, *Ionic current model of the vertebrate rod photoreceptor*, Vision Res., 36 (1996), pp. 4059–4068.
[20] A. Karma, H. Levine, and X. Zou, *Theory of pulse instabilities in electrophysiological models of excitable tissues*, Phys. D, 73 (1994), pp. 113–127.
[21] W. Krassowska, *Field stimulation of cardiac fibers with random spatial structure*, IEEE Trans. Biomed. Engrg., 50 (2003), pp. 33–40.
[22] S. Kubata and K. Sakai, *Relationship between obsessive-compulsive disorder and chaos*, Medical Hypothesis, 59 (2002), pp. 16–23.
[23] W.K. Luk and K. Aihara, *Synchronization and sensitivity enhancement of the Hodgkin–Huxley neurons due to inhibitory inputs*, Biological Cybernetics, 82 (2000), pp. 455–467.
[24] W.W. Lytton, *Adapting a feedforward heteroassociative network to Hodgkin–Huxley dynamics*, J. Comput. Neurosci., 5 (1998), pp. 353–364.
[25] J.M. Meunier, J.C. Eason, and N. Trayanova, *Termination of reentry by long-lasting AC shock in a slice of canine heart: A computational study*, J. Cardiovasc. Electrophys., 13 (2002), pp. 1253–1261.
[26] J. Patlak, *Molecular kinetics of voltage-dependent Na channels*, Physiol. Rev., 71 (1991), pp. 1047–1080.
[27] Z. Qu, J. Kil, F. Xie, A. Garfinkel, and J.N. Weiss, *Scroll wave dynamics in three-dimensional cardiac tissue model: Roles of restitution, thickness, and fiber rotation.*, Biophys. J., 78 (2000), pp. 2761–2775.
[28] Z. Qu, J.N. Weiss, and A. Garfinkel, *Cardiac electrical restitution properties and stability of reentrant spiral waves: A simulation study*, Am. J. Physiol., 276 (1999), pp. H269–H283.
[29] F.H. Samie, O. Berenfeld, J. Anumonwo, S.F. Mironov, S. Udassi, J. Beaumont, S. Taffet, and A.M. Pertsov, *Rectification of the background potassium current. A determinant of rotor dynamics in ventricular fibrillation.* Circ. Res., 86 (2000), pp. 1216–1223.

[30] P.R. Shorten and D.J. Wall, *A Hodgkin–Huxley model exhibiting bursting oscillations*, Bull. Math. Biol., 62 (2000), pp. 695–715.
[31] R.M Siegel and H.L. Read, *Deterministic dynamics emerging from a cortical functional architecture*, Neural Networks, 14 (2001), pp. 697–713.
[32] M. Stemmler and C. Koch, *How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate*, Nature Neurosci., 2 (1999), pp. 521–517.
[33] E.J. Vigmond, R. Ruckdeschel, and N. Trayanova, *Reentry in a morphologically realistic atrial model*, J. Cardiovasc. Electrophys., 12 (2001), pp. 1046–1054.
[34] A.R. Willms, D.J. Barro, R.M. Harris-Warrick, and J. Guckenheimer, *An improved parameter estimation method for Hodgkin–Huxley model*, J. Comput. Neurosci., 6 (1999), pp. 145–168.
[35] A.R. Willms, *NEUROFIT: Software for fitting Hodgkin–Huxley models to voltage-clamp data*, J. Neurosci. Methods., 121 (2002), pp. 139–150.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1268–1293

SINGULAR PERTURBATIONS FOR BOUNDARY VALUE PROBLEMS ARISING FROM EXOTIC OPTIONS*

AYTAC ILHAN†, MATTIAS JONSSON‡, AND RONNIE SIRCAR†

Abstract. We study the pricing of three exotic derivative securities (barrier, lookback, and passport options) which can be characterized by boundary value PDE problems in the context of popular Markovian stochastic volatility models of stock prices. By extending the fast mean-reverting asymptotic analysis in [J.-P. Fouque, G. Papanicolaou, and K. R. Sircar, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, London, 2000], the usual "Greek" correction to the Black–Scholes prices of these contracts is further corrected by a boundary integral term that is rapidly computed numerically. In the case of the passport option, the asymptotic method is effective in accounting for stochastic volatility effects in a simple and robust fashion even in the presence of a highly nonlinear embedded stochastic control problem.

Key words. stochastic volatility, asymptotic approximations, option pricing

AMS subject classifications. 91B28, 93E20, 41A60, 30E25

DOI. 10.1137/S0036139902420043

1. Introduction.

In this paper we describe a framework for approximating the prices of certain path-dependent derivative securities to take into account the observed "implied volatility skew," which contains information about the market's view of the asymmetry and leptokurtosis in stock price returns. The pricing problems for these exotic options are characterized by boundary value problems for PDEs, under the class of stochastic volatility diffusion models we consider here. Our examples are a barrier option, a lookback option, and a passport option, whose prices solve Dirichlet, mixed, and Neumann boundary value problems, respectively. From the point of view of the practical application, there is a need for a quick calculation from which a trader can quote a price to a client. The approximation method used here is computationally fast and robust to specific modeling of the unobserved stochastic volatility process. The analysis extends the singular perturbation approximations for stochastic volatility models studied in [13].

1.1. Empirical foundation.

The basis of the approximations is a rapid time-scale of fluctuation in the stock price volatility relative to the time horizon of the options contract. Such a fast scale has been identified in market data in [16, 1, 4], for example, and is convenient for constructing approximations over times when other, slower, factors in the volatility can be considered relatively benign. Extension of the approach in [13] to incorporate a slower scale is begun in [15]. Asymptotic analysis of a different type of exotic path-dependent contract, Asian options, is studied in [12].

In [16], we studied high-frequency S&P 500 data over the period of a year. The result, using both variogram and spectral methods, was the clear presence of a fast

*Received by the editors December 18, 2002; accepted for publication (in revised form) September 1, 2003; published electronically May 5, 2004.
http://www.siam.org/journals/siap/64-4/42004.html
†Department of Operations Research & Financial Engineering, Princeton University, E-Quad, Princeton, NJ 08544 (ailhan@princeton.edu, sircar@princeton.edu). The research of these authors was partially supported by NSF grant SES-0111499.
‡Department of Mathematics, University of Michigan, Ann Arbor, MI 48109-1109 (mattiasj@umich.edu). The research of this author was partially supported by NSF grant DMS-0200614.

time-scale of volatility fluctuation, corresponding to a characteristic mean-reversion time on the order of a few days.

There is also considerable evidence of a slower scale. Many empirical studies have looked at low-frequency (daily) data, with the data necessarily ranging over a period of years, and have found a *low* rate of volatility mean-reversion. (These analyses of data at lower frequencies over longer time periods primarily pick up a slower time-scale of fluctuation and do not identify scales at the same order as the sampling frequency.) Often, the slow factor half-life is found to be on the order of 70–90 days in equity indices [10].

The combined conclusion has led recently to the study of *two-factor stochastic volatility models* [4], where one factor is slowly mean-reverting and the other is fast mean-reverting. Another recent empirical study [1], this time of exchange rate dynamics, finds "the evidence points strongly toward two-factor [volatility] models with one highly persistent factor and one quickly mean-reverting factor."

In the option-pricing asymptotics presented here, we focus on the effect of the fast volatility scale. This corresponds to assuming the lifetime of the derivative is on the order of the typical half-life of the slow factor, or less, because that factor would act approximately like a constant.

1.2. Alternative approaches. There are a number of other approaches to stochastic volatility modeling, and we briefly discuss only the more common ones here. Volatility models built on diffusions were introduced in the literature in the late 1980s by Hull and White [23], among others. One popular class of models builds on the Feller process model introduced in this context by Heston [20] because call option prices can be solved for in closed form up to a Fourier inversion.

Typically a lot of emphasis is placed on fitting the models very closely to observed implied volatilities (see subsection 1.3 for the definition), and not surprisingly, models with more degrees of freedom perform better in this regard. For example, the models studied in [2, 9] include jumps in stochastic volatility on top of a Heston-type model. However, little attention is paid to the stability of the estimated parameters over time, and it is the usual practice in the industry simply to recalibrate each day.

The approach taken here, based on modeling volatility in terms of its characteristic scales rather than specific distributions, sacrifices some of the goodness of in-sample fit to current data for greater stability properties. It also allows for efficient computation of approximations to prices of exotic contracts, such as considered here. Otherwise, these prices have to be found by simulations or numerical solution of a high-dimensional PDE associated with the full stochastic volatility model. Other authors have also studied this approximation [5] or applied it in different contexts, for example, pricing volatility derivatives [22] or analyzing trading volume [21].

1.3. Calibration from market-implied volatilities. The three options studied here are called exotic (and are listed in increasing order of "exoticness") because they are less heavily traded than standard "vanilla" call and put options. Lookbacks and passport options are usually sold as over-the-counter products. However, market vanilla option prices contain valuable information about the market's perception of future risks. This is typically expressed in units of implied volatility. Given the observed price C_{obs} of a European call option, which gives the holder the right but not the obligation to buy one unit of stock for strike price K on date T, the implied volatility I is defined as that volatility which equates the Black–Scholes option pricing

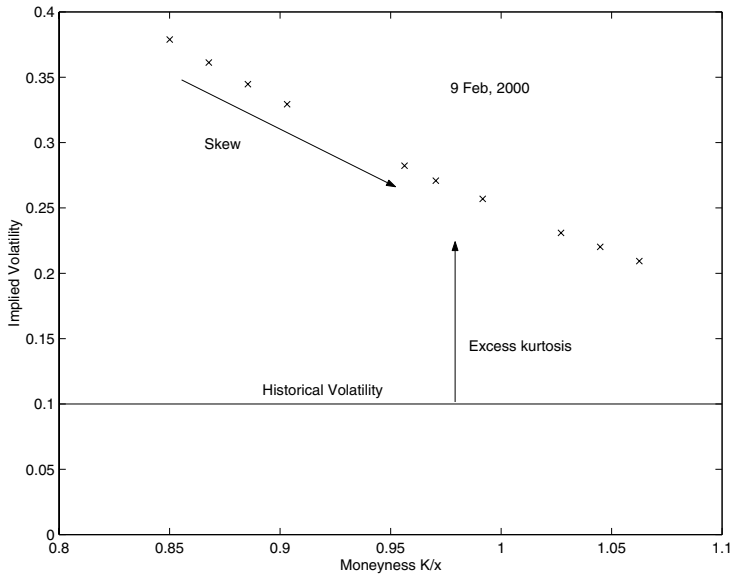


FIG. 1.1. *Implied volatility as a function of moneyness for fixed maturity options. The skew represents asymmetry in the returns distribution of the stock, and the increase in level over historical volatility is due to the excess kurtosis over lognormal models.*

formula $C^{BS}(t, X_t; K, T; I)$ to this price:

$$C^{BS}(t, X_t; K, T; I) = C_{\text{obs}}.$$

Here, t denotes the current time and X_t the current stock price.

A basic problem in financial engineering, given the market's implied volatilities, is to find prices of exotic contracts that are consistent with the principle of *no arbitrage*. (In general, there is no unique solution without making further assumptions on an underlying model.)

Under a large class of fast mean-reverting stochastic volatility models, it is shown in [13] that the implied volatility surface $I(K, T)$ (that is, I considered as a function of the option's strike price K and maturity date T for fixed t and X_t) is approximated by an affine function of the log-moneyness-to-maturity ratio (LMMR):

$$I \approx a \times \text{LMMR} + b, \quad \text{LMMR} = \frac{\log(K/X_t)}{T - t}, \tag{1.1}$$

where a and b are some market constants to be estimated by fitting this formula to option implied volatility data. See Figure 1.1.

Then, given estimates of the slope a and the intercept b, we consider the problem of finding consistent approximations for various exotic options. The cases of American and Asian options, as well as barriers, were studied in [13]. The latter contained an error in the calculation, and we include it here, corrected and in a somewhat different format from the subsequent erratum to [13], as our starting point.

When the formula (1.1) is fitted to certain regimes of S&P 500 implied volatility data, the estimated parameters a and b have good stability properties [13, 14]. This is particularly important for pricing path-dependent securities as considered here, because they depend not just on a one-time distribution of the stock price, but also

on the evolution of the process. The stability and goodness-of-fit, particularly to short-dated options, can be improved by including time-dependent periodic factors, as is done in [14].

Given the condensation of the pricing measure contained in a and b, we show that the asymptotic correction term in these boundary value problems is explicit up to a one-dimensional integral, which can be computed very quickly. It is then easy to gauge the impact of, for example, the slope of the implied volatility skew, measured by a, on the prices of these path-dependent contracts, as we illustrate numerically. The upshot of the analysis is that one obtains the usual asymptotic correction terms to the Black–Scholes prices of the exotics (which can be expressed in terms of the "Greeks," or partial derivatives of the Black–Scholes prices) plus an additional boundary integral term that corrects the Greek correction for skew effects.

1.4. Hedging. Another important problem is hedging the risk of a position in exotic options using the underlying asset and perhaps other vanilla options which are liquidly traded. While this is a well-defined problem in a complete market model (for example, the constant volatility Black–Scholes model), and the hedging ratio, the number of stocks held in order to hedge perfectly, is a by-product of the pricing problem, this is not the case in incomplete market models, such as with stochastic volatility, where the additional randomness driving the volatility cannot be hedged because volatility is not a tradeable asset. Many studies on hedging in incomplete markets model the preferences of the individual or institutional hedger through a loss function. The hedging strategies are found as solutions of a stochastic control problem of minimizing the expected loss for a given initial hedging capital. We do not discuss these approaches in this paper. The example of minimizing expected shortfall is studied in [11], and asymptotic approximations in the case of fast mean-reverting stochastic volatility are constructed in [26].

For the Black–Scholes model, the hedging ratio for a barrier or lookback option is given by the Delta (the partial derivative with respect to x, the stock price) of the option price. It is natural to consider the same quantity, replacing the Black–Scholes price by the asymptotic approximation of the stochastic volatility price we derive here; such an approach was developed in [13, Chapter 7]. This type of strategy is not self-financing, but the value of the hedging portfolio is close to the value of the option. These are discussed in sections 2.5 and 3.5. The hedging problem for a passport option is discussed in section 4.6.

2. Barrier options. A *barrier option* is a path-dependent claim whose payoff depends on whether or not the underlying asset price hits a specified value before the maturity date. One example of a barrier option is the *down-and-out call option*, which gives the holder the right to buy the underlying asset on expiration date T for strike price K unless the asset price has hit the barrier B at some time before T, in which case the contract expires worthless. The payoff at expiration T can be written as

$$h(X_T) = (X_T - K)^+ \mathbf{1}_{\left\{\min_{0\le t\le T} X_t \ge B\right\}},$$

where $\mathbf{1}$ denotes the indicator function.

2.1. Asymptotic approximation. The fast mean-reverting stochastic volatility approximation for barrier options was studied in [13]. In this paper, we give a brief review, which derives the relevant PDE problems to solve for the terms in the

asymptotic expansion. In this case, the boundary condition arises naturally due to the structure of the option.

We shall look at stochastic volatility models in which volatility (σ_t) is driven by an ergodic process (Y_t) that approaches its unique invariant distribution at an exponential rate $1/\varepsilon$. The size of this rate captures the volatility decorrelation speed, and in particular we shall be interested in asymptotic approximations when ε is small, which describes *fast* mean-reverting volatility.

As explained in [13], it is convenient for exposition to take a specific simple example for (Y_t) and allow the generality of the modeling to be in the unspecified relation between volatility and this process: $\sigma_t = f(Y_t)$, where f is some positive (and sufficiently regular) function, bounded above and away from zero. Further, taking (Y_t) to be a Markovian Itô process allows us to simply model the asymmetry, or fatter left-tails of returns distributions, by incorporating a negative correlation between asset price and volatility shocks. We shall thus take (Y_t) to be a mean-reverting Ornstein–Uhlenbeck (OU) process, so that the stochastic volatility models we consider are

$$\begin{aligned} dX_t &= \mu X_t\,dt + f(Y_t)X_t\,dW_t, \\ dY_t &= \frac{1}{\varepsilon}(m - Y_t)\,dt + \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\left(\rho\,dW_t + \sqrt{1-\rho^2}\,dZ_t\right), \end{aligned} \tag{2.1}$$

where (X_t) is the stock price process. Here (W_t) and (Z_t) are independent standard Brownian motions on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and ρ is the instantaneous correlation between asset price and volatility shocks that captures the skew, asymmetry, or leverage effect. The asymptotic results as they are used are *not* specific to the choice of the OU diffusion process, nor do they depend on specifying f. In this scaling, the invariant density of Y is Gaussian, $\mathcal{N}(m, \nu^2)$, which does not depend on ε. Throughout the paper, we adopt the standard practice of denoting the initial values of the processes X and Y with lowercase letters, x and y.

The model (2.1) describes an *incomplete* market, meaning that not all contingent claims can be replicated by trading only in the underlying stock, the volatility process being untradeable. This has profound consequences for pricing, hedging, and calibration problems for derivative securities. By standard no-arbitrage pricing theory [8], there is more than one possible equivalent martingale (or risk-neutral pricing) measure $\mathbb{P}^{\star(\gamma)}$ because the volatility is not a traded asset; the nonuniqueness is denoted by the dependence on γ, which we identify as the market price of volatility risk.

By Girsanov's theorem, $(W_t^\star, Z_t^\star)$ defined by

$$W_t^\star = W_t + \int_0^t \frac{(\mu - r)}{f\,(Y_s)} ds, \qquad Z_t^\star = Z_t + \int_0^t \gamma_s ds$$

are independent Brownian motions under a measure $\mathbb{P}^{\star(\gamma)}$ defined by

$$\frac{d\mathbb{P}^{\star\,(\gamma)}}{d\mathbb{P}} = \exp\left(-\int_0^T \frac{(\mu - r)}{f\,(Y_s)} dW_s - \int_0^T \gamma_s dZ_s - \frac{1}{2}\int_0^T \left[\left(\frac{(\mu - r)}{f\,(Y_s)}\right)^2 + \gamma_s^2\right] ds\right),$$

assuming (γ_t) is a nonanticipating process with sufficient regularity.

In particular, γ_t is the risk premium factor from the *second* source of randomness Z that drives the volatility. As explained in [13], we take the view that the market selects a pricing measure identified by a particular γ which is reflected in liquidly

traded around-the-money European option prices. Other derivative securities must be priced with respect to this measure, if there are to be no arbitrage opportunities. We shall assume that the market price of volatility risk γ_t is a bounded function of the state Y_t: $\gamma_t = \gamma(Y_t)$. In general, it may be a more complicated function of time and the paths of the two Brownian motions, but we make the assumption that (X, Y) remains a time-homogeneous Markov process under the pricing measure and that Y remains a Markov process by itself under $\mathbb{P}^{\star(\gamma)}$. This simplification implies, in particular, that the market parameters V_2^{ε} (defined in (2.11) below) and, consequently, the implied volatility calibration parameter b in (1.1) do not depend on time or the level of the stock price x. The effects of this assumption are reflected in fitting the model to data; the performance study in [14] shows that the fitted b is typically quite stable.

Under $\mathbb{P}^{\star(\gamma)}$, (X, Y) evolves according to

$$dX_t = rX_t dt + f(Y_t)X_t dW_t^{\star}, \tag{2.2}$$

$$dY_t = \left[\frac{1}{\varepsilon}(m - Y_t) - \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\Lambda(Y_t)\right] dt + \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\left(\rho\, dW_t^{\star} + \sqrt{1-\rho^2}\, dZ_t^{\star}\right), \tag{2.3}$$

where

$$\Lambda(Y_t) = \rho\frac{(\mu - r)}{f\,(Y_t)} + \sqrt{1-\rho^2}\,\gamma(Y_t).$$

We also define the infinitesimal generator $\mathcal{L}^{\varepsilon}$ of (X, Y) under this measure and write it grouped in powers of ε as follows:

$$\mathcal{L}^{\varepsilon} = \frac{1}{\varepsilon}\mathcal{L}_0 + \frac{1}{\sqrt{\varepsilon}}\mathcal{L}_1 + \mathcal{L}_2, \tag{2.4}$$

$$\mathcal{L}_0 = \nu^2\frac{\partial^2}{\partial y^2} + (m - y)\frac{\partial}{\partial y}, \tag{2.5}$$

$$\mathcal{L}_1 = \sqrt{2}\nu\rho f(y)x\frac{\partial^2}{\partial x \partial y} - \sqrt{2}\nu\Lambda(y)\frac{\partial}{\partial y}, \tag{2.6}$$

$$\mathcal{L}_2 = \frac{\partial}{\partial t} + \frac{1}{2}f(y)^2x^2\frac{\partial^2}{\partial x^2} + r\left(x\frac{\partial}{\partial x} - \cdot\right). \tag{2.7}$$

Here $\mathcal{L}_0$ is the infinitesimal generator of the mean-reverting OU process, $\mathcal{L}_1$ contains the mixed derivative (from the correlation) and the market price of risk γ, and $\mathcal{L}_2$ is the Black–Scholes partial differential operator $\mathcal{L}_{BS}(f(y))$ at the volatility level $f(y)$.

The price $P(t, x, y)$ of the down-and-out barrier call option satisfies

$$\mathcal{L}^{\varepsilon}P = 0 \quad \text{in } x > B \text{ and } t < T,$$

with a terminal condition at $t = T$, $P(T, x, y) = (x - K)^+$, and a boundary condition at $x = B$, $P(t, B, y) = 0$. The latter expresses the knock-out condition on the barrier.

An asymptotic analysis detailed in [13] shows that the fast mean-reverting approximation (in the limit $\varepsilon \downarrow 0$) for the barrier option is given by

$$P(t, x, y) \approx P^{(0)}(t, x) + \widetilde{P^{(1)}}(t, x).$$

Here, $P^0(t, x)$ is the Black–Scholes price of the option with constant volatility parameter $\bar{\sigma}$, which is related to the original volatility model by

$$\bar{\sigma}^2 = \langle f^2 \rangle,$$

where $\langle\cdot\rangle$ denotes averaging with respect to the invariant density of Y, $\mathcal{N}(m,\nu^2)$. Notice that, under the fast volatility scaling, the first two terms of the expansion do not depend on y, the level of the unobservable process Y. Therefore, $P^{(0)}$ is the solution of the homogenized boundary value problem

$$\begin{aligned} \mathcal{L}_{BS}(\bar\sigma)P^{(0)} &= 0 \quad \text{in } x > B \text{ and } t < T, \\ P^{(0)}(T,x) &= (x-K)^+, \\ P^{(0)}(t,B) &= 0. \end{aligned} \tag{2.8}$$

We can obtain a formula for $P^{(0)}(t,x)$ by the method of images (see [29], for example)

$$P^{(0)}(t,x) = C^{BS}(t,x;\bar\sigma) - \left(\frac{x}{B}\right)^{1-k} C^{BS}(t,B^2/x;\bar\sigma), \tag{2.9}$$

where $k = 2r/\bar\sigma^2$ and $C^{BS}(t,x;\bar\sigma)$ is the Black–Scholes pricing formula for *a call option*, with the volatility parameter $\bar\sigma$:

$$\begin{aligned} C^{BS}(t,x;\bar\sigma) &= xN(d_1) - Ke^{-r(T-t)}N(d_2), \\ d_1 &= \frac{\log(x/K) + \left(r + \frac{1}{2}\bar\sigma^2\right)(T-t)}{\bar\sigma\sqrt{T-t}}, \\ d_2 &= d_1 - \bar\sigma\sqrt{T-t}, \end{aligned}$$

and N denotes the standard cumulative normal distribution function.

2.2. First-order correction. From the asymptotic calculations in [13], the stochastic volatility correction $\widetilde{P^{(1)}}(t,x)$, which is of order $\sqrt{\varepsilon}$, satisfies the PDE problem

$$\begin{aligned} \mathcal{L}_{BS}(\bar\sigma)\widetilde{P^{(1)}} &= \mathcal{A}P^{(0)} \quad \text{in } x > B \text{ and } t < T, \\ \widetilde{P^{(1)}}(T,x) &= 0, \\ \widetilde{P^{(1)}}(t,B) &= 0, \end{aligned}$$

where $\mathcal{A}$ is defined as

$$\mathcal{A} = V_3^\varepsilon x\frac{\partial}{\partial x}\left(x^2\frac{\partial^2}{\partial x^2}\right) + V_2^\varepsilon x^2\frac{\partial^2}{\partial x^2}, \tag{2.10}$$

$$V_2^\varepsilon = -\frac{\nu\sqrt{\varepsilon}}{\sqrt{2}}\left\langle \Lambda\phi'\right\rangle, \tag{2.11}$$

$$V_3^\varepsilon = \frac{\rho\nu\sqrt{\varepsilon}}{\sqrt{2}}\left\langle f\phi'\right\rangle, \tag{2.12}$$

and $\phi(y)$ is a solution of $\mathcal{L}_0\phi(y) = f(y)^2 - \bar\sigma^2$. As shown in [13], the boundedness assumptions on f and γ imply that we can choose ϕ to have a bounded first derivative.

The interpretations of the two market constants above are as follows: V_2^ε contains the effect of the market price of volatility risk; V_3^ε contains the effect of the correlation, or skew, ρ. In the case of zero correlation, $V_3^\varepsilon = 0$ and our correction formulas (2.19), (3.10), and (4.18) below simplify and do not require numerical integration. However, in equity markets, ρ is typically estimated to be negative.

In practice, we do not use the homogenization formulas (2.11) and (2.12) to obtain V_3^ε and V_2^ε from a specific stochastic volatility model. Rather, they are calibrated

from liquid European options prices, or the implied volatility surface using the LMMR formula (1.1). As computed in [13], V_3^ε and V_2^ε are obtained from a and b in (1.1), and from the long-run mean historical volatility $\bar{\sigma}$ estimated from stock returns, by

$$V_2^\varepsilon = -\bar{\sigma}\left(a\left(r - \frac{1}{2}\bar{\sigma}^2\right) + (b - \bar{\sigma})\right),$$
$$V_3^\varepsilon = -a\bar{\sigma}^3.$$

The problem of solving this boundary value problem with a source term can be simplified to a one-dimensional integral by defining

$$\hat{P}(t,x) = \widetilde{P^{(1)}} + \frac{V_3^\varepsilon}{\bar{\sigma}} x P^{(0)}_{x\bar{\sigma}} + \frac{V_2^\varepsilon}{\bar{\sigma}} P^{(0)}_{\bar{\sigma}}$$

for $x \geq B$. Then $\hat{P}(t,x)$ solves

$$\begin{aligned} \mathcal{L}_{BS}(\bar{\sigma})\hat{P}(t,x) &= 0 \quad \text{in } x > B \text{ and } t < T, \\ \hat{P}(T,x) &= 0, \\ \hat{P}(t,B) &= \frac{V_3^\varepsilon}{\bar{\sigma}} g(t), \end{aligned} \tag{2.13}$$

where we define

$$g(t) = x P^{(0)}_{x\bar{\sigma}}\big|_{x=B}. \tag{2.14}$$

This is because the barrier option Vega $\mathcal{V} = P^{(0)}_{\bar{\sigma}}$ solves the PDE problem

$$\begin{aligned} \mathcal{L}_{BS}(\bar{\sigma})\mathcal{V} &= -\bar{\sigma}x^2 P^{(0)}_{xx} \quad \text{in } x > B \text{ and } t < T, \\ \mathcal{V}(T,x) &= 0, \\ \mathcal{V}(t,B) &= 0, \end{aligned}$$

as can be seen by formally differentiating (2.8) with respect to $\bar{\sigma}$. Differentiating again with respect to x, we can see that the Vega of the hedge $U = x P^{(0)}_{x\bar{\sigma}}$ satisfies

$$\mathcal{L}_{BS}(\bar{\sigma})U = -\bar{\sigma}x\frac{\partial}{\partial x}(x^2 P^{(0)}_{xx}), \qquad U(T,x) = 0,$$

but $U(t,B) \neq 0$ in general.

2.3. Interpretation of the Greeks. In the case of a regular option without a barrier boundary condition, the correction to the price is given by

$$\widetilde{P^{(1)}} = -\frac{V_3^\varepsilon}{\bar{\sigma}} x P^{(0)}_{x\bar{\sigma}} - \frac{V_2^\varepsilon}{\bar{\sigma}} P^{(0)}_{\bar{\sigma}},$$

which corresponds to the alternative formulation

$$\widetilde{P^{(1)}} = -(T-t)\left(V_3^\varepsilon x\frac{\partial}{\partial x}\left(x^2\frac{\partial^2}{\partial x^2}\right) + V_2^\varepsilon x^2\frac{\partial^2}{\partial x^2}\right)P^{(0)}$$

given in [13] because

$$\mathcal{V} = \bar{\sigma}(T-t)x^2 P^{(0)}_{xx}, \tag{2.15}$$

$$x P^{(0)}_{x\bar{\sigma}} = \bar{\sigma}(T-t)x\frac{\partial}{\partial x}(x^2 P^{(0)}_{xx}). \tag{2.16}$$

While it is intuitive to present the asymptotic correction in terms of the so-called Greeks $P^{(0)}_{\bar\sigma}$ and $P^{(0)}_{x\bar\sigma}$, intuition can be misleading because, here, these terms are evaluated at the long-run mean volatility $\bar\sigma$, and not at (an estimate of) the current volatility level $f(Y_t)$. In other words, these terms represent sensitivity to the global mean volatility rather than local sensitivity, which is how the Greeks are usually employed in practice. The asymptotic calculation has highlighted the Vega and Vega of the Delta $P^{(0)}_{x\bar\sigma}$ as primary measures of the effect of stochastic volatility on pricing in the fast mean-reversion limit, but the current volatility level is unimportant to this order. It is analogous to a central limit theorem correction to a law of large numbers.

In the case of path-dependent options considered here, these Greek terms do not comprise the whole correction, and the term $\hat P$, which can be represented as a boundary integral as we shall see below, plays an important role.

2.3.1. Calculation. The problem (2.13) can be transformed to a constant coefficient backward heat equation by the simple transformations $\eta = \log(x/B)$ and

$$\hat P(t,x) = \frac{V_3^\varepsilon}{\bar\sigma} v(t,\eta) \exp\left(-\frac{1}{8}\bar\sigma^2(1+k)^2(T-t) + \frac{1}{2}(1-k)\eta\right).$$

Then $v(t,\eta)$ solves

$$\begin{aligned} v_t + \frac{1}{2}\bar\sigma^2 v_{\eta\eta} &= 0 \quad \text{in } \eta > 0 \text{ and } t < T,\\ v(T,\eta) &= 0,\\ v(t,0) &= \tilde g(t), \end{aligned} \tag{2.17}$$

where

$$\tilde g(t) = e^{\frac{1}{8}\bar\sigma^2(1+k)^2(T-t)} B^{-\frac{1}{2}(1-k)} g(t).$$

By Duhamel's theorem (see [3, page 31], for instance), the solution is given by the one-dimensional integral

$$v(t,\eta) = \frac{1}{\bar\sigma\sqrt{2\pi}} \int_t^T \frac{\eta}{(s-t)^{3/2}} e^{-\eta^2/2\bar\sigma^2(s-t)} \tilde g(s)\,\mathrm{d}s. \tag{2.18}$$

We obtain the correction to the barrier price as

$$\begin{aligned} \widetilde{P^{(1)}}(t,x) = &-\frac{V_3^\varepsilon}{\bar\sigma} x P^{(0)}_{x\bar\sigma}(t,x) - \frac{V_2^\varepsilon}{\bar\sigma} P^{(0)}_{\bar\sigma}(t,x)\\ &+\frac{V_3^\varepsilon}{\bar\sigma}\frac{x}{B}\log\left(\frac{x}{B}\right)\frac{1}{\bar\sigma\sqrt{2\pi}}\int_t^T e^{-\frac{1}{2}d_B(s-t)^2}\frac{g(s)}{(s-t)^{3/2}}\,\mathrm{d}s, \end{aligned} \tag{2.19}$$

where

$$d_B(\tau) = \frac{\log(x/B)}{\bar\sigma\sqrt{\tau}} + \frac{1}{2}(1+k)\bar\sigma\sqrt{\tau}.$$

Explicit formulas for $g(t)$ and $\widetilde{P^{(1)}}(t,x)$ are given in Appendix A. These are illustrated in Figures 2.1 and 2.2. As depicted in Figure 2.1, the effect of changing the slope of the

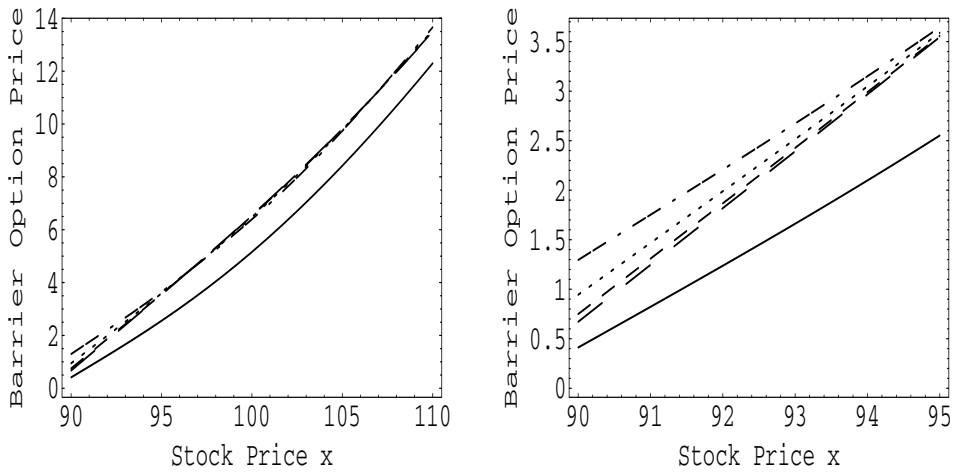


FIG. 2.1. *Effect of changing the slope of the skew a on down-and-out call option price. The parameters used for pricing the contract are* $K = 100$, $B = 89$, $T = 0.5$, $\bar{\sigma} = 0.17$, $b = 0.23$. *As shown more closely in the right figure, near the barrier, making a more negative increases the price. This effect reverses at higher stock prices. In the figures, the solid line shows the corresponding Black–Scholes price. In the right figure, the values of a reading upwards after the Black–Scholes pricing curve are* $a = -0.02, -0.04, -0.09, -0.18$.

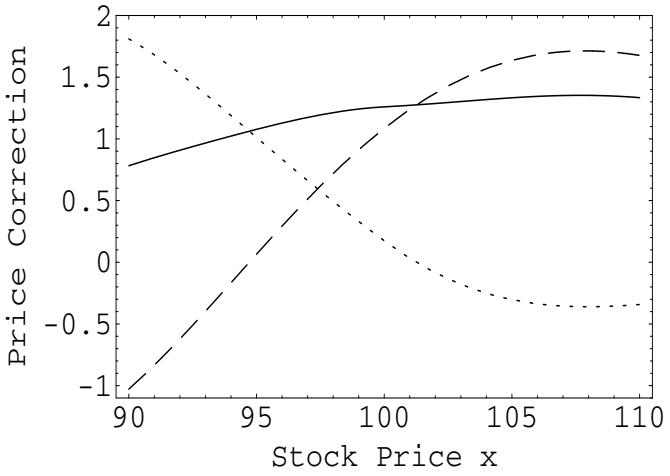


FIG. 2.2. *The first-order correction for down-and-out call option at time* $t = 0$. *The parameters used for pricing the contract are as in Figure* 2.1 *with* $a = -0.154$. *The solid line shows the first-order correction* $\widetilde{P^{(1)}}$, *the dotted line is* $\hat{P}$, *and the dashed line is the contribution of the Greek terms to* $\widetilde{P^{(1)}}$ *in* (2.19).

implied volatility curve has a mixed structure. For the stock prices that are further away from the strike price, where there is a small possibility of having a positive payoff, making a more negative increases the price. For the stock prices that are around the strike price, this effect is reversed. An increase in b, which can also be interpreted as using a higher volatility as we push up the implied volatility curve, increases the corrected price.

2.4. Convergence. In the case of a smooth payoff-at-maturity function, the proof of the convergence result

$$|P(t,x,y) - (P^{(0)}(t,x) + \widetilde{P^{(1)}}(t,x))| = \mathcal{O}(\varepsilon)$$

at a fixed point (t,x,y) is obtained by an adaptation of the proof given in [13, section 5.4]. Here the sense of convergence is with $t < T$, $x > B$, and $y \in \mathbb{R}$ fixed and as $\varepsilon \downarrow 0$, as is needed in the finance application.

The error $Z^\varepsilon(t,x,y)$ defined by

$$P = P^{(0)} + \sqrt{\varepsilon}P^{(1)} + \varepsilon P^{(2)} + \varepsilon^{3/2}P^{(3)} - Z^\varepsilon$$

satisfies

$$\begin{aligned}
\mathcal{L}^\varepsilon Z^\varepsilon &= \varepsilon(\mathcal{L}_1 P^{(3)} + \mathcal{L}_2 P^{(2)}) + \varepsilon^{3/2}\mathcal{L}_2 P^{(3)},\\
Z^\varepsilon(T,x,y) &= \varepsilon P^{(2)}(T,x,y) + \varepsilon^{3/2}P^{(3)}(T,x,y),\\
Z^\varepsilon(t,B,y) &= \varepsilon P^{(2)}(t,B,y) + \varepsilon^{3/2}P^{(3)}(t,B,y),
\end{aligned}$$

using the definitions of $P^{(0)}(t,x)$, $P^{(1)}(t,x)$ and choosing $P^{(2)}(t,x,y)$ and $P^{(3)}(t,x,y)$ as solutions of

$$\mathcal{L}_0 P^{(2)} + \mathcal{L}_2 P^{(0)} = 0, \tag{2.20}$$

$$\mathcal{L}_0 P^{(3)} = -(\mathcal{L}_1 P^{(2)} + \mathcal{L}_2 P^{(1)}), \tag{2.21}$$

respectively. The latter can be chosen to be at most logarithmically growing in y by the properties of the Poisson equations (2.20) and (2.21) and the assumed boundedness of f and Λ. The result follows from the maximum principle because smoothness of the payoff implies $P^{(2)}$ and $P^{(3)}$ are smooth with bounded derivatives.

When the payoff is only continuous as in the case of the barrier call option here, the argument of [17] can be adapted to show that

$$|P(t,x,y) - (P^{(0)}(t,x) + \widetilde{P^{(1)}}(t,x))| = \mathcal{O}(\varepsilon^{1-p})$$

for any $p > 0$. This involves a regularization of the payoff, which can be conveniently done by replacing the nonsmooth call payoff $(x-K)^+$ by the Black–Scholes barrier option price $P^{(0)}(T-\delta, x; \bar{\sigma})$ a small time $\delta > 0$ from maturity. This payoff is smooth and zero at the barrier $x = B$, and we can utilize the explicit Black–Scholes barrier option pricing formula (2.9) to easily estimate the blow-up rates of derivatives at $x = K$ as $\delta \downarrow 0$ and $t \to T$.

The important point is that the barrier price $P^{(0)}$ is smooth in $x > B$ and its derivatives have finite limits as $x \to B^+$. Therefore the presence of the knock-out barrier introduces no further complications.

The only further adaptation to the proof in [17] that needs to be made is in showing that the solution of the regularized problem converges to the solution of the unregularized problem as $\delta \downarrow 0$ at a rate independent of ε. This can be achieved by a rotation of coordinates so that the two solutions can be written as expectations of functionals of independent processes (ξ, Y), where $\xi = X - F(Y)$ and $F' = \frac{\sqrt{\varepsilon}\rho}{\nu\sqrt{2}}f$, stopped on a curved boundary. (Such a transformation is not computationally convenient but is useful to derive regularity properties.) The result follows by conditioning on the subordinating process Y and ε-independent moments of this process.

2.5. Hedging. As mentioned in section 1.4, one possible hedging strategy is to hold

$$\Delta_t^\varepsilon := \frac{\partial}{\partial x}(P^{(0)} + \widetilde{P^{(1)}})(t, X_t)$$

stocks at time t, and $e^{-rt}(P^{(0)} + \widetilde{P^{(1)}} - \Delta_t^\varepsilon X_t)$ units of account in the bank. This is not a self-financing strategy, but the value of the hedging portfolio remains close to the price of the barrier option under the assumption of fast volatility mean-reversion, as shown in [13, Chapter 7].

It is well known that certain popular hedging strategies based on the Greeks Delta, Vega, and Gamma run into trouble when the stock price is close to the barrier because these derivatives do not converge to zero as $x \downarrow B$. The asymptotic methods we consider here are not intended to remedy this problem, and indeed we would expect the approximations to be less effective close to the barrier because the assumption of the remaining lifetime of the option being long compared with the characteristic half-life of the volatility breaks down. This is similar to the situation when the option is close to maturity (see [13, section 5.5] for a discussion). In these cases, boundary layer effects are important (see [22] for an analysis).

3. Lookback options. *Lookback options* are path-dependent options whose payoff depends on the realized maximum or minimum of the underlying asset price during the life of the option. One example of this class of options is the *floating strike lookback put*, which pays the difference of the realized maximum of the underlying asset during the option's life and the asset price itself at the expiration time T. Its payoff is $J_T - X_T$, where we define the running maximum $J_t = \max_{0 \le s \le t} X_s$. Pricing equations for lookback options in the Black–Scholes constant volatility model were first given and solved in [18]. A combination of a lookback call (paying the difference between the terminal stock price and the minimum) and a lookback put can be used to model trading strategies employed by many trend-following hedge funds, as discussed in [6], for example. Davydov and Linetsky [7] derive the closed-form solutions for Laplace transforms of the prices of lookback options in the case of the constant elasticity of variance (CEV) model. This is a complete market model in which the volatility process is a power function of the stock price: $\sigma_t = \kappa S_t^\gamma$. Linetsky [27] studies the construction of spectral expansions for the same model.

In a stochastic volatility environment, the price $P(t, x, J, y)$ of this option satisfies

$$\mathcal{L}^\varepsilon P = 0 \quad \text{in } x < J \text{ and } t < T,$$

with a terminal condition $P(T, x, J, y) = J - x$ and a boundary condition at $x = J$, $P_J(t, J, J, y) = 0$. The derivation of the boundary condition is given in [18], expressing the fact that the price of the lookback option for $X_t = J_t$ is insensitive to small changes in J_t because the realized maximum at time T is larger than the realized maximum at time t with probability one.

The problem of finding $P(t, x, J, y)$ can be reduced to a two (space)-dimensional boundary value problem with the following similarity reduction:

$$\xi = x/J \qquad \text{and} \qquad P(t, x, J, y) = JQ(t, \xi, y).$$

We can express $Q(t,\xi,y)$ as the solution of

$$\begin{aligned}\left(\frac{1}{\varepsilon}\mathcal{L}_0+\frac{1}{\sqrt{\varepsilon}}\mathcal{L}_1+\mathcal{L}_2\right)Q=0 \quad &\text{for } \xi<1 \text{ and } t<T,\\ Q(T,\xi,y)&=1-\xi,\\ (Q_\xi-Q)(t,1,y)&=0,\end{aligned}$$

where, in a slight abuse of notation, we redefine $\mathcal{L}_1$ and $\mathcal{L}_2$ to be the same as (2.6) and (2.7), but with ξ replacing x.

3.1. Asymptotic approximation. Our approximation for the lookback price is

$$Q(t,\xi,y)\approx Q^{(0)}(t,\xi)+\widetilde{Q^{(1)}}(t,\xi),$$

where $P^{(0)}(t,x,J)=JQ^{(0)}(t,x/J)$ is the Black–Scholes price of the option with volatility parameter $\bar\sigma$. That is, following the usual asymptotic analysis such as in [13], $Q^{(0)}$ solves

$$\begin{aligned}\langle\mathcal{L}_2\rangle Q^{(0)}=Q^{(0)}_t+\frac{1}{2}\bar\sigma^2\xi^2Q^{(0)}_{\xi\xi}+r(\xi Q^{(0)}_\xi-Q^{(0)})=0 \quad &\text{in } \xi<1 \text{ and } t<T,\\ Q^{(0)}(T,\xi)&=1-\xi,\\ (Q^{(0)}_\xi-Q^{(0)})(t,1)&=0.\end{aligned}\tag{3.1}$$

The correction term solves

$$\widetilde{Q^{(1)}_t}+\frac{1}{2}\bar\sigma^2\xi^2\widetilde{Q^{(1)}_{\xi\xi}}+r\left(\xi\widetilde{Q^{(1)}_\xi}-\widetilde{Q^{(1)}}\right)=\mathcal{A}Q^{(0)} \quad \text{in } \xi<1 \text{ and } t<T,\tag{3.2}$$

with $\widetilde{Q^{(1)}}(T,\xi)=1-\xi$ and

$$\left(\widetilde{Q^{(1)}}-\widetilde{Q^{(1)}_\xi}\right)(t,1)=0.$$

Here, the operator $\mathcal{A}$ is as in (2.10), but with ξ replacing x.

3.2. Zero-order term. Although the pricing formula $P^{(0)}(t,x,J)$ for a lookback put is well known, we will start by deriving $P^{(0)}(t,x,J)$, as the transformations will also be useful in the derivation of $\widetilde{P^{(1)}}(t,x,J)=J\widetilde{Q^{(1)}}(t,x/J)$.

The PDE problem (3.1) can be transformed to a PDE with constant coefficients by using logarithmic variables. That is, defining

$$\eta=\log\xi, \qquad u^{(0)}(t,\eta)=Q^{(0)}(t,\xi),$$

we find $u^{(0)}(t,\eta)$ to satisfy

$$u^{(0)}_t+\frac{1}{2}\bar\sigma^2u^{(0)}_{\eta\eta}+\left(r-\frac{1}{2}\bar\sigma^2\right)u^{(0)}_\eta-ru^{(0)}=0 \quad \text{in } \eta<0 \text{ and } t<T,\tag{3.3}$$

with the conditions

$$\begin{aligned}u^{(0)}(T,\eta)&=1-e^\eta,\\ (u^{(0)}_\eta-u^{(0)})(t,0)&=0.\end{aligned}$$

We first find $w^{(0)}(t,\eta) = u_\eta^{(0)}(t,\eta) - u^{(0)}(t,\eta)$, which solves the (Dirichlet) boundary value problem

$$w_t^{(0)} + \frac{1}{2}\bar{\sigma}^2 w_{\eta\eta}^{(0)} + \left(r - \frac{1}{2}\bar{\sigma}^2\right) w_\eta^{(0)} - r w^{(0)} = 0 \quad \text{in } \eta < 0 \text{ and } t < T,$$

with the conditions $w^{(0)}(T,\eta) = -1$ and $w^{(0)}(t,0) = 0$. The solution for $w^{(0)}(t,\eta)$ can be found via the method of images

$$w^{(0)}(t,\eta) = e^{-r(T-t)}[e^{(1-k)\eta} N(c_1(T-t)) - N(c_2(T-t))], \tag{3.4}$$

where

$$c_1(\tau) = \frac{\eta}{\bar{\sigma}\sqrt{\tau}} + \frac{1}{2}(1-k)\bar{\sigma}\sqrt{\tau} \qquad \text{and} \qquad c_2(\tau) = \frac{-\eta}{\bar{\sigma}\sqrt{\tau}} + \frac{1}{2}(1-k)\bar{\sigma}\sqrt{\tau}.$$

Restoring all transformations, we get, in the notation of [29],

$$P^{(0)} = -x + x(1+k^{-1})N(d_7) + Je^{-r(T-t)}\left(N(d_5) - k^{-1}\left(\frac{x}{J}\right)^{1-k} N(d_6)\right), \tag{3.5}$$

where

$$d_5 = \frac{\log(J/x) - (r - \frac{1}{2}\bar{\sigma}^2)(T-t)}{\bar{\sigma}\sqrt{T-t}}, \qquad d_6 = \frac{\log(x/J) - (r - \frac{1}{2}\bar{\sigma}^2)(T-t)}{\bar{\sigma}\sqrt{T-t}},$$
$$d_7 = \frac{\log(x/J) + (r + \frac{1}{2}\bar{\sigma}^2)(T-t)}{\bar{\sigma}\sqrt{T-t}}.$$

3.3. First-order correction. Analogous to the zero-order calculation, we define

$$\eta = \log\xi, \qquad u^{(1)}(t,\eta) = \widetilde{Q^{(1)}}(t,\xi),$$

and from (3.2) find $u^{(1)}(t,\eta)$ to satisfy

$$u_t^{(1)} + \frac{1}{2}\bar{\sigma}^2 u_{\eta\eta}^{(1)} + \left(r - \frac{1}{2}\bar{\sigma}^2\right) u_\eta^{(1)} - r u^{(1)} = \tilde{\mathcal{A}} u^{(0)} \quad \text{in } \eta < 0 \text{ and } t < T, \tag{3.6}$$
$$u^{(1)}(T,\eta) = 0,$$
$$(u_\eta^{(1)} - u^{(1)})(t,0) = 0,$$

where

$$\tilde{\mathcal{A}} = V_3^\varepsilon\left(\frac{\partial^3}{\partial\eta^3} - \frac{\partial^2}{\partial\eta^2}\right) + V_2^\varepsilon\left(\frac{\partial^2}{\partial\eta^2} - \frac{\partial}{\partial\eta}\right).$$

Defining $\mathcal{L}_{LB}$ by

$$\mathcal{L}_{LB} = \frac{\partial}{\partial t} + \frac{1}{2}\bar{\sigma}^2\frac{\partial^2}{\partial\eta^2} + \left(r - \frac{1}{2}\bar{\sigma}^2\right)\frac{\partial}{\partial\eta} - r\cdot,$$

we can verify by differentiating (3.3) with respect to $\bar{\sigma}$ that

$$\mathcal{L}_{LB} u_{\bar{\sigma}}^{(0)} = -\bar{\sigma}(u_{\eta\eta}^{(0)} - u_\eta^{(0)}) \qquad \text{and} \qquad \mathcal{L}_{LB} u_{\eta\bar{\sigma}}^{(0)} = -\bar{\sigma}(u_{\eta\eta\eta}^{(0)} - u_{\eta\eta}^{(0)})$$

with $u^{(0)}_{\bar\sigma}(T,\eta) = (u^{(0)}_{\eta\bar\sigma} - u^{(0)}_{\bar\sigma})(t,0) = 0$ and with $u^{(0)}_{\eta\bar\sigma}(T,\eta) = 0$, but $(u^{(0)}_{\eta\eta\bar\sigma} - u^{(0)}_{\eta\bar\sigma})(t,0) \neq 0$ in general.

This motivates us to define $\hat u(t,\eta)$ by

$$\frac{V_3^\varepsilon}{\bar\sigma}\hat u = u^{(1)} + \frac{1}{\bar\sigma}(V_3^\varepsilon u^{(0)}_{\eta\bar\sigma} + V_2^\varepsilon u^{(0)}_{\bar\sigma}).$$

We find that $\hat u$ solves

$$\begin{gathered}\hat u_t + \frac{1}{2}\bar\sigma^2\hat u_{\eta\eta} + \left(r - \frac{1}{2}\bar\sigma^2\right)\hat u_\eta - r\hat u = 0 \quad \text{in } \eta < 0 \text{ and } t < T,\\ \hat u(T,\eta) = 0,\\ (\hat u_\eta - \hat u)(t,0) = g(t),\end{gathered}$$

where we define

$$g(t) = (u^{(0)}_{\eta\eta\bar\sigma} - u^{(0)}_{\eta\bar\sigma})\big|_{\eta=0} = w^{(0)}_{\eta\bar\sigma}\big|_{\eta=0}. \tag{3.7}$$

Defining $\hat w = \hat u_\eta - \hat u$, $\hat w(t,\eta)$ solves the Dirichlet boundary value problem

$$\begin{gathered}\hat w_t + \frac{1}{2}\bar\sigma^2\hat w_{\eta\eta} + \left(r - \frac{1}{2}\bar\sigma^2\right)\hat w_\eta - r\hat w = 0 \quad \text{in } \eta < 0 \text{ and } t < T,\\ \hat w(T,\eta) = 0,\\ \hat w(t,0) = g(t).\end{gathered}$$

Following the analysis leading to (2.18), we can write

$$\hat w(t,\eta) = -\frac{\eta e^\eta}{\bar\sigma\sqrt{2\pi}}\int_t^T e^{-\frac{1}{2}c_3(s-t)^2}\frac{g(s)}{(s-t)^{3/2}}\,\mathrm{d}s, \tag{3.8}$$

where

$$c_3(\tau) = \frac{\eta}{\bar\sigma\sqrt{\tau}} + \frac{1}{2}\bar\sigma(1+k)\sqrt{\tau}.$$

This formula, together with a Taylor expansion of g, yields $\hat w_\eta(t,0) = \frac{1}{2}(1-k)g(t)$.

To recover $\hat u(t,\eta)$, we use

$$\hat u(t,\eta) = \int_0^\eta e^{\eta-z}\hat w(t,z)\,\mathrm{d}z + e^\eta h(t), \tag{3.9}$$

where

$$h'(t) = -\frac{1}{2}\bar\sigma^2\hat w_\eta(t,0) - rg(t), \quad h(T) = 0.$$

Therefore

$$h(t) = \frac{1}{2}\left(r + \frac{1}{2}\bar\sigma^2\right)\int_t^T g(s)\,ds.$$

Restoring the transformations, the first-order correction is given as

$$\widetilde{P^{(1)}}(t,x,J) = -\frac{V_3^\varepsilon}{\bar\sigma}xP^{(0)}_{x\bar\sigma}(t,x) - \frac{V_2^\varepsilon}{\bar\sigma}P^{(0)}_{\bar\sigma}(t,x) + \frac{V_3^\varepsilon}{\bar\sigma}J\hat u(t,\log(x/J)). \tag{3.10}$$

Explicit formulas for the Greeks and g are given in Appendix B.

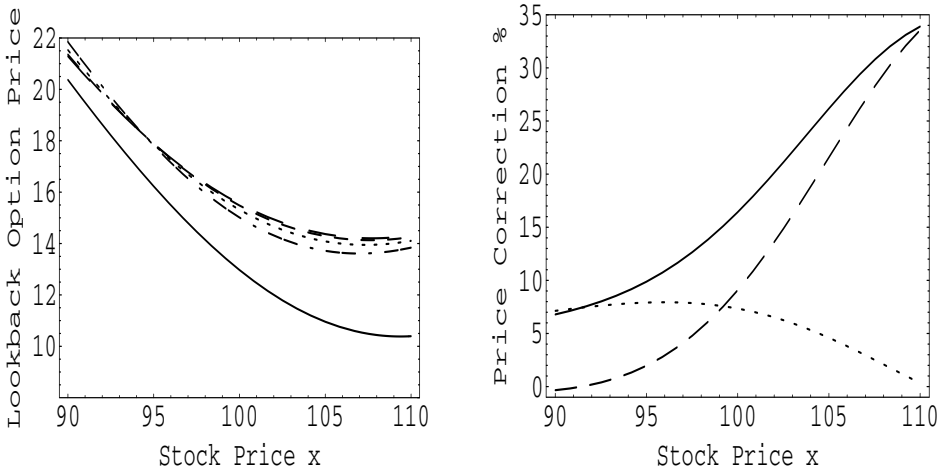


FIG. 3.1. *The left graph shows the effect of changing the slope of the skew a on the lookback put option price. The parameters of the contract are* $T = 0.5$, $\bar{\sigma} = 0.17$, $b = 0.23$. *The current running maximum is* $J = 111$. *The solid line shows the corresponding Black–Scholes price; the values of a reading downward at the right of the graph are* $a = -0.02, -0.04, -0.09, -0.18$. *When the stock price is near to its running maximum, making a more negative decreases the option price. The right graph shows the percentage of first-order correction to the Black–Scholes price for the lookback put option at time* $t = 0$. *The parameters of the contract are as in the left figure with* $a = -0.154$. *The solid line shows the whole first-order correction, the dashed line shows the contribution of the Greek terms in* (3.10), *and the dotted line shows the remainder, i.e., the boundary correction.*

Figure 3.1 illustrates the effect of the two parts of the correction, and for various skew slopes. As in the case of the barrier option, the effect of changing a has a mixed structure. For stock prices that are close to the running maximum, that is, when there is a potential of going above the running maximum, making a more negative decreases the price. For smaller stock prices, the effect is reversed. An increase in the intercept of the implied volatility curve increases the price, as in the barrier option case.

3.4. Convergence. From (3.5), second and higher derivatives of $P^{(0)}$ with respect to x blow up as $t \to T$ and $x \to J$, similar to the Black–Scholes price of an at-the-money European call option. Therefore the proof of a convergence result of the form

$$\left| P(t,x,J,y) - (P^{(0)}(t,x,J) + \widetilde{P^{(1)}}(t,x,J)) \right| = \mathcal{O}(\varepsilon^{1-p})$$

for any $p > 0$ at a fixed point (t,x,J,y) requires the regularization techniques in [17], which are discussed in section 2.4.

3.5. Hedging. In the Black–Scholes model, the hedging strategy for the lookback option is again given by the Delta of the option price (see [29], for example). Following our discussion in section 1.4, we can again consider the hedging strategy consisting of

$$\Delta_t^\varepsilon = \frac{\partial}{\partial x}\left(P^{(0)} + \widetilde{P^{(1)}}\right)(t, X_t, J_t)$$

stocks and the amount $e^{-rt}((P^{(0)} + \widetilde{P^{(1)}})(t, X_t, J_t) - \Delta_t^\varepsilon X_t)$ in the bank account. As in [13, Chapter 7], the portfolio is not self-financing but its value is close to the value

of the lookback option. We do not consider optimal hedges involving other options and loss measurement functions in this paper.

4. Passport options. A passport option allows its holder to trade the stock continuously, starting with initial capital v, and collect his or her profit at the expiration date T, if any, with losses written off. Its price is studied by Hyer, Lipton-Lifschitz, and Pugachevsky [24], where they assumed a log-normal process for the underlying stock. They derive and solve the Hamilton–Jacobi–Bellman equation for the price. Shreve and Vecer [28] used probabilistic techniques to price this option as well as other variants. Henderson and Hobson [19] analyzed passport option pricing under stochastic volatility models where they assume independence of the volatility driving process from the stock price process. They give the price analytically using power series expansion methods for different volatility models.

Let $(q_t)_{0\le t\le T}$ be a possible trading strategy, where q_t is the number of stocks held in the trading account at time t. Additionally, $-1 \le q_t \le 1$ at all times, so the trader is restricted to be at most long or short one stock at any time. Let (V_t) be the value of the holder's portfolio so that

$$dV_t = rV_t\,dt + q_t f(Y_t) X_t\,dW_t^\star, \tag{4.1}$$

written in terms of the risk-neutral Brownian motion $W^\star$ because cash flows are priced under $\mathbb{P}^{\star(\gamma)}$. The payoff of the passport option is simply V_T^+ and so the *no-arbitrage* pricing function $P(t,x,y,v)$ of the contract is given by

$$P(t,x,y,v) = \sup_{|q|\le 1} \mathbb{E}^{\star(\gamma)}\{e^{-r(T-t)}V_T^+ \mid X_t = x,\ Y_t = y,\ V_t = v\}.$$

Assuming P has one continuous derivative in t and is twice continuously differentiable in the spatial variables, P solves the Hamilton–Jacobi–Bellman PDE

$$\frac{\partial P}{\partial t} + \sup_{|q|\le 1} \mathcal{L}_{x,y,v}^{(q)} P = 0,$$

where $\mathcal{L}_{x,y,v}^{(q)}$ is the infinitesimal generator of (X,Y,V), plus the discounting term:

$$\begin{aligned}\mathcal{L}_{x,y,v}^{(q)} = {} & \frac{1}{2} f(y)^2 x^2 \left(\frac{\partial^2}{\partial x^2} + 2q\frac{\partial^2}{\partial x \partial v} + q^2 \frac{\partial^2}{\partial v^2}\right) \\ & + \rho \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}} f(y) x \left(\frac{\partial^2}{\partial x \partial y} + q\frac{\partial^2}{\partial y \partial v}\right) + \frac{1}{2}\frac{\nu^2}{\varepsilon}\frac{\partial^2}{\partial y^2} \\ & + r\left(x\frac{\partial}{\partial x} + v\frac{\partial}{\partial v}\right) + \left(\frac{1}{\varepsilon}(m-y) - \frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\Lambda(y)\right)\frac{\partial}{\partial y} - r\cdot .\end{aligned}$$

The terminal condition is $P(T,x,y,v) = v^+$, and the domain is $t < T$, $x > 0$, $-\infty < y, v < \infty$.

The PDE above can be derived directly (see [24], for instance) by setting up a hedged portfolio of the passport option, another vanilla option, and the underlying stock, and assuming that the holder of the passport option trades the stock optimally.

4.1. Similarity reduction. We first take advantage of a natural homogeneity in the problem. From (2.2) and (4.1), we see that scaling X and V by a common factor, say θ, as in

$$X \mapsto \theta X, \qquad V \mapsto \theta V,$$

does not change those equations. In other words, $P(t,\theta x,y,\theta v)=\theta P(t,x,y,v)$, and so we look for a solution of the form

$$P(t,x,y,v)=xQ(t,\xi,y),\quad \xi=v/x,$$

for some function Q.

Substituting this form gives that $Q(t,\xi,y)$ solves the PDE problem

$$\begin{aligned} Q_t+\sup_{|q|\le 1}\left\{\frac{1}{2}f(y)^2(q-\xi)^2Q_{\xi\xi}+\rho\frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}f(y)(q-\xi)Q_{\xi y}\right\}\\ +\frac{1}{2}\frac{\nu^2}{\varepsilon}Q_{yy}+\left[\frac{1}{\varepsilon}(m-y)-\frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\Lambda(y)+\rho\frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}f(y)\right]Q_y=0,\\ Q(T,\xi,y)=\xi^+ \end{aligned} \tag{4.2}$$

in the domain $t<T$, $-\infty<\xi,y<\infty$. Notice that r has disappeared from the problem because the transformations $Q=P/x$ and $\xi=v/x$ mean that we are using the stock price as our numeraire, which grows at rate r under the risk-neutral measure.

Consider the quadratic (in q) term in (4.2):

$$\frac{1}{2}f(y)^2(q-\xi)^2Q_{\xi\xi}+\rho\frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}f(y)(q-\xi)Q_{\xi y}. \tag{4.3}$$

Assuming that $Q_{\xi\xi}>0$ for $t<T$, the maximum of this quadratic over $q\in[-1,1]$ is at the boundaries: $q^*=\pm1$ at each point in the domain. Therefore $(q^*)^2=1$ and the sup term in (4.2) can be replaced by

$$\frac{1}{2}f(y)^2(1+\xi^2)Q_{\xi\xi}-\rho\frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}f(y)\xi Q_{\xi y}+f(y)^2\left|\xi Q_{\xi\xi}-\frac{\rho\nu\sqrt{2}}{\sqrt{\varepsilon}f(y)}Q_{\xi y}\right|.$$

Let $R(t,\xi,y)$ be the solution to PDE (4.2) with terminal condition $R(T,\xi,y)=|\xi|$. It is then straightforward to verify that the function $\frac{1}{2}(\xi+R(t,\xi,y))$ satisfies both the PDE and the terminal condition in (4.2), so we have

$$Q(t,\xi,y)=\frac{1}{2}(\xi+R(t,\xi,y)). \tag{4.4}$$

The PDE problem for R is therefore

$$\begin{aligned} R_t+\frac{1}{2}f(y)^2(1+\xi^2)R_{\xi\xi}-\rho\frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}f(y)\xi R_{\xi y}+\frac{1}{2}\frac{\nu^2}{\varepsilon}R_{yy}\\ +\left[\frac{1}{\varepsilon}(m-y)-\frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}\Lambda(y)+\rho\frac{\nu\sqrt{2}}{\sqrt{\varepsilon}}f(y)\right]R_y\\ +f(y)^2\left|\xi R_{\xi\xi}-\frac{\rho\nu\sqrt{2}}{\sqrt{\varepsilon}f(y)}R_{\xi y}\right|=0,\\ R(T,\xi,y)=|\xi|. \end{aligned} \tag{4.5}$$

Observe that (4.5) is unchanged by the transformation $\xi\mapsto-\xi$. As a consequence, $R(t,\xi,y)$ is an even function of ξ. This property carries over to the first two terms of our expansion, where we will take advantage of it.

4.2. Asymptotic expansion. We derive the asymptotic expansion for this option pricing function to highlight the differences with the calculation for the previous two cases given in [13]. See [26, 25] for approximations for stochastic control problems from hedging and portfolio optimization.

The expansion is written here for R, but applies, with obvious modifications to the terminal condition, for Q. Under the usual fast mean-reversion scaling, we rewrite the PDE (4.2) as

$$\left(\frac{1}{\varepsilon}\mathcal{L}_0 + \frac{1}{\sqrt{\varepsilon}}\mathcal{L}_1 + \mathcal{L}_2\right) R + \mathrm{NL}^\varepsilon = 0,$$

where $\mathcal{L}_0$, $\mathcal{L}_1$, and $\mathcal{L}_2$ are the linear differential operators

$$\begin{aligned}
\mathcal{L}_0 &= \nu^2 \frac{\partial^2}{\partial y^2} + (m-y)\frac{\partial}{\partial y},\\
\mathcal{L}_1 &= -\sqrt{2}\,\rho\nu f(y)\xi \frac{\partial^2}{\partial \xi \partial y} + \sqrt{2}\,\nu(\rho f(y) - \Lambda(y))\frac{\partial}{\partial y},\\
\mathcal{L}_2 &= \frac{\partial}{\partial t} + \frac{1}{2} f(y)^2 (1+\xi^2)\frac{\partial^2}{\partial \xi^2},
\end{aligned}$$

and NL^ε is the nonlinear term

$$\mathrm{NL}^\varepsilon = f(y)^2 \left| \xi R_{\xi\xi} - \frac{\rho\nu\sqrt{2}}{\sqrt{\varepsilon}\, f(y)} R_{\xi y} \right|.$$

We look for an expansion

$$R = R^{(0)} + \sqrt{\varepsilon} R^{(1)} + \varepsilon R^{(2)} + \cdots,$$

valid for small $\varepsilon > 0$. Inserting the series and comparing terms of $\mathcal{O}(\varepsilon^{-1})$ gives

$$\mathcal{L}_0 R^{(0)} = 0.$$

This is an ODE in y, and from the properties of $\mathcal{L}_0$, the only solutions with reasonable growth at infinity are constants in y. Therefore we take $R^{(0)} = R^{(0)}(t,\xi)$, independent of y. Hence, in the expansion of NL^ε,

$$\mathrm{NL}^\varepsilon = f(y)^2 \left| -\frac{\rho\nu\sqrt{2}}{\sqrt{\varepsilon}\, f(y)} R^{(0)}_{\xi y} + \xi R^{(0)}_{\xi\xi} - \frac{\rho\nu\sqrt{2}}{f(y)} R^{(1)}_{\xi y} + \mathcal{O}(\sqrt{\varepsilon}) \right|,$$

the $\mathcal{O}(\varepsilon^{-1/2})$ disappears.

Comparing terms of $\mathcal{O}(\varepsilon^{-1/2})$ therefore gives

$$\mathcal{L}_0 R^{(1)} + \mathcal{L}_1 R^{(0)} = 0.$$

Since $\mathcal{L}_1$ takes y-derivatives, this reduces to

$$\mathcal{L}_0 R^{(1)} = 0,$$

which implies that $R^{(1)}$ also does not depend on y. Now

$$\mathrm{NL}^\varepsilon = f(y)^2 \left| \xi R^{(0)}_{\xi\xi} + \sqrt{\varepsilon}\left(\xi R^{(1)}_{\xi\xi} - \frac{\rho\nu\sqrt{2}}{f(y)} R^{(2)}_{\xi y}\right) + \mathcal{O}(\varepsilon) \right|, \tag{4.6}$$

including the next order. The $\mathcal{O}(1)$ terms of the expansion in the PDE give

$$\mathcal{L}_0 R^{(2)} + \mathcal{L}_1 R^{(1)} + \mathcal{L}_2 R^{(0)} + f(y)^2 |\xi| R^{(0)}_{\xi\xi} = 0, \tag{4.7}$$

where we have assumed that $R^{(0)}_{\xi\xi} \geq 0$; that is, the leading term inherits the convexity

of R in ξ. The second term $\mathcal{L}_1 R^{(1)} = 0$ because $R^{(1)}$ does not depend on y. We view (4.7) as a Poisson equation for $R^{(2)}$. For there to be a solution, the source term must be centered with respect to the invariant distribution of the OU process (Y_t), namely,

$$\langle \widetilde{\mathcal{L}_2} \rangle R^{(0)} = 0, \tag{4.8}$$

where we define

$$\begin{aligned}\widetilde{\mathcal{L}_2} &= \mathcal{L}_2 + f(y)^2 |\xi| \frac{\partial^2}{\partial \xi^2} \\ &= \frac{\partial}{\partial t} + \frac{1}{2} f(y)^2 (1 + |\xi|)^2 \frac{\partial^2}{\partial \xi^2}.\end{aligned}$$

The averaged operator simply replaces $f(y)^2$ by the constant $\bar{\sigma}^2$. Therefore $R^{(0)}$, when transformed back, gives the passport option pricing function with the constant long-run average volatility $\bar{\sigma}$. To move to the next order, we formally linearize the expression (4.6) for NL^ε as follows:

$$\mathrm{NL}^\varepsilon = f(y)^2 \left(|\xi| R^{(0)}_{\xi\xi} + \sqrt{\varepsilon}\, \mathrm{sgn}(\xi) \left[\xi R^{(1)}_{\xi\xi} - \frac{\rho\nu\sqrt{2}}{f(y)} R^{(2)}_{\xi y} \right] \right) + \mathcal{O}(\varepsilon).$$

Now comparing terms in the expanded PDE of $\mathcal{O}(\sqrt{\varepsilon})$ gives

$$\mathcal{L}_0 R^{(3)} + \mathcal{L}_1 R^{(2)} + \mathcal{L}_2 R^{(1)} + \mathrm{sgn}(\xi) \left[\xi f(y)^2 R^{(1)}_{\xi\xi} - \rho\nu\sqrt{2} f(y) R^{(2)}_{\xi y} \right] = 0.$$

This is a Poisson equation for $R^{(3)}$ whose solvability condition gives

$$\begin{aligned}\langle \widetilde{\mathcal{L}_2} \rangle R^{(1)} &= -\left\langle \left(\mathcal{L}_1 - \rho\nu\sqrt{2}\, f\, \mathrm{sgn}(\xi) \frac{\partial^2}{\partial \xi \partial y} \right) R^{(2)} \right\rangle \\ &= \nu\sqrt{2} \left\langle \left(\rho f\, \mathrm{sgn}(\xi)(1 + |\xi|) \frac{\partial^2}{\partial \xi \partial y} - (\rho f - \Lambda) \frac{\partial}{\partial y} \right) R^{(2)} \right\rangle.\end{aligned} \tag{4.9}$$

As in section 2.2, let $\phi(y)$ be a solution to $\mathcal{L}_0 \phi = f(y)^2 - \bar{\sigma}^2$. Then (4.7) gives

$$R^{(2)} = -\frac{1}{2} \phi(y)(1 + |\xi|)^2 R^{(0)}_{\xi\xi} + D(t, \xi)$$

for some function D that does not depend on y. Substituting and computing the right side of (4.9) gives a combination of second and third derivatives of $R^{(0)}$ in the ξ variable.

As usual, we absorb the $\sqrt{\varepsilon}$ term into the correction and call $\widetilde{R^{(1)}} = \sqrt{\varepsilon} R^{(1)}$. Then $\widetilde{R^{(1)}}(t, \xi)$ solves

$$\begin{aligned}\widetilde{R^{(1)}_t} + \frac{1}{2} \bar{\sigma}^2 (1 + |\xi|)^2 \widetilde{R^{(1)}_{\xi\xi}} = &- (V_3^\varepsilon - V_2^\varepsilon)(1 + |\xi|)^2 R^{(0)}_{\xi\xi} \\ &- V_3^\varepsilon\, \mathrm{sgn}(\xi)(1 + |\xi|)^3 R^{(0)}_{\xi\xi\xi},\end{aligned} \tag{4.10}$$

where V_2^ε and V_3^ε are the market group parameters defined in (2.11) and (2.12). The terminal condition is $\widetilde{R^{(1)}}(T, \xi) = 0$.

4.3. Zero-order term. Again we start by finding the zero-order approximation. We work with $R(t, \xi, y)$ and recover $Q(t, \xi, y)$ using (4.4). Thus $R^{(0)}$ satisfies

$$R^{(0)}_t + \frac{1}{2} \bar{\sigma}^2 (1 + |\xi|)^2 R^{(0)}_{\xi\xi} = 0 \quad \text{in } -\infty < \xi < \infty \text{ and } t < T, \tag{4.11}$$

$$R^{(0)}(T, \xi) = |\xi|. \tag{4.12}$$

It follows that $R^{(0)}(t,\cdot)$ is even at all times, so by the smoothing properties of (4.11) we have $R^{(0)}_\xi(t,0)=0$ for $t<T$. Hence we can solve

$$\begin{aligned} R^{(0)}_t+\frac{1}{2}\bar\sigma^2(1+\xi)^2R^{(0)}_{\xi\xi}&=0 \quad \text{in } \xi>0 \text{ and } t<T,\\ R^{(0)}(T,\xi)&=\xi,\\ R^{(0)}_\xi(t,0)&=0, \end{aligned} \tag{4.13}$$

and obtain the solution in $\xi<0$ by the even extension.

We transform to constant coefficients via

$$\eta=\log(1+\xi), \qquad R^{(0)}(t,\xi)=u^{(0)}(t,\eta).$$

Then $u^{(0)}(t,\eta)$ solves the Neumann boundary value problem

$$\begin{aligned} u^{(0)}_t+\frac{1}{2}\bar\sigma^2(u^{(0)}_{\eta\eta}-u^{(0)}_\eta)&=0 \quad \text{in } \eta>0 \text{ and } t<T,\\ u^{(0)}(T,\eta)&=e^\eta-1,\\ u^{(0)}_\eta(t,0)&=0. \end{aligned} \tag{4.14}$$

Similar to the case of lookback option, we first find the partial derivative $w^{(0)}=u^{(0)}_\eta$, which solves a Dirichlet boundary value problem. Using the method of images, we find $w^{(0)}(t,\eta)$ as

$$w^{(0)}(t,\eta)=e^\eta N(c_5(T-t))-N(c_6(T-t)),$$

where

$$c_5(\tau)=\frac{\eta}{\bar\sigma\sqrt{\tau}}+\frac{1}{2}\bar\sigma\sqrt{\tau} \qquad \text{and} \qquad c_6(\tau)=\frac{-\eta}{\bar\sigma\sqrt{\tau}}+\frac{1}{2}\bar\sigma\sqrt{\tau}. \tag{4.15}$$

Restoring all the transformations gives $P^{(0)}(t,x,v)$ as

$$P^{(0)}=\frac{1}{2}\left[v+xu^{(0)}\left(t,\log\left(1+\frac{|v|}{x}\right)\right)\right],$$

which can be written in the notation of [28] as

$$\begin{aligned} P^{(0)}=\frac{1}{2}\Big[&v+(x+|v|)N(d_+)-xN(d_-)\\ &+x\bar\sigma\sqrt{T-t}\,N'(d_-)-x\bar\sigma\sqrt{T-t}\,d_-N(-d_-)\Big], \end{aligned}$$

where

$$d_\pm=\frac{\log\left(1+\frac{|v|}{x}\right)}{\bar\sigma\sqrt{T-t}}\pm\frac{1}{2}\bar\sigma\sqrt{T-t}.$$

4.4. First-order correction. The first-order correction $\widetilde{R^{(1)}}$ satisfies the PDE

$$\widetilde{R^{(1)}_t}+\frac{1}{2}\bar\sigma^2(1+|\xi|)^2\widetilde{R^{(1)}_{\xi\xi}}=-(V^\varepsilon_3-V^\varepsilon_2)(1+|\xi|)^2R^{(0)}_{\xi\xi}-V^\varepsilon_3(1+|\xi|)^3R^{(0)}_{\xi\xi\xi},$$

with terminal condition $\widetilde{R^{(1)}}(T,\xi)=0$. Thus $\widetilde{R^{(1)}}$ is again an even function of ξ, and we can solve

$$\widetilde{R^{(1)}_t}+\frac{1}{2}\bar\sigma^2(1+\xi)^2\widetilde{R^{(1)}_{\xi\xi}}=-(V^\varepsilon_3-V^\varepsilon_2)(1+\xi)^2R^{(0)}_{\xi\xi}-V^\varepsilon_3(1+\xi)^3R^{(0)}_{\xi\xi\xi},$$

with the terminal condition $\widetilde{R^{(1)}}(T,\xi)=0$ and the boundary condition $\widetilde{R^{(1)}_{\xi}}(t,0)=0$ in $\xi>0$, $t<T$. We will subtract the "particular solution" at a later stage when it becomes easier to identify.

Applying the same set of transformations, namely,

$$\eta=\log(1+\xi),\qquad \widetilde{R^{(1)}}(t,\xi)=u^{(1)}(t,\eta),$$

we get from (4.7) that

$$\begin{aligned} u^{(1)}_t+\frac{1}{2}\bar{\sigma}^2(u^{(1)}_{\eta\eta}-u^{(1)}_{\eta})&=\tilde{\mathcal{A}}u^{(0)}\quad \text{in } \eta>0 \text{ and } t<T,\\ u^{(1)}(T,\eta)&=0,\\ u^{(1)}_{\eta}(t,0)&=0, \end{aligned}\tag{4.16}$$

where

$$\tilde{\mathcal{A}}=-V_3^{\varepsilon}\left(\frac{\partial^3}{\partial\eta^3}-\frac{\partial^2}{\partial\eta^2}\right)+(V_2^{\varepsilon}+V_3^{\varepsilon})\left(\frac{\partial^2}{\partial\eta^2}-\frac{\partial}{\partial\eta}\right).$$

Defining $\mathcal{L}_p$ by

$$\mathcal{L}_p=\frac{\partial}{\partial t}+\frac{1}{2}\bar{\sigma}^2\left(\frac{\partial^2}{\partial\eta^2}-\frac{\partial}{\partial\eta}\right),$$

we can verify by differentiating (4.14) that

$$\mathcal{L}_p u^{(0)}_{\bar{\sigma}}=-\bar{\sigma}(u^{(0)}_{\eta\eta}-u^{(0)}_{\eta})\qquad\text{and}\qquad \mathcal{L}_p u^{(0)}_{\eta\bar{\sigma}}=-\bar{\sigma}(u^{(0)}_{\eta\eta\eta}-u^{(0)}_{\eta\eta}).$$

Moreover, $u^{(0)}_{\bar{\sigma}}=u^{(0)}_{\eta\bar{\sigma}}=u^{(0)}_{\eta\eta\bar{\sigma}}=0$ for $t=T$ and $u^{(0)}_{\eta\bar{\sigma}}(t,0)=0$ but $u^{(0)}_{\eta\eta\bar{\sigma}}(t,0)\neq 0$ in general. This motivates us to define $\hat{u}$ by

$$\frac{V_3^{\varepsilon}}{\bar{\sigma}}\hat{u}=u^{(1)}-\frac{V_3^{\varepsilon}}{\bar{\sigma}}u^{(0)}_{\eta\bar{\sigma}}-\left(\frac{V_2^{\varepsilon}}{\bar{\sigma}}+\frac{V_3^{\varepsilon}}{\bar{\sigma}}\right)u^{(0)}_{\bar{\sigma}}.$$

Further, defining $\hat{w}=\hat{u}_{\eta}$ to reduce to a Dirichlet boundary value problem, we find that $\hat{w}(t,\eta)$ solves

$$\begin{aligned} \hat{w}_t+\frac{1}{2}\bar{\sigma}^2\left(\hat{w}_{\eta\eta}-\hat{w}_{\eta}\right)&=0\quad \text{in } \eta>0 \text{ and } t<T,\\ \hat{w}(T,z)&=0,\\ \hat{w}(t,0)&=g(t), \end{aligned}$$

where

$$g(t)=-u^{(0)}_{\eta\eta\bar{\sigma}}\big|_{\eta=0}=-w^{(0)}_{\eta\bar{\sigma}}\big|_{\eta=0}.\tag{4.17}$$

The rest of the calculation is similar to the analysis in section 3.3. Following the same steps and restoring all transformations, we obtain the first-order correction for the passport option as

$$\begin{aligned} \widetilde{P^{(1)}}(t,x,v)=&-\frac{V_3^{\varepsilon}}{\bar{\sigma}}\left(1+\frac{x}{|v|}\right)xP^{(0)}_{x\bar{\sigma}}-\left(\frac{V_2^{\varepsilon}}{\bar{\sigma}}-\frac{V_3^{\varepsilon}}{\bar{\sigma}}\frac{x}{|v|}\right)P^{(0)}_{\bar{\sigma}}\\ &+\frac{1}{2}\frac{V_3^{\varepsilon}}{\bar{\sigma}}x\hat{u}\left(t,\log\left(1+\frac{|v|}{x}\right)\right). \end{aligned}\tag{4.18}$$

Explicit formulas for the Greeks and g are given in Appendix C.

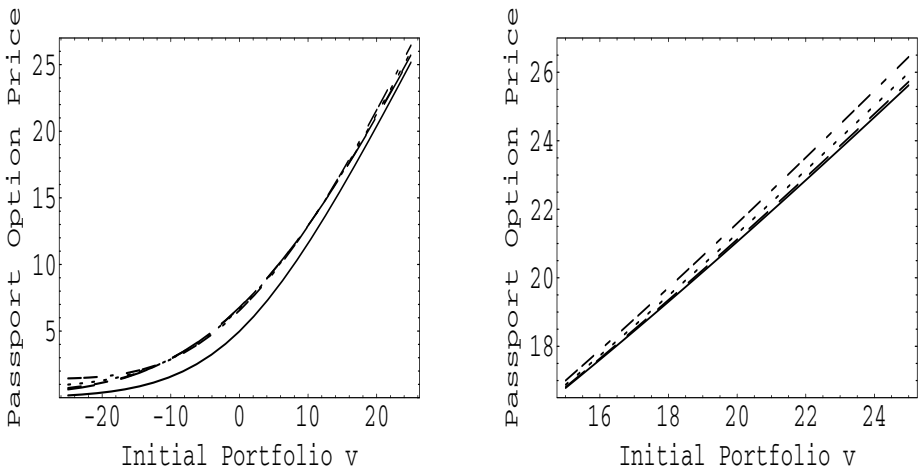


FIG. 4.1. *The left graph illustrates the effect of changing the slope of the skew a on the passport option price. The parameters of the contract are $x = 100$, $T = 0.5$, $\bar{\sigma} = 0.17$, and $b = 0.23$. As $|v|$ gets larger, making a more negative increases the option value, while this effect reverses as $|v|$ gets closer to 0. The right figure shows more closely the upper right corner of the left figure. The solid line shows the corresponding Black–Scholes price; the values of a reading upwards after the Black–Scholes pricing curve are $a = -0.02, -0.04, -0.09, -0.18$.*

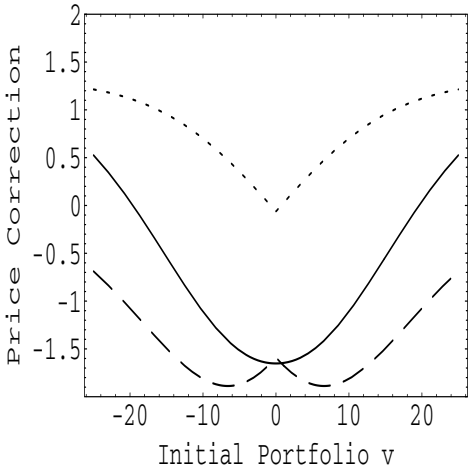


FIG. 4.2. *The first-order correction for the passport option at time $t = 0$. The parameters of the contract are as in Figure* 4.1 *with $a = -0.154$. The solid line shows the full first-order correction, the dashed line shows the contribution of the Greek terms in* (4.18)*, and the dotted line shows the remainder, i.e., the boundary correction.*

Figures 4.1 and 4.2 illustrate the effect of the correction and the slope of the implied volatility skew analyzed as a function of current wealth. As in the previous cases, the effect of changing a has a mixed structure. When the wealth is too large or too small, making a more negative increases the option value. For the wealth level that is close to zero, the reverse is observed. The effect of a higher b is to increase the price, as in the other cases.

4.5. Convergence. The existence of a classical solution to the strongly nonlinear PDE (4.5) is an open question. If such a solution exists so that $R_\xi = 0$ at $\xi = 0$ for $t < T$, then one can write a *nonlinear* error equation for Z^ε defined by

$$R = R^{(0)} + \sqrt{\varepsilon} R^{(1)} + \varepsilon R^{(2)} + \varepsilon^{3/2} R^{(3)} - Z^\varepsilon$$

with suitable terminal and boundary conditions. Under regularity hypotheses on the solution, a *weak* convergence result may be obtained by introducing test functions in (ξ, y). The convergence is necessarily weak in this case because of the sense in which the expansion of the $|\cdot|$ function can be used.

4.6. Hedging. Hedging a passport option differs from the viewpoint of the writer and the holder. The trading strategy of the holder is defined as the maximizer in the definition of the price. This quantity q^* is always plus or minus one, according to the optimizer of the quadratic expression (4.3). The hedging strategy of the seller in the case of a complete market with constant volatility $\bar\sigma$ is $P_x^{(0)} + q^* P_v^{(0)}$, given the strategy of the holder. In the case of a complete market this type of strategy perfectly replicates the payoff if the holder is trading optimally.

As in the barrier and lookback cases, this strategy could be adapted to our corrected price so that the value of the hedging portfolio remains close to the price of the option. Another strategy that also depends only on the calibrated parameters V_2^ε and V_3^ε but focuses on reducing the hedging error, the difference between the option payout, and the hedging portfolio at maturity, is discussed in [13, section 7.2].

Appendix A. Formulas for barrier option correction. Here we include the following explicit formulas for the terms in (2.19). Namely, $g(t)$, defined in (2.14), is given by

$$g(t) = -\left[\frac{2\log\frac{B}{K}}{\bar\sigma^2(T-t)}\mathcal{V}^{BS}(t,B) + \frac{4r}{\bar\sigma^3}C^{BS}(t,B)\right],$$

and the Greeks of the barrier option are given by

$$\begin{aligned}
P_{\bar\sigma}^{(0)}(t,x) &= \mathcal{V}^{BS}(t,x) - \left(\frac{x}{B}\right)^{1-k}\mathcal{V}^{BS}\left(t,\frac{B^2}{x}\right) - \frac{4r}{\bar\sigma^3}\log\frac{x}{B}\left(\frac{x}{B}\right)^{1-k} C^{BS}\left(t,\frac{B^2}{x}\right),\\
P_{x\bar\sigma}^{(0)}(t,x) &= C_{x\bar\sigma}^{BS}(t,x) + \left(\frac{x}{B}\right)^{-k}\left(\frac{k-1}{B}\mathcal{V}^{BS}\left(t,\frac{B^2}{x}\right) + \frac{B}{x}C_{x\bar\sigma}^{BS}\left(t,\frac{B^2}{x}\right)\right)\\
&\quad -\frac{4r}{\bar\sigma^3}\left(\frac{x}{B}\right)^{-k}\frac{1}{B}\left(1-k\log\frac{x}{B}\right)C^{BS}\left(t,\frac{B^2}{x}\right)\\
&\quad +\frac{4r}{\bar\sigma^3}\left(\frac{x}{B}\right)^{-(k+1)}\log\frac{x}{B}\Delta^{BS}\left(t,\frac{B^2}{x}\right),
\end{aligned}$$

where $\Delta^{BS}(t,x)$, $\mathcal{V}^{BS}(t,x)$, $C_{x\bar\sigma}^{BS}(t,x)$ are the Greeks of a call option that has the same parameters:

$$\begin{aligned}
\Delta^{BS}(t,x) &= C_x^{BS}(t,x) = N(d_1),\\
\mathcal{V}^{BS}(t,x) &= C_{\bar\sigma}^{BS}(t,x;\bar\sigma) = xe^{-\frac{1}{2}d_1^2}\frac{\sqrt{T-t}}{\sqrt{2\pi}},\\
C_{x\bar\sigma}^{BS}(t,x) &= \frac{\mathcal{V}^{BS}(t,x)}{x}\left(1-\frac{d_1}{\bar\sigma\sqrt{T-t}}\right).
\end{aligned}$$

Appendix B. Formulas for lookback option correction. The formula for $g(t)$ in the case of the lookback option, defined in (3.7), is

$$g(t) = \frac{2k}{\bar{\sigma}} e^{-r(T-t)} N\left(\frac{1}{2}(1-k)\bar{\sigma}\sqrt{T-t}\right) - 2\frac{e^{-\frac{1}{8}(1+k)^2\bar{\sigma}^2(T-t)}}{\bar{\sigma}^2\sqrt{2\pi(T-t)}}.$$

And the Greek terms for this option referred to in (3.10) are

$$P^{(0)}_{\bar{\sigma}}(t,x) = -Je^{-r(T-t)}\left(\frac{x}{J}\right)^{1-k}\left(\frac{\bar{\sigma}}{r} + \frac{2}{\bar{\sigma}}\log\frac{x}{J}\right)N(d_6(T-t)) + \frac{\bar{\sigma}}{r}xN(d_7(T-t)),$$

$$\begin{aligned} P^{(0)}_{x\bar{\sigma}}(t,x) = {} & \left(\frac{4r}{\bar{\sigma}^3}\log\frac{x}{J} - \frac{\bar{\sigma}}{r} - \frac{2}{\bar{\sigma}}\log\frac{x}{J}\right)\frac{J}{x}N(d_6(T-t)) \\ & - \frac{2\log\frac{x}{J}}{\bar{\sigma}^2\sqrt{T-t}}\frac{J}{x}N'(d_6(T-t)) + \frac{\bar{\sigma}}{r}\frac{J}{x}N(d_7(T-t)). \end{aligned}$$

Appendix C. Formulas for passport option correction. In the case of the passport option, $g(t)$ defined in (4.17) is simply given by

$$g(t) = 2\frac{e^{-\frac{1}{8}\bar{\sigma}^2(T-t)}}{\bar{\sigma}^2\sqrt{2\pi(T-t)}}.$$

The Greek terms used in (4.18) are as follows:

$$P^{(0)}_{\bar{\sigma}}(t,x) = x\bar{\sigma}(T-t)N(d_-) + 2x\sqrt{T-t}N'(d_-),$$

$$P^{(0)}_{x\bar{\sigma}}(t,x) = \frac{P^{(0)}_{\bar{\sigma}}(t,x)}{x} - \frac{2|v|\log\left(1+\frac{|v|}{x}\right)}{(x+|v|)\bar{\sigma}^2\sqrt{T-t}}N'(d_-).$$

REFERENCES

[1] S. Alizadeh, M. Brandt, and F. Diebold, *Range-based estimation of stochastic volatility models*, J. Finance, 57 (2002), pp. 1047–1091.

[2] G. Bakshi, C. Cao, and Z. Chen, *Empirical performance of alternative option pricing models*, J. Finance, 52 (1997), pp. 2003–2049.

[3] H. Carslaw and J. Jaeger, *Conduction of Heat in Solids*, 2nd ed., Oxford Clarendon Press, Oxford, UK, 1959.

[4] M. Chernov, R. Gallant, E. Ghysels, and G. Tauchen, *Alternative models for stock price dynamics*, J. Econometrics, 116 (2003), pp. 225–257.

[5] J. Conlon and M. Sullivan, *Convergence to Black-Scholes for ergodic volatility models*, preprint, Department of Mathematics, University of Michigan, Ann Arbor, MI, 2003.

[6] D. Darius, A. Ilhan, J. Mulvey, K. Simsek, and K. R. Sircar, *Trend-following hedge funds and multi-period asset allocation*, Quant. Finance, 2 (2002), pp. 354–361.

[7] D. Davydov and V. Linetsky, *Pricing and hedging path-dependent options under the CEV process*, Management Sci., 47 (2001), pp. 949–965.

[8] D. Duffie, *Dynamic Asset Pricing Theory*, 3rd ed., Princeton University Press, Princeton, NJ, 2001.

[9] D. Duffie, J. Pan, and K. Singleton, *Transform analysis and option pricing for affine jump-diffusions*, Econometrica, 68 (2000), pp. 1343–1377.

[10] R. Engle and A. Patton, *What good is a volatility model?*, Quant. Finance, 1 (2001), pp. 237–245.

[11] H. Föllmer and P. Leukert, *Efficient hedging: Cost versus shortfall risk*, Finance Stoch., 4 (2000), pp. 117–146.

[12] J.-P. Fouque and C.-H. Han, *Pricing Asian options with stochastic volatility*, Quant. Finance, 3 (2003), pp. 353–362.

[13] J.-P. Fouque, G. Papanicolaou, and K. R. Sircar, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, London, 2000.
[14] J.-P. Fouque, G. Papanicolaou, K. R. Sircar, and K. Solna, *Maturity cycles in implied volatility*, Finance Stoch., to appear.
[15] J.-P. Fouque, G. Papanicolaou, R. Sircar, and K. Solna, *Multiscale stochastic volatility asymptotics*, Multiscale Model. Simul., 2 (2003), pp. 22–42.
[16] J.-P. Fouque, G. Papanicolaou, K. R. Sircar, and K. Solna, *Short time-scale in S&P* 500 *volatility*, J. Computational Finance, 6 (2003), pp. 1–23.
[17] J.-P. Fouque, G. Papanicolaou, R. Sircar, and K. Solna, *Singular perturbations in option pricing*, SIAM J. Appl. Math., 63 (2003), pp. 1648–1665.
[18] M. Goldman, H. Sozin, and M. Gatto, *Path dependent options*: *Buy at the low, sell at the high*, J. Finance, 34 (1979), pp. 1111–1128.
[19] V. Henderson and D. Hobson, *Passport options with stochastic volatility*, Applied Math. Finance, 8 (2001), pp. 97–118.
[20] S. Heston, *A closed-form solution for options with stochastic volatility with applications to bond and currency options*, Review of Financial Studies, 6 (1993), pp. 327–343.
[21] S. Howison and D. Lamper, *Trading volume in models of financial derivatives*, Applied Math. Finance, 8 (2001), pp. 119–135.
[22] S. Howison, A. Rafailidis, and H. Rasmussen, *On the Pricing and Hedging of Volatility Derivatives*, preprint, Mathematical Institute, University of Oxford, Oxford, UK, 2003.
[23] J. Hull and A. White, *The pricing of options on assets with stochastic volatilities*, J. Finance, 42 (1987), pp. 281–300.
[24] T. Hyer, A. Lipton-Lifschitz, and D. Pugachevsky, *Passport to success*, Risk, 10 (1997), pp. 127–131.
[25] M. Jonsson and K. R. Sircar, *Optimal investment problems and volatility homogenization approximations*, in Modern Methods in Scientific Computing and Applications, NATO Sci. Ser. II Math. Phys. Chem. 75, A. Bourlioux, M. Gander, and G. Sabidussi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 255–281.
[26] M. Jonsson and K. R. Sircar, *Partial hedging in a stochastic volatility environment*, Math. Finance, 12 (2002), pp. 375–409.
[27] V. Linetsky, *Lookback Options and Diffusion Hitting Times*: *A Spectral Expansion Approach*, technical report, IEMS Department, Northwestern University, Evanston, IL, 2002.
[28] S. E. Shreve and J. Vecer, *Options on a traded account*: *Vacation calls, vacation puts, and passport options*, Finance Stoch., 4 (2000), pp. 225–274.
[29] P. Wilmott, S. Howison, and J. Dewynne, *Mathematics of Financial Derivatives*: *A Student Introduction*, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1294–1310

MODELLING ANNULAR MICROMIXERS*

JAMES P. GLEESON†, OLIVIA M. ROCHE†, JONATHAN WEST‡, AND ANNE GELB§

Abstract. Magnetohydrodynamic mixing of two fluids in an annular microchannel is modelled as a two-dimensional laminar convection-diffusion problem and examined using asymptotic analysis and numerical simulation. The time T required for mixing of a plug of solute depends on the Péclet number Pe and on the geometry of the annulus. Three scaling regimes are identified: purely diffusive, Taylor-dispersive, and convection-dominated; each has a characteristic power-law dependence of T upon Pe. Consequences of these results for optimal micromixer design are discussed.

Key words. laminar mixing, convection-diffusion, asymptotic analysis, microfluidics

AMS subject classifications. 76M45, 76R99

DOI. 10.1137/S0036139902420407

1. Introduction. Recent advances in microfluidic and lab-on-a-chip technology have led to increased interest in laminar mixing of fluids [1, 2, 3]. Efficient mixing is vital for chemical reactions, but turbulence is absent at the low Reynolds numbers common in microscale devices, and molecular diffusion mixes on an unacceptably slow timescale. In this paper we discuss the mathematical modelling of an annular magnetohydrodynamic (MHD) micromixer, prototypes of which are under development at the Irish National Microelectronics Research Centre [4]. The device consists of an annular channel (see Figures 1 and 2), with inner and outer walls acting as electrodes and with an electromagnet underneath, which provides a vertical magnetic field. A radial electric field is imposed by applying a potential difference across the inner and outer electrodes, and the electric and magnetic fields produce an azimuthal Lorentz force, which acts as a pumping mechanism for the fluid [5].

The idealized mixing action of this device is illustrated in Figure 1: in the absence of molecular diffusion, the initially separated fluids are convected through each other, increasing the interfacial length between them (linearly in time), and so promoting the mixing action of diffusion. In reality, the actions of convection and diffusion are felt simultaneously, and the goal of this paper is to examine their effect upon the efficiency of the mixer. The discussion here is limited to two dimensions, where the idealized limit of infinite depth has been taken. Previous studies of MHD pumping in an annulus have been motivated by liquid-metal flows and their stability [5, 6], but to our knowledge this is the first investigation of the mixing effects in an annular geometry.

2. Notation and equations. The geometry of the annulus is shown in Figure 2. The radius of the center-line is R, and ρ represents the half-width of the channel; thus the inner wall is located at $r = R - \rho$, and the outer wall at $r = R + \rho$. We describe the geometry using the nondimensional parameter

*Received by the editors December 20, 2002; accepted for publication (in revised form) October 1, 2003; published electronically May 5, 2004. This work was funded by Science Foundation Ireland under Investigator Award 02/IN.1/IM062, the Enterprise Ireland International Collaboration Fund, the National Microelectronics Research Centre, and the Faculty of Arts Research Fund, University College Cork.

http://www.siam.org/journals/siap/64-4/42040.html

†Applied Mathematics, University College Cork, Ireland (j.gleeson@ucc.ie).

‡National Microelectronics Research Centre, Lee Maltings, Cork, Ireland (jonathan.west@newcastle.ac.uk).

§Department of Mathematics, Arizona State University, Tempe, AZ 85287 (ag@math.la.asu.edu).

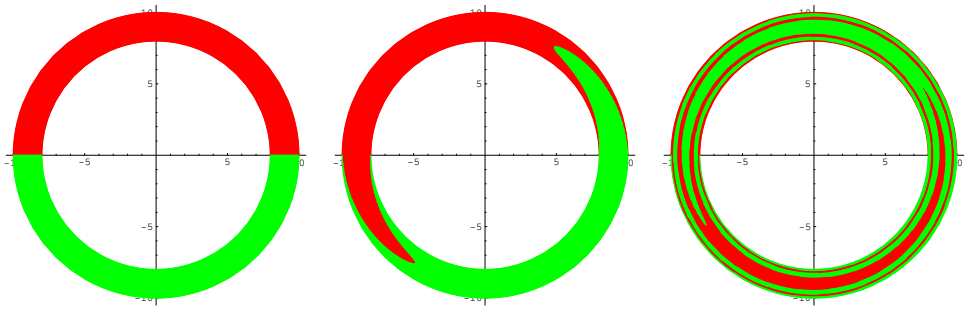


FIG. 1. *Operation of an idealized micromixer at three times, neglecting diffusion.*

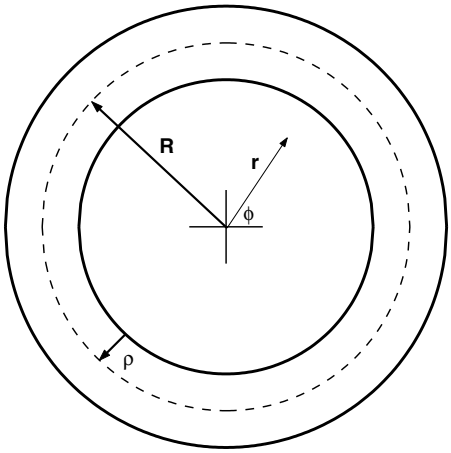


FIG. 2. *Annular geometry, showing the center-line radius R and the channel half-width ρ.*

$$\gamma = \frac{\rho}{R}, \tag{1}$$

which satisfies $0 < \gamma < 1$. Note that the limit $\gamma \to 0$ corresponds to a locally straight channel, and γ approaches 1 as the annulus becomes a punctured disk.

The fluid velocity in the micromixer is found by solving the steady Navier–Stokes equations in the presence of the MHD body force [6]. In the two-dimensional case considered here, this reduces to an ordinary differential equation for the azimuthal velocity $v(r)$, all other velocity components being zero:

$$\frac{d^2v}{dr^2} + \frac{1}{r}\frac{dv}{dr} - \frac{v}{r^2} = -\frac{\alpha}{r}, \tag{2}$$

where α represents the MHD forcing (specifically, $\alpha = -BI/4\pi h\eta$, where I and B are the root-mean-square current and magnetic field strengths, h is the channel depth, and η is the absolute viscosity of the fluid). The solution of (2) satisfying no-slip boundary conditions at the walls $r = R - \rho$ and $r = R + \rho$ is

$$v(r) = \frac{\omega}{8R\rho r}\left[\frac{1}{4} - \frac{(R^2-\rho^2)^2}{16R^2\rho^2}\left(\ln\frac{R-\rho}{R+\rho}\right)^2\right]^{-1} \times \left[(R^2-\rho^2)^2 \ln\frac{R-\rho}{R+\rho} + r^2(R-\rho)^2 \ln\frac{r}{R-\rho} + r^2(R+\rho)^2 \ln\frac{R+\rho}{r}\right], \tag{3}$$

where the velocity is characterized by an average angular velocity ω, which is related to the MHD forcing parameter α by

$$\omega = \alpha \left[\frac{1}{4} - \frac{(R^2 - \rho^2)^2}{16 R^2 \rho^2} \left(\ln \frac{R - \rho}{R + \rho} \right)^2 \right]. \tag{4}$$

The profile (3) reduces to the parabolic Poiseuille profile for a straight channel when $\gamma \to 0$,

$$v(r) \approx \frac{3\omega R}{2\rho^2}(r - R + \rho)(R + \rho - r), \tag{5}$$

with maximum at $r = R$:

$$v_{\max} = \frac{3}{2}\omega R. \tag{6}$$

The mixing of a plug of solute into a surrounding solvent is governed by the convection-diffusion equation (assuming both fluid phases have similar density, viscosity, etc.)

$$\frac{\partial c}{\partial t} + \nabla \cdot (\mathbf{u} c) - \kappa \nabla^2 c = 0, \tag{7}$$

where $\mathbf{u}(\mathbf{x}, t)$ is the fluid velocity vector, $c(\mathbf{x}, t)$ is the concentration of the solvent, and κ is the molecular diffusion coefficient. In the annular micromixer the convection-diffusion equation may be written in polar coordinates,

$$\frac{\partial c}{\partial t} + \frac{v(r)}{r}\frac{\partial c}{\partial \phi} - \frac{\kappa}{r}\frac{\partial}{\partial r}\left(r \frac{\partial c}{\partial r} \right) - \frac{\kappa}{r^2}\frac{\partial^2 c}{\partial \phi^2} = 0, \tag{8}$$

with no-flux boundary conditions at the walls:

$$\begin{aligned} \frac{\partial c}{\partial r}(R - \rho, \phi, t) &= 0, \\ \frac{\partial c}{\partial r}(R + \rho, \phi, t) &= 0. \end{aligned} \tag{9}$$

In the following sections we examine solutions of (8) using analytical, asymptotic, and numerical methods.

The dimensionless parameter used to compare the importance of convection and diffusion in (7) is the Péclet number, which we define for our system as

$$Pe = \frac{\omega R \rho}{\kappa}. \tag{10}$$

Note that ωR is the characteristic linear velocity, while ρ is the smallest linear dimension in the system. Two natural groupings of parameters to give a dimensional time will be used in what follows: the diffusion time R^2/κ, which measures the time for azimuthal mixing by diffusion in the absence of convection, and the convection time ω^{-1}, which is representative of the timescale of rotation of the fluid. The choice of appropriate time-scaling depends on whether diffusion or convection is dominant—we initially choose ω^{-1} but will consider the alternative in section 6.

A completely mixed solute-solvent system has the initial concentration $c(\mathbf{x}, 0)$ of solute spread evenly over the whole annulus. We define mixing efficiency using the timescale over which the concentration profile evolves to the uniform state. In order to measure the deviation from uniformity, we first introduce two averaging procedures, each operating on a function $f(r, \phi)$ defined on the annulus. Radial averaging is denoted by an overbar,

$$\overline{f} = \frac{1}{2R\rho} \int_{R-\rho}^{R+\rho} f(r, \phi) r dr, \tag{11}$$

and angle brackets signify angular averaging,

$$\langle f \rangle = \frac{1}{2\pi} \int_0^{2\pi} f(r, \phi) d\phi. \tag{12}$$

Thus the average value of the concentration over the whole annulus is

$$\langle \overline{c} \rangle = \frac{1}{4\pi R\rho} \int_{R-\rho}^{R+\rho} \int_0^{2\pi} c(r, \phi, t) r \, d\phi \, dr. \tag{13}$$

In fact, it is straightforward to show from the convection-diffusion equation that $\langle \overline{c} \rangle$ is a constant—the total amount of solute is not changed over time but is simply redistributed evenly over the annulus.

In the following sections we adopt a simple initial concentration,

$$c(r, \phi, 0) = 1 + \cos \phi, \tag{14}$$

for ease of asymptotic analysis. Later we show that the asymptotic results also apply to other initial conditions, for instance to the condition used in Figure 1:

$$c(r, \phi, 0) = \begin{cases} 1 & \text{if } 0 \le \phi < \pi, \\ 0 & \text{if } \pi \le \phi < 2\pi. \end{cases} \tag{15}$$

A *mixing measure* $m(t)$ is a positive function of time characterizing the deviation of the concentration at time t from its uniformly mixed state $\langle \overline{c} \rangle$. Define $m(t)$ by

$$m(t) = \frac{\left\langle \overline{(c(r, \phi, t) - \langle \overline{c} \rangle)^2} \right\rangle}{\left\langle \overline{(c(r, \phi, 0) - \langle \overline{c} \rangle)^2} \right\rangle}, \tag{16}$$

so that $m(0) = 1$, and $m(t) \to 0$ as $t \to \infty$. The time T_M for $m(t)$ to decay from 1 to a specified value M is called the *mixing time* and is defined by the condition

$$m(T_M) = M. \tag{17}$$

From the point of view of experimentalists and design engineers, it is desirable to have simple formulas relating the mixing time T_M to the Péclet number Pe and the geometry ratio γ of the micromixer. We proceed to obtain asymptotic approximations to the solution of (8), and hence scaling laws for the mixing times T_M.

3. Low Péclet numbers. When the Péclet number (10) is sufficiently small, diffusive effects completely dominate convective motion. The mixing time by diffusion alone may be calculated by solving the diffusion equation

$$\frac{\partial c}{\partial t} - \frac{\kappa}{r}\frac{\partial}{\partial r}\left(r\frac{\partial c}{\partial r}\right) - \frac{\kappa}{r^2}\frac{\partial^2 c}{\partial \phi^2} = 0 \tag{18}$$

in the annulus, i.e., neglecting the convective term in (8). A series solution can be written as

$$c = \langle \bar{c} \rangle + \sum_{n=1}^{\infty}\sum_{j=1}^{\infty} e^{-\kappa\lambda_{nj}t} G_n\left(r\sqrt{\lambda_{nj}}\right)\left[\alpha_{nj}\cos(n\phi) + \beta_{nj}\sin(n\phi)\right], \tag{19}$$

where the α_{nj} and β_{nj} are constants (determined from the initial condition) and λ_{nm} are eigenvalues determined by the conditions

$$\frac{dG_n}{dr} = 0 \quad \text{at } r = R-\rho \text{ and } r = R+\rho. \tag{20}$$

The eigenfunctions G_n are given in terms of Bessel functions as

$$G_n(r\sqrt{\lambda}) = J_n(r\sqrt{\lambda}) - Y_n(r\sqrt{\lambda})\left[\frac{J_{n-1}\left((R-\rho)\sqrt{\lambda}\right) - J_{n+1}\left((R-\rho)\sqrt{\lambda}\right)}{Y_{n-1}\left((R-\rho)\sqrt{\lambda}\right) - Y_{n+1}\left((R-\rho)\sqrt{\lambda}\right)}\right]. \tag{21}$$

For the single-mode initial condition (14), only the $n = 1$ term of the double sum is present in (19). Consequently, the mixing measure (16) in this case is

$$m(t) = \sum_{j=1}^{\infty} e^{-2\kappa\lambda_{1j}t}\frac{\overline{G_1\left(r\sqrt{\lambda_{1j}}\right)}^2}{\overline{G_1\left(r\sqrt{\lambda_{1j}}\right)}^2}. \tag{22}$$

The form of $m(t)$ at large times t is dominated by the first term ($j = 1$) in this sum. Indeed, in the limit $\gamma \to 0$, we have

$$m(t) \sim \exp(-2\kappa\lambda_{11}t) \quad \text{as } t \to \infty, \tag{23}$$

with the eigenvalue given by

$$\lambda_{11} = \frac{1}{R^2}\left[1 + \frac{1}{3}\gamma^2 + O(\gamma^4)\right] \text{ for } \gamma \ll 1. \tag{24}$$

The asymptotic mixing time T_M as defined in (17) then follows from (23) and (24),

$$T_M \sim \frac{1}{2\kappa\lambda_{11}}\ln\left(\frac{1}{M}\right) \quad \text{as } M \to 0, \tag{25}$$

and so the nondimensional mixing time is

$$\omega T_M \approx \frac{Pe}{2\gamma}\ln\left(\frac{1}{M}\right)\left[1 - \frac{1}{3}\gamma^2 + \cdots\right] \quad \text{for } \gamma \ll 1. \tag{26}$$

4. Intermediate Péclet numbers. In this section we follow and adapt Taylor's arguments [7] (see also [8]) to derive an effective dispersion under certain conditions on the Péclet number, and we clearly identify the limits of validity. We note that Nunge, Lin, and Gill [9] have derived a Taylor dispersion coefficient in a curved channel (matching our (35)), but they do not investigate for which values of Pe the analysis is valid. These bounding values of Pe assume great importance in our later analysis, so we follow Taylor's original formulation to understand when the approximations break down.

Consider a reference frame rotating with angular velocity ω, i.e., with azimuthal angle measured by

$$\phi' = \phi - \omega t. \tag{27}$$

The convection-diffusion equation (8) in this rotating frame is

$$\frac{\partial c}{\partial t} + \left(\frac{v(r)}{r} - \omega\right)\frac{\partial c}{\partial \phi'} - \frac{\kappa}{r}\frac{\partial}{\partial r}\left(r\frac{\partial c}{\partial r}\right) - \frac{\kappa}{r^2}\frac{\partial^2 c}{\partial \phi'^2} = 0, \tag{28}$$

with boundary conditions as before. We will work in this rotating frame for the remainder of this section and thus drop the prime on ϕ. Following Taylor, we assume that the concentration is quasi-steady in the rotating frame, with angular gradient $\frac{\partial c}{\partial \phi}$ independent of r. Moreover, the effect of azimuthal diffusion is neglected compared to radial diffusion. The resulting equation is readily integrated twice to yield

$$c = c(R-\rho,\phi) + \frac{1}{\kappa}\frac{\partial c}{\partial \phi}\int_{R-\rho}^{r}\frac{1}{r_2}\int_{R-\rho}^{r_2}\left[v(r_1) - \omega r_1\right] dr_1\, dr_2. \tag{29}$$

We take the radial mean (as defined in (11)) to allow us to eliminate $c(R-\rho,\phi)$ and obtain

$$c = \overline{c} + \frac{1}{\kappa}\frac{\partial c}{\partial \phi}\left[g(r) - \overline{g}\right], \tag{30}$$

having used the notation

$$g(r) = \int_{R-\rho}^{r}\frac{1}{r_2}\int_{R-\rho}^{r_2}\left[v(r_1) - \omega r_1\right] dr_1\, dr_2 \tag{31}$$

for brevity. The average flux of solvent through the line $\phi =$ constant (in the moving frame) is given by the radial average of the product of the concentration and angular velocity ω' (also relative to the moving frame):

$$\begin{aligned}\overline{J} &= \overline{c\,\omega'} \\ &= \frac{1}{2R\rho}\int_{R-\rho}^{R+\rho}\left[\overline{c} + \frac{1}{\kappa}\frac{\partial c}{\partial \phi}\left(g(r) - \overline{g}\right)\right]\left[\frac{v}{r} - \omega\right] r\, dr \\ &= \frac{1}{\kappa}\frac{\partial c}{\partial \phi}\frac{1}{2R\rho}\int_{R-\rho}^{R+\rho}\left(v(r) - \omega r\right) g(r) dr. \end{aligned} \tag{32}$$

If we assume

$$\frac{\partial c}{\partial \phi} \approx \frac{\partial \overline{c}}{\partial \phi}, \tag{33}$$

this implies that the radial mean concentration in the moving frame is governed by a one-dimensional diffusion equation

$$\frac{\partial \overline{c}}{\partial t} = D \frac{\partial^2 \overline{c}}{\partial \phi^2}, \tag{34}$$

where D is the Taylor dispersion coefficient for the annulus, defined by

$$\begin{aligned} D &= \frac{1}{\kappa} \frac{1}{2R\rho} \int_{R-\rho}^{R+\rho} (v(r) - \omega r)\, g(r)\, dr \\ &= \frac{1}{\kappa} \frac{1}{2R\rho} \int_{R-\rho}^{R+\rho} (v(r) - \omega r) \int_{R-\rho}^{r} \frac{1}{r_2} \int_{R-\rho}^{r_2} [v(r_1) - \omega r_1]\, dr_1\, dr_2\, dr. \end{aligned} \tag{35}$$

This integral may be calculated for the velocity (3); after some algebra we find

$$\begin{aligned} D = {} & \frac{\omega^2 R^2}{24\kappa\gamma} \left[4\gamma^2 - (1-\gamma^2)^2 \ln\left(\frac{1-\gamma}{1+\gamma}\right) \right]^{-2} \\ & \times \left[48\gamma^5(1+\gamma^2) + 120\gamma^4(1-\gamma^2)^2 \ln\frac{1+\gamma}{1-\gamma} - 72\gamma^3(1-\gamma^2)^2(1+\gamma^2) \left(\ln\frac{1+\gamma}{1-\gamma}\right)^2 \right. \\ & \left. - 6\gamma(1-\gamma^2)^4(1+\gamma^2) \left(\ln\frac{1+\gamma}{1-\gamma}\right)^4 + (1-\gamma^2)^4(3+10\gamma^2+3\gamma^4) \left(\ln\frac{1+\gamma}{1-\gamma}\right)^5 \right]. \end{aligned} \tag{36}$$

In the $\gamma \ll 1$ limit, this reduces to

$$D \approx \omega Pe \left[\frac{2}{105}\gamma + \frac{346}{1576}\gamma^3 \cdots \right]. \tag{37}$$

The solution of the one-dimensional diffusion equation (34) with initial condition from (14) is straightforward. Having found $\overline{c}$, we can calculate the concentration c and thus the mixing measure $m(t)$ from (30) and (16), respectively. The mixing measure has an exponential tail,

$$m(t) \sim \exp(-2Dt) \quad \text{as } t \to \infty, \tag{38}$$

and so the mixing time T_M is

$$T_M \sim \frac{1}{2D} \ln\left(\frac{1}{M}\right) \quad \text{as } M \to 0 \tag{39}$$

in the Taylor dispersion regime. From (37), this yields the (nondimensional) asymptotic form

$$\omega T_M \approx \frac{1}{\gamma Pe} \ln\left(\frac{1}{M}\right) \frac{105}{4} \left[1 - \frac{18165}{1576}\gamma^2 + \cdots \right] \text{ for } \gamma \ll 1 \tag{40}$$

when the assumptions made above are valid.

The range of Pe for which the above approximate analysis holds will prove to be of great interest in later sections, so we now examine carefully the two basic assumptions that were made and cast these into conditions on Pe for the result (40) to hold.

The first condition stems from the requirement that the azimuthal diffusion be negligible compared to the radial diffusion; effectively, this requires that the coefficient

of $\frac{\partial^2 c}{\partial \phi^2}$ in (28), actually its radial average, should be much less than the dispersion coefficient (35):

$$\kappa \ll D\overline{r^2}. \tag{41}$$

Consider this condition when $\gamma \ll 1$, using the first term of (37):

$$\begin{aligned} \kappa &\ll \frac{2}{105}\gamma\omega Pe R^2 \\ \iff Pe^2 &\gg \frac{105}{2}. \end{aligned} \tag{42}$$

The resulting condition $Pe \gg 7.2$ is analogous to that in a straight capillary; see, for instance, [8].

The second important approximation is the replacement of $\frac{\partial c}{\partial \phi}$ by $\frac{\partial \overline{c}}{\partial \phi}$ in (33). This is valid if the coefficient of $\frac{\partial c}{\partial \phi}$ in (30) is very small. The order of magnitude of this coefficient is $\overline{g}/\kappa$, and so the condition is

$$\frac{1}{\kappa}\frac{1}{2R\rho}\int_{R-\rho}^{R+\rho} r \int_{R-\rho}^{r} \frac{1}{r_2}\int_{R-\rho}^{r_2} [v(r_1) - \omega r_1]\, dr_1\, dr_2\, dr \ll 1. \tag{43}$$

Again, this yields a simple condition if $\gamma \ll 1$:

$$\begin{aligned} \frac{1}{15}\frac{\omega\rho^2}{\kappa} &\ll 1 \\ \iff Pe &\ll \frac{15}{\gamma}. \end{aligned} \tag{44}$$

Thus (42) and (44) provide rough bounds on Pe for the lower and upper ranges of validity of the Taylor dispersion approximations.

5. Convection-dominated mixing: $Pe \gg 1$. When molecular diffusion is small compared to convective effects, i.e., at large Péclet number, (8) is singularly perturbed. Mixing of a scalar at high Péclet numbers is a topic that has attracted much attention recently, with particular attention paid to the mixing of a passive scalar (as in our case), or vorticity (an active scalar), in the spiral flow field of a vortex [10, 11, 12]. In this section we apply approximation approaches based upon these works to understand the mixing speed in the annular micromixer.

As in previous sections, we examine the asymptotic limit of small γ, i.e., neglecting higher order effects of the channel curvature. Accordingly, (8) is rewritten using the nondimensional variables

$$\begin{aligned} \tilde{r} &= \frac{r-R}{\rho}, \\ \tilde{t} &= \omega t. \end{aligned} \tag{45}$$

Note that the nondimensional radial variable $\tilde{r}$ lies between -1 and 1, with $\tilde{r} = 0$ in the center of the channel. Taking the limit of small γ and using (5), equation (8) reduces to

$$\frac{\partial c}{\partial \tilde{t}} + \frac{3}{2}\left(1 - \tilde{r}^2\right)\frac{\partial c}{\partial \phi} - \epsilon\frac{\partial^2 c}{\partial \tilde{r}^2} = 0, \tag{46}$$

where $\epsilon = (\gamma Pe)^{-1}$ is a small parameter when Pe is sufficiently large.

It is well known [10, 11, 12] that scalar mixing occurs in vortex spiral flows on times scaling as $Pe^{1/3}$. Here we follow [10] to show that a similar scaling arises in the annular mixer, at least for early times. We consider the scalar evolving according to (46), and introduce the notation $\Omega(\tilde{r}) = 3/2(1-\tilde{r}^2)$ for the azimuthal velocity. In the complete absence of molecular diffusion, i.e., if $\epsilon = 0$, the evolution of the nth angular mode from an initial condition of

$$c(\tilde{t}=0) = \exp[in\phi] \tag{47}$$

is given by simple angular convection:

$$c = \exp\left[in\phi - in\Omega(\tilde{r})\tilde{t}\right]. \tag{48}$$

In order to include the effects of nonzero diffusion, we follow [10] and introduce a Lagrangian angular variable:

$$\theta = \phi - \Omega(\tilde{r})\tilde{t}. \tag{49}$$

The derivatives in (46) are then modified as follows:

$$\begin{aligned} \frac{\partial}{\partial \tilde{t}} &\to \frac{\partial}{\partial \tilde{t}} - \Omega\frac{\partial}{\partial\theta}, \\ \frac{\partial}{\partial \phi} &\to \frac{\partial}{\partial\theta}, \\ \frac{\partial}{\partial \tilde{r}} &\to \frac{\partial}{\partial \tilde{r}} - \Omega' \tilde{t}\frac{\partial}{\partial\theta}. \end{aligned} \tag{50}$$

Thus (46) becomes

$$\frac{\partial c}{\partial \tilde{t}} = \epsilon\left(\frac{\partial}{\partial \tilde{r}} - \Omega'\tilde{t}\frac{\partial}{\partial\theta}\right)^2 c, \tag{51}$$

and if the second term in parentheses dominates the first, we obtain

$$\frac{\partial c}{\partial \tilde{t}} = \epsilon\Omega'^2\tilde{t}^2\frac{\partial^2 c}{\partial\theta^2}. \tag{52}$$

The solution of this equation satisfying the initial condition (47) is

$$\begin{aligned} c &= \exp\left[in\theta - \epsilon n^2\Omega'^2\tilde{t}^3/3\right] \\ &= \exp\left[in\phi - in\Omega\tilde{t} - \epsilon n^2\Omega'^2\tilde{t}^3/3\right]. \end{aligned} \tag{53}$$

Noting that $\Omega' = -3\tilde{r}$ and $n = 1$ in our example, the mixing measure $m(t)$ can be calculated from (16) to yield (restoring dimensional variables)

$$m(t) = \frac{\sqrt{\pi}}{2}F\left(\frac{6\omega^3 t^3}{\gamma Pe}\right), \tag{54}$$

where F is defined as the monotonic function

$$F(x) = x^{-1/2}\mathrm{erf}(x^{1/2}). \tag{55}$$

The nondimensional mixing time corresponding to a mixing measure value of M is therefore

$$\omega T_M \approx \left[\frac{\gamma Pe}{6} F^{-1}\left(\frac{2M}{\sqrt{\pi}}\right)\right]^{1/3}, \tag{56}$$

and note in particular that this increases as $Pe^{1/3}$ when M and γ are fixed.

The approximation made to obtain (52) from (51) has limited validity, and so the $Pe^{1/3}$ scaling obtained in (56) is not expected to hold for all times. In particular, we note that $\Omega'(\tilde{r}) = 0$ at the center of the channel, and so (52) is not accurate there. An analogous situation was considered recently [13] for scalar mixing in a vortex: the vanishing azimuthal velocity at the center of the vortex is shown there to render the approximate solution (53) invalid at large times. A different approach is required: we follow [13] and [14] in seeking a solution of (46) of the form

$$c = g(\tilde{t}) \exp\left[in\phi - \frac{3}{2} in\tilde{t} - if(\tilde{t})\tilde{r}^2\right], \tag{57}$$

where the functions $g(\tilde{t})$ and $f(\tilde{t})$ are to be determined, subject to initial conditions $g(0) = 1$, $f(0) = 0$. This form for c can be motivated by noting that if $g \equiv 1$ and $f \equiv 0$, then (57) would be an exact solution of (46) if the azimuthal velocity Ω had no $\tilde{r}$-dependence. In fact, $f(\tilde{t})$ and $g(\tilde{t})$ can be found so that (57) is an exact solution of (46). Substituting (57) into (46) and taking $n = 1$ for the initial condition (14), we find first order equations for f and g:

$$\begin{aligned} \frac{df}{d\tilde{t}} &= -4i\epsilon f^2 - \frac{3}{2}, \\ \frac{dg}{d\tilde{t}} &= -2i\epsilon f g. \end{aligned} \tag{58}$$

Solutions satisfying the initial conditions are given by

$$\begin{aligned} f(\tilde{t}) &= \frac{3(1+i)}{2} \frac{1}{\mu} \tanh\left(\frac{-1+i}{2}\mu\tilde{t}\right), \\ g(\tilde{t}) &= \left[\cosh\left(\frac{-1+i}{2}\mu\tilde{t}\right)\right]^{-\frac{1}{2}}, \end{aligned} \tag{59}$$

where we have written $\mu = 2\sqrt{3\epsilon}$ for clarity. After some manipulation, using (57) in (16) yields a mixing measure

$$m = \sqrt{\frac{\pi}{6}}\mu \left[\sinh(\mu\tilde{t}) - \sin(\mu\tilde{t})\right]^{-\frac{1}{2}} \operatorname{erf}\left[\sqrt{\frac{3}{\mu}}\sqrt{\frac{\sinh(\mu\tilde{t}) - \sin(\mu\tilde{t})}{\cosh(\mu\tilde{t}) + \cos(\mu\tilde{t})}}\right]. \tag{60}$$

When $\mu\tilde{t} \ll 1$, this reduces to

$$\begin{aligned} m &\approx \sqrt{\frac{\pi}{2}} \left(\mu^2\tilde{t}^3\right)^{-\frac{1}{2}} \operatorname{erf}\left[\frac{1}{\sqrt{2}} \left(\mu^2\tilde{t}^3\right)^{\frac{1}{2}}\right] \\ &= \frac{\sqrt{\pi}}{2} F\left(\frac{6\omega^3 t^3}{\gamma Pe}\right), \end{aligned} \tag{61}$$

as in (54). The corresponding scaling for T_M is given by (56) and applies for values of M large enough so that $\mu\tilde{t} \ll 1$.

For $\mu\tilde{t} \gg 1$, the solution (60) decays exponentially in time:

$$m \approx \sqrt{\frac{\pi}{3}\mu}\,\mathrm{erf}\left[\sqrt{\frac{3}{\mu}}\right]\exp\left[-\frac{1}{2}\mu\tilde{t}\right], \tag{62}$$

and for large Péclet numbers $\mu \ll 1$, so the asymptotic mixing time is given (in dimensional variables) by

$$\omega T_M \approx \frac{1}{4}\sqrt{\frac{\gamma Pe}{3}}\ln\left[\frac{4\pi^2}{3M^4}\frac{1}{\gamma Pe}\right]. \tag{63}$$

Note that this mixing time scales as $Pe^{1/2}\ln Pe$ for fixed M and γ. This longer timescale replaces the $Pe^{1/3}$ scaling given by (56) when the mixing time T_M is sufficiently large that the $\mu\tilde{t} \gg 1$ asymptotics are important—this is relevant when we seek a level of mixing M that is sufficiently small. Physically, this new scaling emerges as a consequence of the vanishing of the differential rotation rate at the center of the channel: $\Omega'(0) = 0$; see [13] for a detailed discussion of the analogous problem at the vortex centers. Over short timescales (with $\mu\tilde{t} \ll 1$), the advective stretching and diffusion mix the scalar with time scaling as $Pe^{1/3}$. This convection-enhanced mixing is most effective near the sides of the channel, where the scalar is stretched into thin lamellae; see Figure 1(right-most panel). The more persistent scalar structure at the center of the channel is destroyed only on the longer timescales ($\mu\tilde{t} \gg 1$), leading to the $Pe^{1/2}\ln Pe$ scaling in (63). It is noteworthy that neither (53) nor (57) satisfy the no-flux boundary conditions (9) at the walls of the channel. However, numerical simulations (see section 6) indicate that this does not have an appreciable effect upon the accuracy of the asymptotic formulas. This is a consequence of the fast mixing of the scalar near the walls, so that any error in the boundary conditions is dominated by the slower-mixing scalar structure in the channel center.

6. Numerical simulations. To check the asymptotic results derived in previous sections, and to extend our analysis to cases with initial conditions other than the simple form (14), we solve the convection-diffusion equation (8) numerically. A decomposition into a finite number N of Fourier modes in the azimuthal angle is employed,

$$c(r,\phi,t) = \frac{1}{2}g_0(r,t) + \sum_{n=1}^{N} g_n(r,t)\cos(n\phi) + h_n(r,t)\sin(n\phi), \tag{64}$$

and substitution into (8) yields partial differential equations for g_n and h_n. Having solved the system of equations for the Fourier coefficients, the full concentration field may be constructed from (64), or the mixing measures may be evaluated more directly from (16) by integrating over the angle and normalizing $m(0)$ to unity:

$$m(t) = \overline{\left(\frac{g_0}{2} - \langle\overline{c}\rangle\right)^2 + \frac{1}{2}\sum_{n=1}^{N} g_n^2 + h_n^2}. \tag{65}$$

Logarithmic plots of $m(t)$ as a function of time at various Pe values are shown in Figure 3 for $\gamma = 0.05$. For comparison we plot also the asymptotic forms of $m(t)$, using (23) in the diffusive regime, (38) in the Taylor regime, and (54) and (60) in the convective regime. The asymptotic formulas fit the numerical results well, with

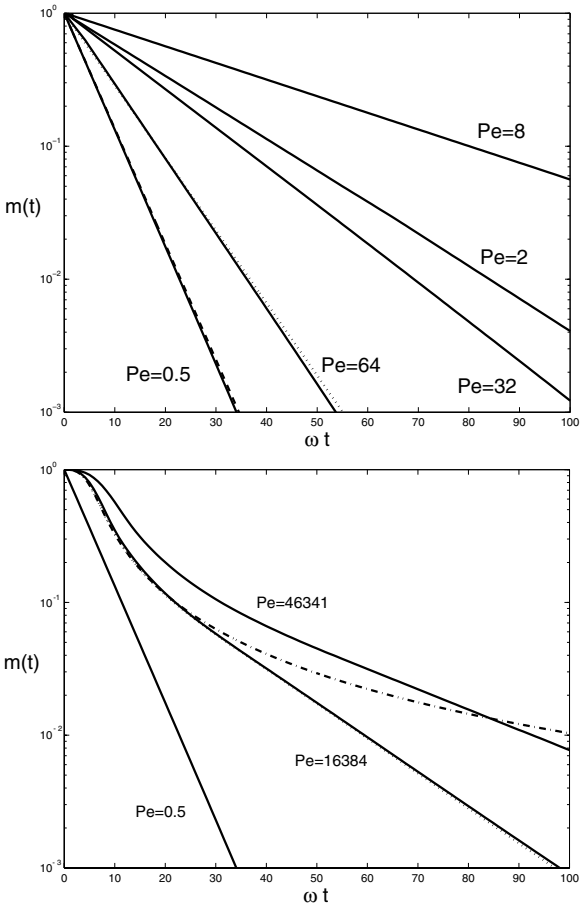


FIG. 3. *Mixing measure $m(t)$ as a function of nondimensional time, calculated in numerical simulations with $\gamma = 0.05$ and various Péclet numbers as shown. Asymptotic predictions are also shown for $Pe = 0.5$ (dashed) using* (23) *and $Pe = 64$ (dotted) using* (38) *in the upper panel. In the lower panel the asymptotic formulas* (54) *(dot-dash) and* (60) *(dotted) are shown for $Pe = 16384$.*

the exception of (54), which, as noted in section 5, is limited to early times. Taking account of the vanishing differential rotation rate at the center of the channel leads to (60), which matches the numerical results very well (the dotted line being almost indistinguishable from the solid).

From the numerical solution for $m(t)$, it is straightforward to calculate the mixing times T_M required for the measure to decay from its initial value of 1 to the value M. We choose three values of M for comparison with the asymptotic predictions—$M = 0.3$, 0.1, and 0.01—and investigate a wide range of Péclet numbers. For fixed values of γ and M, the asymptotic analysis of the previous sections predicts a mixing time proportional to Pe in the diffusion-dominated regime, to Pe^{-1} in the Taylor dispersion regime, and to either $Pe^{1/3}$ or $Pe^{1/2} \ln Pe$ when convection dominates. The numerical values of nondimensional times ωT_M are plotted in Figure 4 along with the straight lines corresponding to the formulas (26), (40), and (56) for all three values of M. Note the excellent agreement with predictions, except for the lowest value of M, when the approximation (54) no longer accurately describes the time evolution of $m(t)$. For this value of M the asymptotic result (63) is plotted with a dotted line and is found

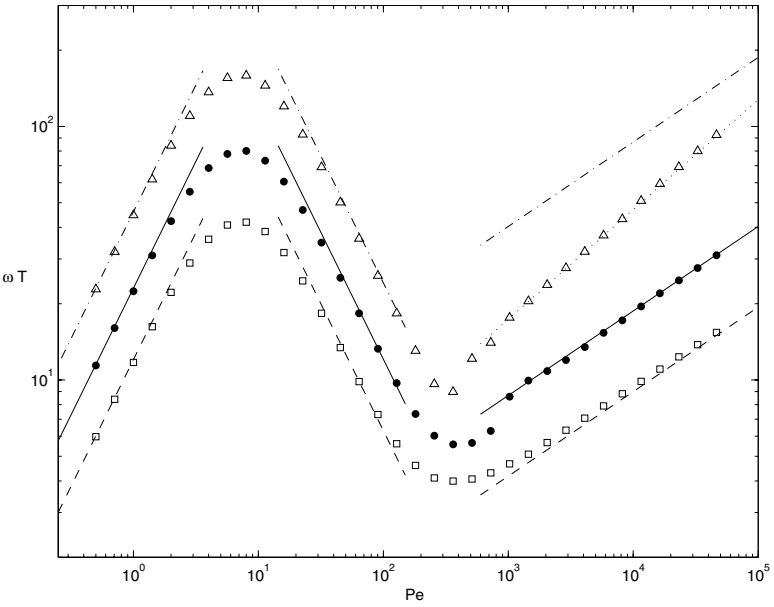


FIG. 4. *Nondimensional mixing times as a function of Péclet number, for* $\gamma = 0.05$. *Asymptotic results are shown as lines, and numerical results as symbols for values of the mixing measure:* $M = 0.3$ *(dashed line; squares),* $M = 0.1$ *(solid line, points), and* $M = 0.01$ *(dot-dash line, triangles). The long-time asymptotic result* (63) *is also plotted as a dotted line for the case of* $M = 0.01$.

to closely match the numerical results.

Because much of our analysis has concentrated on the $\gamma \to 0$ limit, it is worthwhile comparing numerics and asymptotics for a larger value of γ, noting that the maximum possible value of γ is 1. In Figure 5 we plot the results for $\gamma = 0.2$ and find that in general the correspondence between predicted and numerical values is excellent. However, we note that the limits on the Taylor regime now preclude the formation of a clear Pe^{-1} scaling range.

Another simplification in the analysis concerns the initial condition (14), which consists of only a single harmonic of the azimuthal variable. In order to examine the robustness of our predictions, we numerically solve the convection-diffusion equation for initial condition (15), replacing the discontinuous function by hyperbolic tangents to avoid Gibbs oscillations. The decay of the mixing measure is still dominated by the first angular harmonic, and so the mixing times are remarkably close to the asymptotic estimates; see Figure 6. We therefore expect the formulas to be useful for rather general initial distributions of the solute.

The nondimensional time ωT_M used in Figures 3–6 may be replaced by the alternative nondimensionalization mentioned in section 2, i.e., the diffusion time $\kappa T_M / R^2$. The timescales are related by

$$\frac{\kappa T_M}{R^2} = \frac{\gamma}{Pe}\,\omega T_M, \tag{66}$$

and so the data of, for example, Figure 4 is easily recast in terms of the diffusion time; see Figure 7. The relevant mixing times in the diffusive, Taylor, and convective

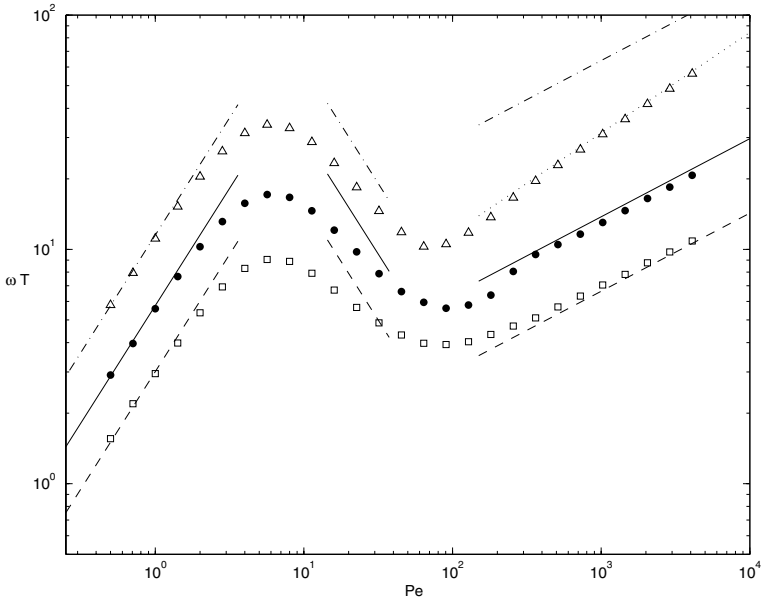


FIG. 5. *Nondimensional mixing times as a function of Péclet number for* $\gamma = 0.2$. *Asymptotic results are shown as lines, and numerical results as symbols for values of the mixing measure:* $M = 0.3$ *(dashed line; squares),* $M = 0.1$ *(solid line, points), and* $M = 0.01$ *(dot-dash line, triangles). The long-time asymptotic result* (63) *is also plotted as a dotted line for the case of* $M = 0.01$.

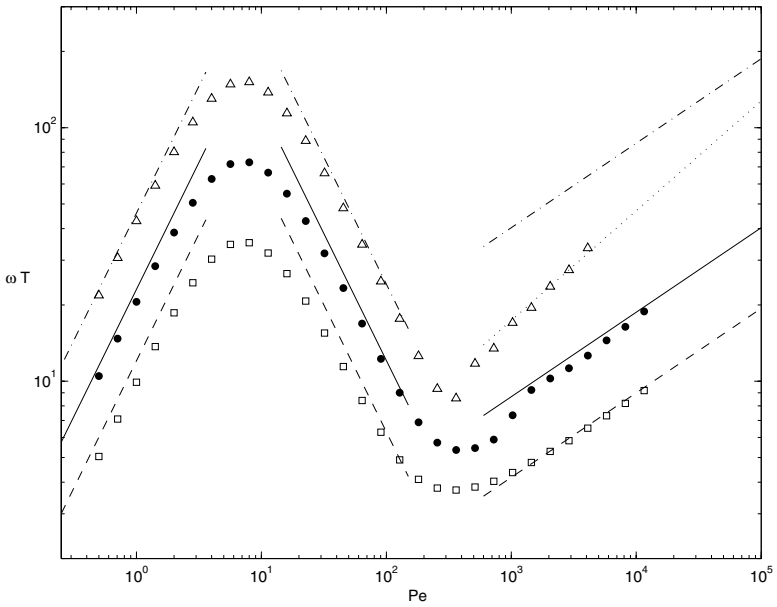


FIG. 6. *Nondimensional mixing times as a function of Péclet number, for* $\gamma = 0.05$ *and an initial concentration given by* (15) *(slightly smoothed). Asymptotic results are shown as lines, and numerical results as symbols for values of the mixing measure:* $M = 0.3$ *(dashed line; squares),* $M = 0.1$ *(solid line, points), and* $M = 0.01$ *(dot-dash line, triangles). The long-time asymptotic result* (63) *is also plotted as a dotted line for the case of* $M = 0.01$.

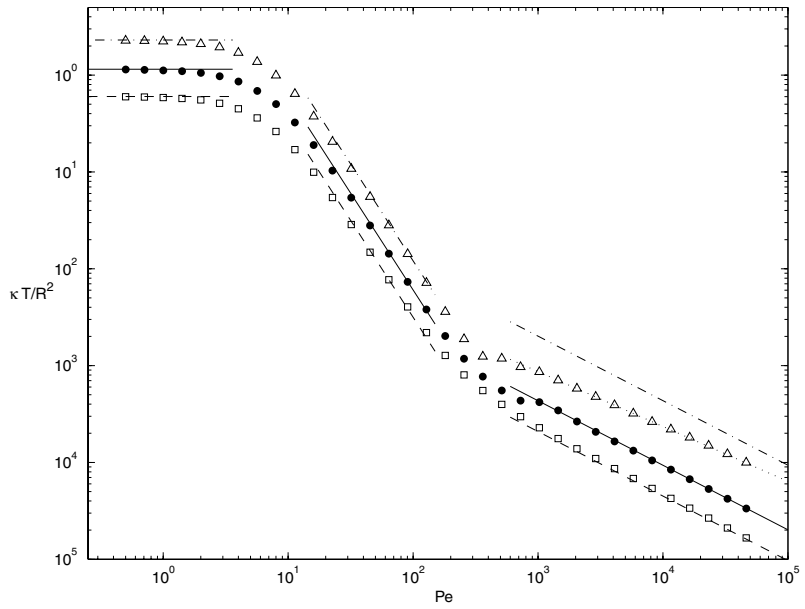


FIG. 7. *Mixing times nondimensionalized by the diffusion time, $\kappa T/R^2$, as a function of Péclet number, for $\gamma = 0.05$. Asymptotic results are shown as lines, and numerical results as symbols for values of the mixing measure: $M = 0.3$ (dashed line; squares), $M = 0.1$ (solid line, points), and $M = 0.01$ (dot-dash line, triangles). The long-time asymptotic result* (70) *is also plotted as a dotted line for the case of $M = 0.01$.*

regimes are thus found by multiplying (26), (40), (56), and (63) by γ/Pe, yielding

$$\frac{\kappa T_M}{R^2} \approx \frac{1}{2}\ln\left(\frac{1}{M}\right)\left[1 - \frac{1}{3}\gamma^2 + \cdots\right] \tag{67}$$

for $Pe \ll 7.2$ and

$$\frac{\kappa T_M}{R^2} \approx \frac{1}{Pe^2}\ln\left(\frac{1}{M}\right)\frac{105}{4}\left[1 - \frac{18165}{1576}\gamma^2 + \cdots\right] \tag{68}$$

for $7.2 \ll Pe \ll 15/\gamma$. For the highest Péclet numbers $Pe \gg 15/\gamma$, we obtained two asymptotic forms, the first valid for short times (moderate M values),

$$\frac{\kappa T_M}{R^2} \approx Pe^{-\frac{2}{3}}\gamma^{\frac{4}{3}}\left[\frac{1}{6}F^{-1}\left(\frac{2M}{\sqrt{\pi}}\right)\right]^{1/3}, \tag{69}$$

and the second for longer times (smaller M values),

$$\frac{\kappa T_M}{R^2} \approx Pe^{-\frac{1}{2}}\gamma^{\frac{3}{2}}\frac{1}{4\sqrt{3}}\ln\left[\frac{4\pi^2}{3M^4}\frac{1}{\gamma Pe}\right]. \tag{70}$$

Figure 7 is especially of interest to experimentalists working with a particular solute and solvent (so that κ is fixed) while varying the rate of rotation velocity ω of the micromixer to change Pe. The mixing time is seen to decrease from the diffusion time through the Taylor regime (at a rate proportional to Pe^{-2}), and then continue to decrease at a slower rate beyond the Taylor regime. The slower rate corresponds

to a $Pe^{-2/3}$ scaling when (54) is valid, i.e., for $M \geq 0.1$, but is closer to $Pe^{-1/2} \ln Pe$ for very small values of M.

From our analysis of the limits of validity of the Taylor dispersion description (see (44)), we can estimate the end of the Taylor regime as $15\gamma^{-1}$ for small γ. Of interest to the micromixer designer is the influence of the geometry ratio γ: since the mixing time decreases as Pe^{-2} until $Pe \approx 15\gamma^{-1}$, it seems advisable to decrease γ as much as possible in order to achieve faster mixing at lower velocities. However, this approach is overly simplistic, since the Péclet number also depends on γ—the linear velocity in the channel decreases with decreasing ρ, all other things being equal. A problem of great relevance to experimentalists is to find the optimal annular geometry to give the lowest possible mixing time for a given species, and given values of the electric and magnetic pumping fields (so κ and α are fixed). Suppose that the "footprint" R of the device is chosen; then we must find the channel half-width ρ to minimize the mixing time. Since Pe may be shown to increase monotonically with ρ, we conclude from Figure 7 that the mixing time decreases as the channel width increases. Practical considerations such as the minimal inner radius for effective electrodes will also provide an upper bound on the acceptable geometry ratio γ.

7. Summary. Laminar mixing in a two-dimensional annulus has been examined by asymptotic analysis and numerical simulation of the convection-diffusion equation (8). The mixing measure defined in (16) has been shown to decay exponentially in the diffusion and Taylor regimes (see (23) and (38)), with time constants given in terms of the Péclet number and the ratio γ characterizing the annulus geometry. Corresponding mixing times T_M are predicted to scale as Pe^0 and Pe^{-2} relative to the azimuthal diffusion time; see (67) and (68). Figures 3–7 demonstrate the robustness of these predictions by comparing them to numerical simulations. In the convection-dominated regime, asymptotic analysis modelled upon studies of mixing in vortices yields the mixing measure (60), with two interesting subregimes for the mixing time: a $Pe^{-2/3}$ dependence for early times, crossing over to a $Pe^{-1/2} \ln Pe$ scaling at longer times, as shown in (69) and (70), respectively. These predictions are again found to agree well with numerical results. Although the asymptotic formulas were derived in the limit $\gamma \to 0$, we find that they also yield good predictions for larger values of γ (Figure 5).

An important limitation of the present analysis is the assumption that all motion is two-dimensional; it is known [15, 16] that the dispersion in a channel of finite depth may not equal the value calculated by assuming the depth to be infinite. Work incorporating three-dimensional effects may yield further insight into this important problem.

Acknowledgment. We acknowledge the helpful comments of the anonymous referees.

REFERENCES

[1] J. P. Gleeson and J. West, *Magnetohydrodynamic micromixing*, in Proceedings of the Fifth International Conference on Modeling and Simulation of Microsystems 2002, Puerto Rico, Computational Publications, Cambridge, MA, 2002, pp. 318–321.

[2] I. Meisel and P. Ehrhard, *Simulation of electrically excited flows in microchannels for mixing application*, in Proceedings of the Fifth International Conference on Modeling and Simulation of Microsystems 2002, Puerto Rico, Computational Publications, Cambridge, MA, 2002, pp. 62–65.

[3] K. Mohseni, *Mixing and impulse extremization in microscale vortex formation*, in Proceedings of the Fifth International Conference on Modeling and Simulation of Microsystems 2002, Puerto Rico, Computational Publications, Cambridge, MA, 2002, pp. 392–395.
[4] J. West, B. Karamata, B. Lillis, J. P. Gleeson, J. Alderman, J. K. Collins, W. Lane, A. Mathewson, and H. Berney, *Application of magnetohydrodynamic actuation to continuous flow chemistry*, Lab on a Chip, 2 (2002), pp. 224–230.
[5] P. Tabeling and J. P. Chabrerie, *Magnetohydrodynamic Taylor vortex flow under a transverse pressure gradient*, Phys. Fluids, 24 (1981), pp. 406–412.
[6] M-H. Chang and C-K. Chen, *Hydromagnetic stability of current-induced flow in a small gap between concentric cylinders*, J. Fluids Engrg., 121 (1999), pp. 548–554.
[7] G. I. Taylor, *Dispersion of soluble matter in solvent flowing slowly through a tube*, Proc. Roy. Soc. London A, 219 (1953), pp. 186–203.
[8] R. F. Probstein, *Physiochemical Hydrodynamics*, Wiley, New York, 1994.
[9] R. J. Nunge, T.-S. Lin, and W. N. Gill, *Laminar dispersion in curved tubes and channels*, J. Fluid Mech., 51 (1972), pp. 363–383.
[10] D. I. Pullin and T. S. Lundgren, *Axial motion and scalar transport in stretched spiral vortices*, Phys. Fluids, 13 (2001), pp. 2553–2563.
[11] P. Flohr and J. C. Vassilicos, *Accelerated scalar dissipation in a vortex*, J. Fluid Mech., 348 (1997), pp. 295–317.
[12] P. B. Rhines and W. R. Young, *How rapidly is a passive scalar mixed within closed streamlines?*, J. Fluid Mech., 133 (1983), pp. 133–145.
[13] K. Bajer, A. P. Bassom, and A. D. Gilbert, *Accelerated diffusion in the centre of a vortex*, J. Fluid Mech., 437 (2001), pp. 395–411.
[14] M. J. Lighthill, *Initial development of diffusion in a Poiseuille flow*, J. Inst. Math. Appl., 2 (1966), pp. 97–108.
[15] H. Brenner and D. A. Edwards, *Macrotransport Processes*, Butterworth–Heinemann, Boston, 1993.
[16] M. R. Doshi, P. M. Daiya, and W. N. Gill, *Three dimensional laminar dispersion in open and closed rectangular conduits*, Chem. Eng. Sci., 33 (1978), pp. 795–804.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1311–1333

TOTAL BOUNDED VARIATION REGULARIZATION AS A BILATERALLY CONSTRAINED OPTIMIZATION PROBLEM*

M. HINTERMÜLLER† AND K. KUNISCH‡

Abstract. It is demonstrated that the predual for problems with total bounded variation regularization terms can be expressed as a bilaterally constrained optimization problem. Existence of a Lagrange multiplier and an optimality system are established. This allows us to utilize efficient optimization methods developed for problems with box constraints in the context of bounded variation formulations. Here, in particular, the primal-dual active set method, considered as a semismooth Newton method, is analyzed, and superlinear convergence is proved. As a by-product we obtain that the Lagrange multiplier associated with the box constraints acts as an edge detector. Numerical results for image denoising and zooming/resizing show the efficiency of the new approach.

Key words. total bounded variation, predual, semismooth Newton methods, box constraints, image reconstruction

AMS subject classifications. 94A08, 49M29, 65K05

DOI. 10.1137/S0036139903422784

1. Introduction and notation.

This work is concerned with the study of the problem

$$\left\{\begin{array}{ll}\min & \frac{1}{2}\int_\Omega |Ku-f|^2dx+\frac{\alpha}{2}\int_\Omega |u|^2dx+\beta\int_\Omega |\mathrm{D}u| \\ \text{over} & u\in \mathrm{BV}(\Omega),\end{array}\right. \tag{1.1}$$

where Ω is a simply connected domain in $\mathbb{R}^2$ with Lipschitz continuous boundary $\partial\Omega$, $f\in \mathrm{L}^2(\Omega)$, $\beta>0$, $\alpha\geq 0$ are given, and $\mathrm{K}\in\mathcal{L}(\mathrm{L}^2(\Omega))$. By K^* we denote the adjoint of K. We assume that K^*K is invertible or $\alpha>0$. Further, $\mathrm{BV}(\Omega)$ denotes the space of functions of bounded variation. A function u is in $\mathrm{BV}(\Omega)$ if the BV seminorm defined by

$$\int_\Omega |Du| = \sup\left\{\int_\Omega u\ \mathrm{div}\,\vec{v} : \vec{v}\in (C_0^\infty(\Omega))^2,\ |\vec{v}(x)|_{\ell^\infty}\leq 1\right\}$$

is finite. It is well known [16] that $\mathrm{BV}(\Omega)\subset \mathrm{L}^2(\Omega)$ for $\Omega\subset\mathbb{R}^2$, and that $u\mapsto |u|_{L^2}+\int_\Omega |\mathrm{D}u|$ defines a norm on $\mathrm{BV}(\Omega)$. If $K=$ *identity*, then (1.1) is the well-known image restoration problem with BV-regularization term. It consists of recovering the true image u from the noisy image f. It is well known [9] that (1.1) admits a unique solution $u^*\in\mathrm{BV}(\Omega)$. BV-regularization, differently from regularization by means of $\int_\Omega |\nabla u|^2\,dx$, for example, is known to be preferable due to its ability to preserve edges in the original image during the reconstruction process. Since the pioneering work in [23], the literature on (1.1) has grown tremendously. We give some selected references [1, 5, 7, 12, 14, 18] and refer to the recent monograph [25] for further references.

*Received by the editors February 14, 2003; accepted for publication (in revised form) December 18, 2003; published electronically May 5, 2004. This research was supported by the special research grant SFB "Optimierung und Kontrolle."

http://www.siam.org/journals/siap/64-4/42278.html

†Department of Computational and Applied Mathematics, Rice University, CAAM - MS 134, 6100 S. Main Street, Houston, TX 77005 (hint@caam.rice.edu).

‡Department of Mathematics, University of Graz, Heinrichstr. 36, A-8010 Graz, Austria (karl.kunisch@uni-graz.at).

The original formulation has been extended in various directions including concepts of reconstruction of images with multiple scales; see, e.g., [2, 4, 6, 19].

Despite its favorable properties for reconstruction of images, and especially images with blocky structure, problem (1.1) poses some severe difficulties. On the analytical level these are related to the fact that (1.1) is posed in a nonreflexive Banach space, the dual of which is difficult to characterize [16, 19], and on the numerical level the optimality system related to (1.1) consists of a nonlinear partial differential equation, which is not directly amenable to numerical implementations.

In the present work we show the remarkable result that while the dual of the nonreflexive Banach space problem (1.1) has a complicated measure theoretic structure, its predual can be characterized in a well-known Hilbert space setting. Specifically, the predual to (1.1) is a quadratic optimization problem with bilateral constraints. For such problems the literature provides a variety of possible algorithms. Here we describe and analyze two variants of semismooth Newton methods. We prove their superlinear convergence and provide numerical examples for some denoising and zooming problems. In practice these algorithms are globally convergent without the need for line searches. As a by-product we obtain that the Lagrange multiplier associated with the box constraints acts as an edge detector. We show numerically that the edge detecting property does not require any postprocessing on the multiplier such as thresholding or sharpening techniques.

Let us briefly mention a few alternatives that have been investigated for treating (1.1) numerically. In [23] a time marching scheme to solve the necessary optimality condition related to (1.1) is used. Time marching is also essential for the work in, e.g., [6]. In [19, 26] fixed point iteration schemes are applied to the optimality system using primal variables only. The optimality system based on the primal and dual variables is the basis for the schemes in [19] and [8]. In the former an augmented Lagrangian-based active set strategy is used; in the latter a Newton method is applied. Compared to the formulations used in earlier work, ours appears to have the advantage of being of significantly simpler structure since only a quadratic problem with affine box constraints must be solved. In earlier work, if analysis is carried out, then frequently $\int_\Omega |\mathrm{D}u|$ is replaced by

$$\int_\Omega \sqrt{\delta + |\nabla u|^2}\, dx, \tag{1.2}$$

for $\delta > 0$. In our approach the algorithms are well posed for $\delta = 0$, and for the discretized formulations we have superlinear convergence, still with $\delta = 0$.

The paper is organized as follows. In the remainder of this section we recall some facts from convex analysis and summarize the function space notation that will be used. In section 2 we characterize the predual of (1.1) in the sense of Fenchel. We shall point out the close connection, for 1D (one–dimensional) problems, between our algorithm and the taut-string algorithm well known in nonparametric regression analysis [11, 21]. Section 3 is devoted to the description and convergence proof for a class of regularized problems. Semismooth Newton methods for the predual problems are developed in section 4. Superlinear convergence for the regularized infinite-dimensional problems, and for the discretized predual problems without extra regularization, is proved. Section 5 is devoted to a numerical feasibility study of our results.

We recall the Fenchel duality theorem in infinite-dimensional spaces in a form that is convenient for our work; see, e.g., [3, 13] for details. Let V and Y be Banach spaces with topological duals denoted by V^* and Y^*, respectively. Further, let $\Lambda \in \mathcal{L}(V, Y)$

and let $\mathcal{F} : V \to \mathbb{R} \cup \{\infty\}$, $\mathcal{G} : Y \to \mathbb{R} \cup \{\infty\}$ be convex lower semicontinuous functionals not identically equal to ∞, and assume that there exists $v_0 \in V$ such that $\mathcal{F}(v_0) < \infty$, $\mathcal{G}(\Lambda v_0) < \infty$, and $\mathcal{G}$ is continuous at Λv_0. Then we have

$$\inf_{u \in V} \mathcal{F}(u) + \mathcal{G}(\Lambda u) = \sup_{p \in Y^*} -\mathcal{F}^*(\Lambda^* p) - \mathcal{G}^*(-p), \tag{1.3}$$

where $\mathcal{F}^* : V^* \to \mathbb{R} \cup \{\infty\}$ denotes the conjugate of $\mathcal{F}$ defined by

$$\mathcal{F}^*(v^*) = \sup_{v \in V} \langle v, v^* \rangle_{V,V^*} - \mathcal{F}(v).$$

Under the conditions imposed on $\mathcal{F}$ and $\mathcal{G}$, it is known that the problem on the right-hand side of (1.3) admits a solution. Moreover, $(\bar{u}, \bar{p})$ are solutions to the two optimization problems in (1.3) if and only if

$$\Lambda^* \bar{p} \in \partial \mathcal{F}(\bar{u}), \tag{1.4a}$$

$$-\bar{p} \in \partial \mathcal{G}(\Lambda \bar{u}), \tag{1.4b}$$

where $\partial \mathcal{F}$ denotes the subdifferential of the convex functional $\mathcal{F}$.

To compute, formally, the Fenchel dual to (1.1) we set $\Lambda = \nabla$,

$$\mathcal{F}(u) = \frac{1}{2}|\mathrm{K}u - f|^2 + \frac{\alpha}{2}|u|^2 \qquad \textit{and} \qquad \mathcal{G}(\vec{p}) = \beta \int_\Omega |\vec{p}|_{\ell^1} dx,$$

where u and $\vec{p}$ denote a scalar and a 2D vector-valued function, respectively. Further, $|\cdot|$ denotes the $L^2(\Omega)$-norm and $|\cdot|_{\ell^1}$ stands for the ℓ^1-norm on $\mathbb{R}^n$. For the convex conjugates we find

$$\mathcal{F}^*(v) = \frac{1}{2}(v + \mathrm{K}^* f, B^{-1}(v + \mathrm{K}^* f)) - \frac{1}{2}|f|^2 \quad \text{and} \quad \mathcal{G}^*(\vec{p}) = \mathrm{I}_{[-\beta\vec{1}, \beta\vec{1}]}(\vec{p}),$$

where $\vec{1}$ is the 2D vector field with 1 in both coordinates, $B = \alpha\,\mathrm{I} + \mathrm{K}^*\mathrm{K}$, and

$$\mathrm{I}_{[-\beta\vec{1}, \beta\vec{1}]}(\vec{p}) = \begin{cases} 0 & \text{if } -\beta\vec{1} \le \vec{p}(x) \le \beta\vec{1} \qquad \text{for almost every (a.e.) } x \in \Omega, \\ \infty & \text{otherwise.} \end{cases}$$

Thus, formally the dual to (1.1) is given by

$$\begin{cases} \inf \ \frac{1}{2}|\operatorname{div} \vec{p} + \mathrm{K}^* f|_B^2 \\ \text{s.t.} \ -\beta\vec{1} \le \vec{p}(x) \le \beta\vec{1} \qquad \text{for a.e. } x \in \Omega, \end{cases} \tag{1.5}$$

where $|v|_B^2 = (v, B^{-1}v)$, and the relationship (1.4) applied to the solutions of (1.1) and (1.5) implies that

$$\operatorname{div} \vec{p} = Bu - \mathrm{K}^* f, \quad \vec{p} = \beta \left(\frac{u_{x_i}}{|u_{x_i}|} \right)_{i=1}^n \qquad \text{on } \{x : u_{x_i}(x) \neq 0 \text{ for all } i\}. \tag{1.6}$$

The functional analytic statement corresponding to (1.6) is given in (2.3), (2.4) below.

We note that nondifferentiability due to the BV-term in (1.1) is replaced by the bilateral constraints in the formal dual (1.5).

In the next section we shall put (1.5) into a proper functional analytical framework. For this purpose we require some notation which we summarize next. Let $\mathbb{L}^2(\Omega) = \mathrm{L}^2(\Omega) \times \mathrm{L}^2(\Omega)$ endowed with the Hilbert space inner product structure and

norm. If the context suggests to do so, then we shall distinguish between vector fields $\vec{v} \in \mathbb{L}^2(\Omega)$ and scalar functions $v \in \mathrm{L}^2(\Omega)$ by using an arrow on top of the letter. Analogously we set $\mathbb{H}_0^1(\Omega) = H_0^1(\Omega) \times H_0^1(\Omega)$. We denote $L_0^2(\Omega) = \{v \in \mathrm{L}^2(\Omega) : \int_\Omega v dx = 0\}$, $H_0(\mathrm{div}) = \{\vec{v} \in \mathbb{L}^2(\Omega) : \mathrm{div}\,\vec{v} \in \mathrm{L}^2(\Omega), \vec{v} \cdot n = 0 \text{ on } \partial\Omega\}$, where n is the outer normal to $\partial\Omega$. The space $H_0(\mathrm{div})$ is endowed with $|\vec{v}|^2_{H_0(\mathrm{div})} = |\vec{v}|^2_{\mathbb{L}^2(\Omega)} + |\,\mathrm{div}\,\vec{v}|^2_{L^2}$ as norm. Further, we put $H_0(\mathrm{div}\,0) = \{\vec{v} \in H_0(\mathrm{div}) : \mathrm{div}\,\vec{v} = 0 \text{ almost everywhere in } \Omega\}$. It is well known that

$$\mathbb{L}^2(\Omega) = \mathrm{grad}\; H^1(\Omega) \oplus H_0(\mathrm{div}\,0); \tag{1.7}$$

cf. [10, p. 216], for example. Moreover,

$$H_0(\mathrm{div}) = H_0(\mathrm{div}\,0)^\perp \oplus H_0(\mathrm{div}\,0), \tag{1.8}$$

with

$$H_0(\mathrm{div}\,0)^\perp = \{\vec{v} \in \mathrm{grad}\, H^1(\Omega) : \mathrm{div}\,\vec{v} \in L^2(\Omega), \vec{v} \cdot n = 0 \text{ on } \partial\Omega\},$$

and $\mathrm{div} : H_0(\mathrm{div}\,0)^\perp \subset H_0(\mathrm{div}) \to L_0^2(\Omega)$ is a homeomorphism. In fact, it is injective by construction, and for every $f \in L_0^2(\Omega)$ there exists, by the Lax–Milgram lemma, $\varphi \in H^1(\Omega)$ such that

$$\mathrm{div}\,\nabla\varphi = f \text{ in } \Omega, \qquad \nabla\varphi \cdot n = 0 \text{ on } \partial\Omega,$$

with $\nabla\varphi \in H_0(\mathrm{div}\,0)^\perp$. Hence, by the closed mapping theorem we have

$$\mathrm{div} \in \mathcal{L}(H_0(\mathrm{div})^\perp, L_0^2(\Omega)).$$

Finally, let $\mathrm{P}_{\mathrm{div}}$ and $\mathrm{P}_{\mathrm{div}^\perp}$ denote the orthogonal projections in $\mathbb{L}^2(\Omega)$ onto $H_0(\mathrm{div}\,0)$ and $\mathrm{grad}\, H^1(\Omega)$, respectively. Note that the restrictions of $\mathrm{P}_{\mathrm{div}}$ and $\mathrm{P}_{\mathrm{div}^\perp}$ to $H_0(\mathrm{div}\,0)$ coincide with the orthogonal projections in $H_0(\mathrm{div})$ onto $H_0(\mathrm{div}\,0)$ and $H_0(\mathrm{div}\,0)^\perp$.

2. The Fenchel predual. The section is devoted to the study of the problems

$$\begin{cases} \min & \dfrac{1}{2}|\,\mathrm{div}\,\vec{p} + \mathrm{K}^* f|_B^2 \quad \text{over } \vec{p} \in H_0(\mathrm{div}) \\ \text{s.t.} & -\beta\vec{1} \le \vec{p} \le \beta\vec{1}, \end{cases} \tag{2.1}$$

and

$$\begin{cases} \min & \dfrac{1}{2}|\,\mathrm{div}\,\vec{p} + \mathrm{K}^* f|_B^2 + \frac{\gamma}{2}|\mathrm{P}_{\mathrm{div}}\vec{p}|^2 \quad \text{over } \vec{p} \in H_0(\mathrm{div}) \\ \text{s.t.} & -\beta\vec{1} \le \vec{p} \le \beta\vec{1}, \end{cases} \tag{2.2}$$

where $\gamma > 0$ is given, and we recall that for $v \in \mathrm{L}^2(\Omega)$ we set $|v|_B^2 = (v, B^{-1}v)_{L^2}$.

PROPOSITION 2.1. *Both* (2.1) *and* (2.2)*, admit a solution. The solution to* (2.2) *is unique.*

Proof. Existence of a solution to (2.1) as well as (2.2) can be proved by standard arguments. To verify uniqueness of the solution to (2.2) we note that the set of feasible $\vec{p}$ is convex. Hence it suffices to verify strict convexity of $J(\vec{p}) = \frac{1}{2}|\,\mathrm{div}\,\vec{p} + \mathrm{K}^* f|_B^2 + \frac{\gamma}{2}|\mathrm{P}_{\mathrm{div}}\vec{p}|^2$. To ascertain strict convexity of J we use the fact that the second derivative satisfies

$$J''(\vec{p}, \vec{p}) = |\,\mathrm{div}\,\vec{p}\,|_B^2 + \gamma|\mathrm{P}_{\mathrm{div}}\vec{p}\,|^2 \ge \kappa|\vec{p}\,|^2_{H_0(\mathrm{div})}$$

for a constant $\kappa > 0$ independent of $\vec{p} \in H_0(\mathrm{div})$. Here we have used (1.6) and the subsequent comments. Hence J is even uniformly convex, and uniqueness follows. ∎

THEOREM 2.2. *The Fenchel dual to* (2.1) *is given by* (1.1)*, and the solutions* u^* *of* (1.1) *and* $\vec{p}^*$ *of* (2.1) *are related by*

$$Bu^* = \mathrm{div}\,\vec{p}^* + \mathrm{K}^* f, \tag{2.3}$$

$$\langle (-\mathrm{div})^* u^*, \vec{p} - \vec{p}^* \rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} \le 0 \quad \textit{for all } \vec{p} \in H_0(\mathrm{div}), \tag{2.4}$$

with $-\beta\vec{1} \le \vec{p} \le \beta\vec{1}$.

Alternatively, (2.1) can be considered as the predual of the original problem (1.1). If (2.1) is a zero-residue problem, i.e., $\vec{p}^*$ satisfies $\mathrm{div}\,\vec{p}^* = -\mathrm{K}^* f$, then the additional penalty term in (2.2) chooses from among all solutions the one which minimizes $|\mathrm{P}_{\mathrm{div}}\vec{p}^*|$.

Proof of Theorem 2.2. We apply Fenchel duality as recalled in section 1 with $V = H_0(\mathrm{div})$, $Y = Y^* = \mathrm{L}^2(\Omega)$, $\Lambda = -\mathrm{div}$, $\mathcal{G} : Y \to \mathbb{R}$ given by $\mathcal{G}(v) = \frac{1}{2}|v - \mathrm{K}^* f|_B^2$, and $\mathcal{F} : V \to \mathbb{R}$ defined by $\mathcal{F}(\vec{p}\,) = \mathrm{I}_{[-\beta\vec{1},\beta\vec{1}]}(\vec{p}\,)$. The convex conjugate $\mathcal{G}^* : L^2(\Omega) \to \mathbb{R}$ of $\mathcal{G}$ is given by

$$\mathcal{G}^*(v) = \frac{1}{2}|\mathrm{K}v + f|^2 + \frac{\alpha}{2}|v|^2 - \frac{1}{2}|f|^2.$$

Further, the conjugate $\mathcal{F}^* : H_0(\mathrm{div})^* \to \mathbb{R}$ of $\mathcal{F}$ is given by

$$\mathcal{F}^*(\vec{q}) = \sup_{\vec{p} \in S_1} \langle \vec{q}, \vec{p}\,\rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} \quad \text{for } \vec{q} \in H_0(\mathrm{div})^*, \tag{2.5}$$

where $S_1 = \{\vec{p} \in H_0(\mathrm{div}) : -\beta\vec{1} \le \vec{p} \le \beta\vec{1}\}$. Let us set

$$S_2 = \{\vec{p} \in C_0^1(\Omega) \times C_0^1(\Omega) : -\beta\vec{1} \le \vec{p} \le \beta\vec{1}\}.$$

The set S_2 is dense in the topology of $H_0(\mathrm{div})$ in S_1. In fact, let $\vec{p}$ be an arbitrary element of S_1. Since $(\mathcal{D}(\Omega))^2$ is dense in $H_0(\mathrm{div})$ (see, e.g., [15, p. 26]), there exists a sequence $\vec{p}_n \in (\mathcal{D}(\Omega))^2$ converging in $H_0(\mathrm{div})$ to $\vec{p}$. Let $\mathcal{P}$ denote the canonical projection in $H_0(\mathrm{div})$ onto the closed convex subset S_1 and note that, since $\vec{p} \in S_1$,

$$\begin{aligned} |\vec{p} - \mathcal{P}\vec{p}_n|_{H_0(\mathrm{div})} &\le |\vec{p} - \vec{p}_n|_{H_0(\mathrm{div})} + |\vec{p}_n - \mathcal{P}\vec{p}_n|_{H_0(\mathrm{div})} \\ &\le 2|\vec{p} - \vec{p}_n|_{H_0(\mathrm{div})} \to 0 \qquad \text{for } n \to \infty. \end{aligned}$$

Hence $\lim_{n\to\infty} |\vec{p} - \mathcal{P}\vec{p}_n|_{H_0(\mathrm{div})} = 0$ and S_2 is dense in S_1. Returning to (2.5), we have for $v \in \mathrm{L}^2(\Omega)$ and $(-\mathrm{div})^* \in \mathcal{L}(\mathrm{L}^2(\Omega), V^*)$,

$$\mathcal{F}^*((-\mathrm{div})^* v) = \sup_{\vec{p} \in S_2} (v, -\mathrm{div}\,\vec{p}\,),$$

which can be $+\infty$. By the definition of the functions of bounded variation it is finite if and only if $v \in \mathrm{BV}(\Omega)$ (see [16, p. 3]) and

$$\mathcal{F}^*((-\mathrm{div})^* v) = \beta \int_\Omega |\mathrm{D}v| < \infty \quad \text{for } v \in \mathrm{BV}(\Omega).$$

The dual problem to (2.1) is found to be

$$\min\ \frac{1}{2}|\mathrm{K}u - f|^2 + \frac{\alpha}{2}|u|^2 + \beta \int_\Omega |\mathrm{D}u| \quad \text{over } u \in \mathrm{BV}(\Omega).$$

From (1.4), moreover, we find

$$\langle(-\operatorname{div})^* u^*, \vec{p} - \vec{p}^*\rangle_{H_0(\operatorname{div})^*, H_0(\operatorname{div})} \le 0 \quad \text{for all } p \in S_1$$

and

$$Bu^* = \operatorname{div} \vec{p}^* + \mathrm{K}^* f. \qquad \square$$

We obtain the following optimality system.

COROLLARY 2.3. *Let $\vec{p}^* \in H_0(\operatorname{div})$ be a solution to (2.1). Then there exists $\vec{\lambda}^* \in H_0(\operatorname{div})^*$ such that*

$$\operatorname{div}^* B^{-1} \operatorname{div} \vec{p}^* + \operatorname{div}^* B^{-1} \mathrm{K}^* f + \vec{\lambda}^* = 0, \tag{2.6}$$

$$\langle \vec{\lambda}^*, \vec{p} - \vec{p}^*\rangle_{H_0(\operatorname{div})^*, H_0(\operatorname{div})} \le 0 \quad \textit{for all } \vec{p} \in H_0(\operatorname{div}), \tag{2.7}$$

with $-\beta\vec{1} \le \vec{p} \le \beta\vec{1}$.

For convenience we also specify the variational form of (2.6) which holds in $H_0(\operatorname{div})^*$:

$$(B^{-1} \operatorname{div} \vec{p}^*, \operatorname{div} \vec{v}\,)_{L^2} + (B^{-1}\mathrm{K}^* f, \operatorname{div} \vec{v}\,)_{L^2} + \langle \vec{\lambda}^*, \vec{v}\,\rangle_{H_0(\operatorname{div})^*, H_0(\operatorname{div})} = 0$$

for all $\vec{v} \in H_0(\operatorname{div})$.

Proof of Corollary 2.3. Set $\vec{\lambda}^* = -\operatorname{div}^* u^* \in H_0(\operatorname{div})^*$ and apply $\operatorname{div}^* B^{-1}$ to obtain (2.6). For this choice of $\vec{\lambda}^*$, equation (2.7) follows from (2.4). $\square$

The optimality system for (2.2) is given next.

COROLLARY 2.4. *Let $\vec{p}^* \in H_0(\operatorname{div})^*$ denote the solution to (2.2). Then there exists $\vec{\lambda}^* \in H_0(\operatorname{div})^*$ such that*

$$\operatorname{div}^* B^{-1} \operatorname{div} \vec{p}^* + \operatorname{div}^* B^{-1} \mathrm{K}^* f + \gamma P_{\operatorname{div}} \vec{p}^* + \vec{\lambda}^* = 0, \tag{2.8}$$

$$\langle \vec{\lambda}^*, \vec{p} - \vec{p}^*\rangle_{H_0(\operatorname{div})^*, H_0(\operatorname{div})} \le 0 \quad \textit{for all } \vec{p} \in H_0(\operatorname{div}), \tag{2.9}$$

with $-\beta\vec{1} \le \vec{p} \le \beta\vec{1}$.

Proof. We only sketch the proof here since the assertion will also follow from the proof of Theorem 3.1 below. By (1.6), every $\vec{v} \in H_0(\operatorname{div})$ can be decomposed according to $\vec{v} = \vec{v}_1 + \vec{v}_2 \in H_0(\operatorname{div} 0)^\perp \oplus H_0(\operatorname{div} 0)$. The functional in (2.2) is then separable, and (2.2) can be expressed as

$$\min_{\vec{p} \in H_0(\operatorname{div})} \mathcal{F}(\vec{p}) + \mathcal{G}_1(\Lambda_1 \vec{p}_1) + \mathcal{G}_2(\Lambda_2 \vec{p}_2),$$

where $\mathcal{F}$ is defined in the proof of Theorem 2.2, $\mathcal{G}_1$ and Λ_1 coincide with $\mathcal{G}$ and Λ from the proof of Theorem 2.2, and we set

$$\mathcal{G}_2 : \mathbb{L}^2(\Omega) \to \mathbb{R}, \quad \mathcal{G}_2(\vec{p}) = \frac{\gamma}{2} |\vec{p}\,|^2_{\mathbb{L}^2(\Omega)},$$

$\Lambda_2 \in \mathcal{L}(H_0(\operatorname{div} 0), \mathbb{L}^2(\Omega))$ with Λ_2 the canonical injection. From general results in convex analysis (e.g., [13, p. 61]), there exist $\vec{u}_1^* \in \mathbb{L}^2(\Omega)$ and $\vec{u}_2^* \in \mathbb{L}^2(\Omega)$ such that

$$\begin{aligned} B\vec{u}_1^* &= \operatorname{div} \vec{p}_1^* + \mathrm{K}^* f = \operatorname{div} \vec{p}^* + \mathrm{K}^* f, \\ -\vec{u}_2^* &= \gamma \vec{p}_2^* = \gamma P_{\operatorname{div}} \vec{p}^* \end{aligned}$$

and

$$\langle -\operatorname{div}^* \vec{u}_1^* + \vec{u}_2^*, \vec{p} - \vec{p}^* \rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} \le 0 \qquad \text{for all } \vec{p} \in S_1.$$

The claim follows with $\vec{\lambda}^* = -\operatorname{div}^* \vec{u}_1^* + \vec{u}_2^*$. □

We end this section with the following remarks.

Remark 1.

- In our numerical tests, in many cases we can set $\gamma = 0$. This suggests the conjecture that the constraints $-\beta\vec{1} \le \vec{p} \le \beta\vec{1}$ imply some type of uniqueness.
- We point out the close connection between (2.1) and the taut-string algorithm well known in regression analysis [11, 21]. Here we have K = I, $\alpha = 0$. A continuous version of the taut-string algorithm can be expressed as

$$\begin{cases} \min \displaystyle\int_0^1 \sqrt{1 + |w_x|^2}\, dx, \\ \text{s.t. } F - \beta \le w \le F + \beta, \end{cases} \tag{2.10}$$

where $F(x) = \int_0^x f(s)ds$. The denoised image u is obtained from $u = w_x$. Observe that the change of variables $p = w - F$ transforms (2.10) into

$$\begin{cases} \min \displaystyle\int_0^1 \sqrt{1 + |p_x + f|^2}\, dx, \\ \text{s.t. } -\beta \le p \le \beta \end{cases} \tag{2.11}$$

and $u = p_x + f$. Thus, except for the square root in (2.11), we obtain (2.1).

3. A family of regularized problems. To treat (2.1) and (2.2) numerically one can discretize these box constrained problems and implement one's algorithm of choice for the resulting finite-dimensional quadratic optimization problems with affine constraints. With such an approach the infinite-dimensional structure tends to get covered up. One of the features that can be pointed out by considering (2.6) and (2.8) of the optimality systems is that the leading differential operator is not smoothing (see (1.7)) as it is for obstacle-type problems, nor is it a compact perturbation of the identity operator as, for instance, for control constrained optimal control problems [17]. This complicates the convergence analysis for semismooth Newton algorithms; see [17, 24]. Therefore we describe in this section a family of approximating problems which have more amenable properties for Newton-type algorithms in an infinite-dimensional setting. A second difficulty with (2.1), (2.2) is related to the fact that β will typically be chosen as a small constant so that the resulting problems are close to bottleneck problems. We shall see in section 5 that the algorithms we propose are able to deal efficiently with such constraints.

As announced above, we focus in this section on a family of approximating problems given by

$$\begin{cases} \min \ \dfrac{1}{2c}|\nabla\vec{p}|^2 + \dfrac{1}{2}|\operatorname{div}\vec{p} + \mathrm{K}^* f|_B^2 + \dfrac{\gamma}{2}|\mathrm{P}_{\mathrm{div}}\vec{p}|^2 \\ \qquad + \dfrac{1}{2c}|\max(0, c(\vec{p} - \beta\vec{1}))|^2 + \dfrac{1}{2c}|\min(0, c(\vec{p} + \beta\vec{1}))|^2 \text{ over } \vec{p} \in \mathbb{H}_0^1(\Omega), \end{cases} \tag{3.1}$$

where $c > 0$. Let $\vec{p}_c$ denote the unique solution to (3.1). It satisfies the optimality

condition

$$-\frac{1}{c}\Delta\vec{p}_c - \nabla B^{-1}\operatorname{div}\vec{p}_c - \nabla B^{-1}\mathrm{K}^* f + \gamma \mathrm{P}_{\mathrm{div}}\vec{p}_c + \vec{\lambda}_c = 0, \tag{3.2a}$$

$$\vec{\lambda}_c = \max(0, c(\vec{p}_c - \beta\vec{1})) + \min(0, c(\vec{p}_c + \beta\vec{1})). \tag{3.2b}$$

Next we address convergence as $c \to \infty$.

THEOREM 3.1. *The family $\{(\vec{p}_c, \vec{\lambda}_c)\}_{c>0}$ converges weakly in $H_0(\mathrm{div}) \times \mathbb{H}_0^1(\Omega)^*$ to the unique solution $(\vec{p}^*, \vec{\lambda}^*)$ of (2.8), (2.9). Moreover, the convergence of $\vec{p}_c$ to $\vec{p}^*$ is strong in $H_0(\mathrm{div})$.*

Proof. Recall the variational form of (2.8) given by

$$(\operatorname{div}\vec{p}^*, \operatorname{div}\vec{v}\,)_B + (\mathrm{K}^* f, \operatorname{div}\vec{v}\,)_B + \gamma(\mathrm{P}_{\mathrm{div}}\vec{p}^*, \mathrm{P}_{\mathrm{div}}\vec{v}\,) + \langle \vec{\lambda}^*, \vec{v}\,\rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} = 0 \tag{3.3}$$

for all $\vec{v} \in H_0(\mathrm{div})$. To verify uniqueness, let us suppose that $(\vec{p}_i, \vec{\lambda}_i) \in H_0(\mathrm{div}) \times H_0(\mathrm{div})^*$, $i = 1, 2$, are two solution pairs to (2.8), (2.9). For $\delta\vec{p} = \vec{p}_2 - \vec{p}_1$, $\delta\vec{\lambda} = \vec{\lambda}_2 - \vec{\lambda}_1$ we have

$$(B^{-1}\operatorname{div}\delta\vec{p}, \operatorname{div}\vec{v}\,) + \gamma(\mathrm{P}_{\mathrm{div}}\delta\vec{p}, \mathrm{P}_{\mathrm{div}}\vec{v}\,) + \langle \delta\vec{\lambda}, \vec{v}\,\rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} = 0 \tag{3.4}$$

for all $\vec{v} \in H_0(\mathrm{div})$, and

$$\langle \delta\vec{\lambda}, \delta\vec{p}\,\rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} \geq 0.$$

With $\vec{v} = \delta\vec{p}$ in (3.4) we obtain

$$|B^{-1}\operatorname{div}\delta\vec{p}\,|^2 + \gamma|\mathrm{P}_{\mathrm{div}}\delta\vec{p}\,|^2 \leq 0,$$

and hence $\vec{p}_1 = \vec{p}_2$. From (3.3) we deduce that $\vec{\lambda}_1 = \vec{\lambda}_2$. Thus uniqueness is established, and we can henceforth rely on subsequential arguments.

In the following computation we consider the coordinates $\vec{\lambda}_c^i$, $i = 1, 2$, of $\vec{\lambda}_c$. We have for the pointwise a.e. evaluation at $x \in \Omega$

$$\begin{aligned}\vec{\lambda}_c^i\,\vec{p}_c^i &= \big(\max(0, c(\vec{p}_c^i - \beta)) + \min(0, c(\vec{p}_c^i + \beta))\big)\,\vec{p}_c^i \\ &= \begin{cases} c(\vec{p}_c^i - \beta)\vec{p}_c^i & \text{if } \vec{p}_c^i \geq \beta, \\ 0 & \text{if } |\vec{p}_c^i| = \beta, \\ c(\vec{p}_c^i + \beta)\vec{p}_c^i & \text{if } \vec{p}_c^i \leq \beta. \end{cases}\end{aligned}$$

It follows that

$$(\vec{\lambda}_c^i, \vec{p}_c^i)_{\mathrm{L}^2(\Omega)} \geq \frac{1}{c}|\vec{\lambda}_c^i|^2_{\mathrm{L}^2(\Omega)} \quad \text{for } i = 1, 2,$$

and consequently

$$(\vec{\lambda}_c, \vec{p}_c)_{\mathbb{L}^2(\Omega)} \geq \frac{1}{c}|\vec{\lambda}_c|^2_{\mathbb{L}^2(\Omega)} \quad \text{for every } c > 0. \tag{3.5}$$

From (3.2) and (3.5) we deduce that

$$\frac{1}{c}|\nabla\vec{p}_c|^2 + |\operatorname{div}\vec{p}_c|_B^2 + \gamma|\mathrm{P}_{\mathrm{div}}\vec{p}_c|^2 \leq |\operatorname{div}\vec{p}_c|_B|\mathrm{K}^* f|_B$$

and hence

$$\frac{1}{c}|\nabla \vec{p}_c|^2 + \frac{1}{2}|\operatorname{div} \vec{p}_c|_B^2 + \gamma |\mathrm{P}_{\mathrm{div}} \vec{p}_c|^2 \le \frac{1}{2}|\mathrm{K}^* f|_B. \tag{3.6}$$

We further estimate

$$\begin{aligned} |\vec{\lambda}_c|_{\mathbb{H}_0^1(\Omega)^*} &= \sup_{|\vec{v}\,|_{\mathbb{H}_0^1(\Omega)}=1} \langle \vec{\lambda}_c, \vec{v} \rangle_{\mathbb{H}_0^1(\Omega)^*, \mathbb{H}_0^1(\Omega)} \\ &\le \sup_{|\vec{v}|_{\mathbb{H}_0^1(\Omega)}=1} \Big\{ \frac{1}{c}|\nabla \vec{p}_c||\nabla \vec{v}\,| + |\operatorname{div} \vec{p}_c|_B |\operatorname{div} \vec{v}\,|_B + |\mathrm{K}^* f|_B |\operatorname{div} \vec{v}\,|_B \\ &\qquad + \gamma |\mathrm{P}_{\mathrm{div}} \vec{p}_c|\,|\mathrm{P}_{\mathrm{div}} \vec{v}\,| \Big\}. \end{aligned}$$

From (3.6) we deduce the existence of a constant K independent of $c \ge 1$ such that

$$|\vec{\lambda}_c|_{\mathbb{H}_0^1(\Omega)^*} \le K. \tag{3.7}$$

Combining (3.6) and (3.7), we can assert the existence of $(\vec{p}^*, \vec{\lambda}^*) \in H_0(\mathrm{div}) \times \mathbb{H}_0^1(\Omega)^*$ such that for a subsequence denoted by the same symbol

$$(\vec{p}_c, \vec{\lambda}_c) \rightharpoonup (\vec{p}^*, \vec{\lambda}^*) \qquad \text{weakly in } H_0(\mathrm{div}) \times \mathbb{H}_0^1(\Omega)^*. \tag{3.8}$$

We recall the variational form of (3.2), i.e.,

$$\begin{aligned} \frac{1}{c}(\nabla \vec{p}_c, \nabla \vec{v}\,) + (\operatorname{div} \vec{p}_c, \operatorname{div} \vec{v}\,)_B + (\mathrm{K}^* f, \operatorname{div} \vec{v}\,)_B + \gamma (\mathrm{P}_{\mathrm{div}} \vec{p}_c, \mathrm{P}_{\mathrm{div}} \vec{v}\,) \\ + (\vec{\lambda}_c, \vec{v}\,) = 0 \quad \text{for all } \vec{v} \in \mathbb{H}_0^1(\Omega). \end{aligned}$$

Passing to the limit $c \to \infty$, using (3.6) and (3.8) we have

$$\begin{aligned} (\operatorname{div} \vec{p}^*, \operatorname{div} \vec{v}\,)_B + (\mathrm{K}^* f, \operatorname{div} \vec{v}\,)_B + \gamma (\mathrm{P}_{\mathrm{div}} \vec{p}^*, \mathrm{P}_{\mathrm{div}} \vec{v}\,) \\ + \langle \vec{\lambda}^*, \vec{v}\,\rangle_{\mathbb{H}_0^1(\Omega)^*, \mathbb{H}_0^1(\Omega)} = 0 \quad \text{for all } \vec{v} \in \mathbb{H}_0^1(\Omega). \end{aligned} \tag{3.9}$$

Since $\mathbb{H}_0^1(\Omega)$ is dense in $H_0(\mathrm{div})$ and $\vec{p}^* \in H_0(\mathrm{div})$, we have that (3.9) holds for all $\vec{v} \in H_0(\mathrm{div})$. Consequently $\vec{\lambda}^*$ can be identified with an element in $H_0(\mathrm{div})^*$, and $\langle \cdot, \cdot \rangle_{\mathbb{H}_0^1(\Omega)^*, \mathbb{H}_0^1(\Omega)}$ in (3.9) can be replaced by $\langle \cdot, \cdot \rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})}$. We next verify that $\vec{p}^*$ is feasible. For this purpose note that

$$(\vec{\lambda}_c, \vec{p} - \vec{p}_c) = \big(\max(0, c(\vec{p}_c - \beta \vec{1})) + \min(0, c(\vec{p}_c + \beta \vec{1}), \vec{p} - \vec{p}_c \big) \le 0 \tag{3.10}$$

for all $-\beta \vec{1} \le \vec{p} \le \beta \vec{1}$. From (3.1) we have

$$\frac{1}{c}|\nabla \vec{p}_c|^2 + |\operatorname{div} \vec{p}_c + \mathrm{K}^* f|_B^2 + \gamma |\mathrm{P}_{\mathrm{div}} \vec{p}_c|^2 + \frac{1}{c}|\vec{\lambda}_c|^2 \le |\mathrm{K}^* f|_B^2. \tag{3.11}$$

Consequently, $\frac{1}{c}|\vec{\lambda}_c|^2 \le |\mathrm{K}^* f|_B^2$ for all $c > 0$. Note that

$$\frac{1}{c}|\vec{\lambda}_c|_{\mathbb{L}^2(\Omega)}^2 = c|\max(0, \vec{p}_c - \beta \vec{1})|_{\mathbb{L}^2(\Omega)}^2 + c|\min(0, \vec{p}_c + \beta \vec{1})|_{\mathbb{L}^2(\Omega)}^2$$

and thus

$$\begin{aligned} |\max(0, (\vec{p}_c - \beta \vec{1}))|_{\mathbb{L}^2(\Omega)}^2 &\overset{c \to \infty}{\to} 0, \\ |\min(0, (\vec{p}_c + \beta \vec{1}))|_{\mathbb{L}^2(\Omega)}^2 &\overset{c \to \infty}{\to} 0. \end{aligned} \tag{3.12}$$

Recall that $\vec{p}_c \rightharpoonup \vec{p}^*$ weakly in $\mathbb{L}^2(\Omega)$. Weak lower semicontinuity of the convex functional $\vec{p} \mapsto |\max(0, \vec{p} - \beta\vec{1})|_{\mathbb{L}^2(\Omega)}$ and (3.12) imply that

$$\int_\Omega |\max(0, \vec{p}^* - \beta\vec{1})|^2 dx \le \liminf_{c\to\infty} \int_\Omega |\max(0, \vec{p}_c - \beta\vec{1})|^2 dx = 0.$$

Consequently, $\vec{p}^* \le \beta\vec{1}$, and analogously one verifies that $-\beta\vec{1} \le \vec{p}^*$. In particular, $\vec{p}^*$ is feasible, and from (3.10) we conclude that

$$\langle \vec{\lambda}_c, \vec{p}^* - \vec{p}_c \rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} \le 0 \quad \text{for all } c > 0. \tag{3.13}$$

By optimality of $\vec{p}_c$ for (3.1) we have

$$\limsup_{c\to\infty} \left(\frac{1}{2} |\operatorname{div} \vec{p}_c + \mathrm{K}^* f|_B^2 + \frac{\gamma}{2} |\mathrm{P}_{\mathrm{div}} \vec{p}_c|^2 \right) \le \frac{1}{2} |\operatorname{div} \vec{p} + \mathrm{K}^* f|_B^2 + \frac{\gamma}{2} |\mathrm{P}_{\mathrm{div}} \vec{p}\,|^2 \tag{3.14}$$

for all $\vec{p} \in S_2 = \{\vec{p} \in (C_0^1(\Omega))^2 : -\beta\vec{1} \le \vec{p} \le \beta\vec{1}\}$. Density of S_2 in $S_1 = \{\vec{p} \in H_0(\mathrm{div}) : -\beta\vec{1} \le \vec{p} \le \beta\vec{1}\}$ in the norm of $H_0(\mathrm{div})$ implies that (3.14) holds for all $\vec{p} \in S_1$ and consequently

$$\begin{aligned} \limsup_{c\to\infty} \left(\frac{1}{2} |\operatorname{div} \vec{p}_c + \mathrm{K}^* f|_B^2 + \frac{\gamma}{2} |\mathrm{P}_{\mathrm{div}} \vec{p}_c|^2 \right) &\le \frac{1}{2} |\operatorname{div} \vec{p}^* + \mathrm{K}^* f|_B^2 + \frac{\gamma}{2} |\mathrm{P}_{\mathrm{div}} \vec{p}^*|^2 \\ &\le \liminf_{c\to\infty} \left(\frac{1}{2} |\operatorname{div} \vec{p}_c + \mathrm{K}^* f|_B^2 + \frac{\gamma}{2} |\mathrm{P}_{\mathrm{div}} \vec{p}_c|^2 \right), \end{aligned}$$

where for the last inequality weak lower semicontinuity of norms is used. The above inequalities together with weak convergence of $\vec{p}_c$ to $\vec{p}^*$ in $H_0(\mathrm{div})$ imply strong convergence of $\vec{p}_c$ to $\vec{p}^*$ in $H_0(\mathrm{div})$. Finally we aim at passing to the limit in (3.13). This is impeded by the fact that we only established $\vec{\lambda}_c \rightharpoonup \vec{\lambda}^*$ in $\mathbb{H}_0^1(\Omega)^*$. Note from (3.2) that $\{-\frac{1}{c}\Delta\vec{p}_c + \vec{\lambda}_c\}_{c\ge 1}$ is bounded in $H_0(\mathrm{div})$. Hence there exists $\vec{\mu}^* \in H_0(\mathrm{div})^*$ such that

$$-\frac{1}{c}\Delta\vec{p}_c + \vec{\lambda}_c \rightharpoonup \vec{\mu}^* \quad \text{weakly in } H_0(\mathrm{div})^*,$$

and consequently also in $\mathbb{H}_0^1(\Omega)^*$. Moreover, $\{\frac{1}{\sqrt{c}}|\nabla\vec{p}_c|\}_{c\ge 1}$ is bounded and hence

$$-\frac{1}{c}\Delta\vec{p}_c \rightharpoonup 0 \quad \text{weakly in } \mathbb{H}_0^1(\Omega)^*$$

as $c \to \infty$. Since $\vec{\lambda}_c \rightharpoonup \vec{\lambda}^*$ weakly in $\mathbb{H}_0^1(\Omega)^*$, it follows that

$$\langle \vec{\lambda}^* - \vec{\mu}^*, \vec{v} \rangle_{\mathbb{H}_0^1(\Omega)^*, \mathbb{H}_0^1(\Omega)} = 0 \quad \text{for all } \vec{v} \in \mathbb{H}_0^1(\Omega).$$

Since both $\vec{\lambda}^*$ and $\vec{\mu}^*$ are elements of $H_0(\mathrm{div})^*$ and since $\mathbb{H}_0^1(\Omega)$ is dense in $H_0(\mathrm{div})$, it follows that $\vec{\lambda}^* = \vec{\mu}^*$ in $H_0(\mathrm{div})^*$. For $\vec{p} \in S_2$ we have

$$\begin{aligned} \langle \vec{\lambda}^*, \vec{p} - \vec{p}^* \rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} &= \langle \mu^*, \vec{p} - \vec{p}^* \rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} \\ &= \lim_{c\to\infty} \left\langle -\frac{1}{c}\Delta\vec{p}_c + \vec{\lambda}_c, \vec{p} - \vec{p}_c \right\rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} \\ &= \lim_{c\to\infty} \left(\frac{1}{c}(\nabla\vec{p}_c, \nabla(\vec{p} - \vec{p}_c)) + (\vec{\lambda}_c, \vec{p} - \vec{p}_c) \right) \\ &\le \lim_{c\to\infty} \left(\frac{1}{c}(\nabla\vec{p}_c, \nabla\vec{p}) + (\vec{\lambda}_c, \vec{p} - \vec{p}_c) \right) \le 0 \end{aligned}$$

by (3.10) and (3.11). Since S_2 is dense in S_1, we find

$$\langle \vec{\lambda}^*, \vec{p} - \vec{p}^* \rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} \leq 0 \quad \text{for all } \vec{p} \in S_1. \qquad \square$$

The problem formulation (3.1) contains two limiting processes: the $\mathbb{H}_0^1(\Omega)$ smoothing, which will be used to guarantee superlinear convergence of semismooth Newton methods applied to the first order optimality conditions (3.2) of (3.1) in function spaces (see section 4), and a penalization of the constraints $-\beta\vec{1} \leq \vec{p} \leq \beta\vec{1}$ resulting in the max- and min-terms. There is no need to utilize the same parameter c for both limiting processes. Rather, if $\frac{1}{2c}|\nabla\vec{p}|^2$ is replaced by $\frac{1}{2\bar{c}}|\nabla\vec{p}|^2$, then $(\vec{p}_{\bar{c},c}, \vec{\lambda}_{\bar{c},c})$ converges to $(\vec{p}^*, \vec{\lambda}^*)$ weakly in $H_0(\mathrm{div}) \times \mathbb{H}_0^1(\Omega)^*$, where $(\vec{p}_{\bar{c},c}, \vec{\lambda}_{\bar{c},c})$ denotes the solution of (3.2) with $\frac{1}{c}\Delta\vec{p}_c$ replaced by $\frac{1}{\bar{c}}\Delta\vec{p}_{\bar{c},c}$, as $c \to \infty$ and $\bar{c} \to \infty$.

4. Semismooth Newton methods. Here we shall describe two algorithms, one for a discretized form of (2.2) and another one for (3.1). Both algorithms are locally superlinearly convergent.

First we consider the unregularized problem (2.2). After discretization it is of the form

$$\begin{cases} \min \quad \dfrac{1}{2}|A_1 p + \tilde{f}|^2 + \dfrac{\gamma}{2}|A_2 p|^2 \\ \text{s.t.} \quad -\beta\mathbf{1} \leq p \leq \beta\mathbf{1}, \end{cases} \tag{4.1}$$

where $p \in \mathbb{R}^m$, for some $m \in \mathbb{N}$ with coordinates p_i. Further, A_1, A_2 are $m \times m$-matrices, $\tilde{f} \in \mathbb{R}^m$, and $\mathbf{1} \in \mathbb{R}^m$ denotes the vector with all entries equal to 1. We assume that $\ker A_1 \cap \ker A_2 = 0$. The optimality condition for (4.1) is given by

$$\begin{aligned} &A_1^T A_1 p + \gamma A_2^T A_2 p + A_1^T \tilde{f} + \lambda = 0, \\ &\lambda = \max(0, \lambda + c(p - \beta\mathbf{1})) + \min(0, \lambda + c(p + \beta\mathbf{1})), \end{aligned} \tag{4.2}$$

where $c > 0$ is arbitrary and fixed. The primal-dual active set strategy, or equivalently the semismooth Newton algorithm applied to (4.2), is specified next.

ALGORITHM A.

(1) *Choose* $p_0, \lambda_0 \in \mathbb{R}^m$ *and set* $k = 0$.

(2) *Define*

$$\begin{aligned} \mathcal{A}_{k+1}^+ &= \{i : (\lambda_k + c(p_k - \beta\mathbf{1}))_i > 0\}, \\ \mathcal{A}_{k+1}^- &= \{i : (\lambda_k + c(p_k + \beta\mathbf{1}))_i < 0\}, \\ \mathcal{I}_{k+1}^i &= \{i : i \notin \mathcal{A}_{k+1}^\pm\}. \end{aligned}$$

(3) *Solve for* p_{k+1}, λ_{k+1}

$$\begin{aligned} &A_1^T A_1 p_{k+1} + \gamma A_2^T A_2 p_{k+1} + A_1^T \tilde{f} + \lambda_{k+1} = 0, \\ &(\lambda_{k+1})_i = 0 \text{ for } i \in \mathcal{I}_{k+1}, \\ &(p_{k+1})_i = \beta \text{ for } i \in \mathcal{A}_{k+1}^+, \quad (p_{k+1})_i = -\beta \text{ for } i \in \mathcal{A}_{k+1}^-. \end{aligned}$$

(4) *Stop, or set* $k = k + 1$ *and go to* (2).

This algorithm can be obtained by applying a formal Newton step to (4.2), choosing as generalized derivative for the function $s \mapsto \max(0, s)$ the value 1 if $s \geq 0$ and 0 if $s < 0$, and making an analogous choice for $s \mapsto \min(0, s)$.

For the following result we suppose that, given p_0, the first equation in (4.2) is used to compute λ_0. Then we have the following result.

THEOREM 4.1. *If $|p_0 - p^*|_{\mathbb{R}^m}$ is sufficiently small, then the iterates $\{(p_k, \lambda_k)\}_{k=1}^{\infty}$ of Algorithm A converge superlinearly to the solution (p^*, λ^*) of* (4.2).

The result can be verified by standard techniques from semismooth Newton methods; see, e.g., [17]. We do not enter into the details here but rather for Algorithm B below, where they are more involved.

We turn to the algorithmic treatment of the infinite-dimensional problem (3.1), for which we propose the following algorithm.

ALGORITHM B.

(1) *Choose $\vec{p}_0 \in \mathbb{H}_0^1(\Omega)$ and set $k = 0$.*

(2) *Set, for $i = 1, 2$,*

$$\begin{aligned} \mathcal{A}_{k+1}^{+,i} &= \{x : (\vec{p}_k^{\,i} - \beta\vec{1})(x) > 0\}, \\ \mathcal{A}_{k+1}^{-,i} &= \{x : (\vec{p}_k^{\,i} + \beta\vec{1})(x) < 0\}, \\ \mathcal{I}_{k+1}^{i} &= \Omega \setminus (\mathcal{A}_{k+1}^{+,i} \cup \mathcal{A}_{k+1}^{-,i}). \end{aligned}$$

(3) *Solve for $\vec{p} \in \mathbb{H}_0^1(\Omega)$ and set $\vec{p}_{k+1} = \vec{p}$, where*

$$\begin{aligned} \frac{1}{c}(\nabla\vec{p}, \nabla\vec{v}\,) + (\operatorname{div}\vec{p}, \operatorname{div}\vec{v}\,)_B + (\mathrm{K}^* f, \operatorname{div}\vec{v}\,)_B + \gamma(\mathrm{P}_{\mathrm{div}}\vec{p}, \mathrm{P}_{\mathrm{div}}\vec{v}\,) \\ + (c(\vec{p} - \beta\vec{1}\,)\chi_{\mathcal{A}_{k+1}^{+}}, \vec{v}\,) + (c(\vec{p} + \beta\vec{1}\,)\chi_{\mathcal{A}_{k+1}^{-}}, \vec{v}\,) = 0 \end{aligned} \tag{4.3}$$

for all $\vec{v} \in \mathbb{H}_0^1(\Omega)$.

(4) *Set*

$$\vec{\lambda}_{k+1}^{i} = \begin{cases} 0 & \text{on } \mathcal{I}_{k+1}^{i}, \\ c(\vec{p}_{k+1}^{\,i} - \beta\vec{1}) & \text{on } \mathcal{A}_{k+1}^{+,i}, \\ c(\vec{p}_{k+1}^{\,i} + \beta\vec{1}) & \text{on } \mathcal{A}_{k+1}^{-,i}, \end{cases}$$

for $i = 1, 2$.

(5) *Stop, or set $k = k + 1$ and go to* (2).

In the above, $\chi_{\mathcal{A}_{k+1}^{+}}$ stands for

$$\chi_{\mathcal{A}_{k+1}^{+}}^{i} = \begin{cases} 1 & \text{if } x \in \mathcal{A}_{k+1}^{+,i}, \\ 0 & \text{if } x \notin \mathcal{A}_{k+1}^{+,i}, \end{cases}$$

and analogously for $\mathcal{A}_{k+1}^{-}$. The superscript i, $i = 1, 2$, refers to the respective component. We note that (4.3) admits a solution $\vec{p}_{k+1} \in \mathbb{H}_0^1(\Omega)$. Step (4) is included for the sake of the analysis of the algorithm. Let $C : \mathbb{H}_0^1(\Omega) \to H^{-1}(\Omega) \times H^{-1}(\Omega)$ stand for the operator

$$C = -\frac{1}{c}\Delta - \nabla B^{-1} \operatorname{div} + \gamma \mathrm{P}_{\mathrm{div}}.$$

It is a homeomorphism for every $c > 0$ and allows us to express (3.2) as

$$C\vec{p} - \nabla B^{-1}\mathrm{K}^* f + c\max(0, \vec{p} - \beta\vec{1}) + c\min(0, \vec{p} + \beta\vec{1}) = 0, \tag{4.4}$$

where we drop the index in the notation for $\vec{p}_c$. For $\varphi \in \mathrm{L}^2(\Omega)$ we define

$$\mathrm{Dmax}(0,\varphi)(x) = \begin{cases} 1 & \text{if } \varphi(x) > 0, \\ 0 & \text{if } \varphi(x) \le 0, \end{cases} \tag{4.5}$$

and

$$\mathrm{Dmin}(0,\varphi)(x) = \begin{cases} 1 & \text{if } \varphi(x) < 0, \\ 0 & \text{if } \varphi(x) \ge 0. \end{cases} \tag{4.6}$$

These operators define semismooth derivatives; see, e.g., [22] for the finite-dimensional case and [17, 24] for the infinite dimensions. Using (4.5), (4.6) as generalized derivatives for the max and the min operations in (4.4), the semismooth Newton step can be expressed as

$$C\vec{p}_{k+1} + c(\vec{p}_{k+1} - \beta\vec{1})\chi_{\mathcal{A}^+_{k+1}} + c(\vec{p}_{k+1} + \beta\vec{1})\chi_{\mathcal{A}^-_{k+1}} - \nabla B^{-1}\mathrm{K}^* f = 0, \tag{4.7}$$

and $\vec{\lambda}_{k+1}$ from step (4) of Algorithm B is given by

$$\vec{\lambda}_{k+1} = c(\vec{p}_{k+1} - \beta\vec{1})\chi_{\mathcal{A}^+_{k+1}} + c(\vec{p}_{k+1} + \beta\vec{1})\chi_{\mathcal{A}^-_{k+1}}. \tag{4.8}$$

The iteration of Algorithm B can also be expressed with respect to the variable $\vec{\lambda}$ rather than $\vec{p}$. For this purpose we define

$$F(\vec{\lambda}) = \vec{\lambda} - c\max(0, C^{-1}(\nabla\hat{f} - \vec{\lambda}) - \beta\vec{1}) - c\min(0, C^{-1}(\nabla\hat{f} - \vec{\lambda}) + \beta\vec{1}), \tag{4.9}$$

where we put $\hat{f} = B^{-1}\mathrm{K}^* f$. Setting $\vec{p}_k = C^{-1}(\nabla\hat{f} - \vec{\lambda}_k)$, the semismooth Newton step applied to $F(\vec{\lambda}) = 0$ at $\vec{\lambda} = \vec{\lambda}_k$ results in

$$\vec{\lambda}_{k+1} = c(C^{-1}(\nabla\hat{f} - \vec{\lambda}_{k+1}) - \beta\vec{1})\chi_{\mathcal{A}^+_{k+1}} + c(C^{-1}(\nabla\hat{f} - \vec{\lambda}_{k+1}) + \beta\vec{1})\chi_{\mathcal{A}^-_{k+1}},$$

which coincides with (4.8). Therefore the semismooth Newton iterations according to Algorithm B and for $F(\vec{\lambda}) = 0$ coincide, provided that the initializations are related by $C\vec{p}_0 - \nabla\hat{f} + \vec{\lambda}_0 = 0$. The mapping F is slantly differentiable; i.e., for every $\vec{\lambda} \in \mathbb{L}^2(\Omega)$

$$|F(\vec{\lambda} + \vec{h}) - F(\vec{\lambda}) - DF(\vec{\lambda} + \vec{h})h|_{\mathbb{L}^2(\Omega)} = o(|\vec{h}|_{\mathbb{L}^2(\Omega)}) \tag{4.10}$$

for $|\vec{h}|_{\mathbb{L}^2(\Omega)} \to 0$ (see [17]). Here D denotes the derivative of F defined by means of (4.5) and (4.6). For (4.10) to hold the smoothing property of C^{-1} in the sense of an embedding from $\mathbb{L}^2(\Omega)$ into $\mathbb{L}^p(\Omega)$ for some $p > 2$ is essential. The following result now follows from standard arguments.

THEOREM 4.2. *If $|\vec{\lambda}_c - \vec{\lambda}_0|_{\mathbb{L}^2(\Omega)}$ is sufficiently small, then the iterates $\{(\vec{p}_k, \vec{\lambda}_k)\}_{k=1}^{\infty}$ of Algorithm B converge superlinearly in $\mathbb{H}^1_0(\Omega) \times \mathbb{L}^2(\Omega)$ to the solution $(\vec{p}_c, \vec{\lambda}_c)$ of* (3.1).

5. Discretization and numerical examples. We report now on numerical results attained by Algorithms A and B. In the examples below we choose $K = \mathrm{I}$ and $\alpha = 0$ for image denoising. We also include results for image zooming; see, e.g., [20] for a general description. In this case, we use given data f which correspond to a coarse (low-pixel–based) approximation of a given image. Then the aim is to

reconstruct the original image at the original pixel-scale. As a consequence, $K \neq \mathrm{I}$ with $\ker(K)$ typically nontrivial, which requires us to choose $\alpha > 0$.

In our tests, for the div-operator we use backward differences with quadratic extrapolation on the left boundary for Algorithm B. For Algorithm A we use symmetric differences with quadratic extrapolation on the boundary, where A_1 denotes the discretized divergence operator. The discrete grad $-$ div-operator is taken as $A_1^T A_1$. For Algorithm B we need the discrete Laplacian with homogeneous Dirichlet boundary conditions. We use the standard five-point stencil for its discretization. The projection $P_{\mathrm{div}}\vec{p}$ is obtained by solving a Neumann problem as stated at the end of section 1. Again we use the five-point stencil for discretizing the Laplace operator with symmetric differences for the discretization of the Neumann boundary condition.

To investigate possible ill-conditioning due to the parameter c appearing in Algorithm B we also tested a first order augmented Lagrangian variant of Algorithm B. To specify the algorithm we define

$$L(\vec{p},\vec{\lambda}) = \frac{1}{2\bar{c}}|\nabla\vec{p}\,|^2 + \frac{1}{2}|\operatorname{div}\vec{p} + \mathrm{K}^* f|_B^2 + \tfrac{\gamma}{2}|\mathrm{P}_{\mathrm{div}}\vec{p}\,|^2 + \phi_c(\vec{p},\vec{\lambda}),$$

where ϕ_c is the generalized Moreau–Yosida regularization of the indicator function ϕ of the set $\{\vec{p} \in \mathbb{L}^2(\Omega) : -\beta\vec{1} \le \vec{p} \le \beta\vec{1}\}$. We have

$$\phi_c(\vec{p},\vec{\lambda}) = \inf_{\vec{q}\in\mathbb{L}^2(\Omega)} \phi(\vec{p}-\vec{q}\,) + (\vec{\lambda},\vec{q}\,)_{\mathbb{L}^2(\Omega)} + \frac{c}{2}|\vec{q}\,|^2_{\mathbb{L}^2(\Omega)}$$

for $c > 0$ and $\vec{\lambda} \in \mathbb{L}^2(\Omega)$. Some simple manipulations result in

$$\begin{aligned}\phi_c(\vec{p},\vec{\lambda}) = \frac{1}{2c}\big|\max\big(0,\vec{\lambda}+c(\vec{p}-\beta\vec{1})\big)\big|^2_{\mathbb{L}^2(\Omega)} \\ + \frac{1}{2c}\big|\min\big(0,\vec{\lambda}+c(\vec{p}+\beta\vec{1})\big)\big|^2_{\mathbb{L}^2(\Omega)} - \frac{1}{2c}|\vec{\lambda}|^2_{\mathbb{L}^2(\Omega)}.\end{aligned}$$

Augmented Lagrangian Method (ALM).

(1) *Choose* $\vec{\lambda}_0 \in \mathbb{L}^2(\Omega)$, $c > 0$, *and* $n = 0$.

(2) *Given* $\vec{\lambda}_n \in \mathbb{L}^2(\Omega)$, *determine*

$$\vec{p}_n = \operatorname{argmin}\{L(\vec{p},\vec{\lambda}_n) : \vec{p} \in \mathbb{L}^2(\Omega)\}.$$

(3) *Update* $\vec{\lambda}_n$ *by* $\vec{\lambda}_{n+1} = \phi_c'(\vec{p}_n,\vec{\lambda}_n)$.

(4) *If convergence is not achieved, set* $n = n+1$ *and go to step* (2).

In step (3) we have

$$\phi_c'(\vec{p},\vec{\lambda}) = \max\big(0,\vec{\lambda}+c(\vec{p}-\beta\vec{1})\big) + \min\big(0,\vec{\lambda}+c(\vec{p}+\beta\vec{1})\big).$$

Note that the auxiliary problems in step (2) of ALM coincide with (3.1) except for the shift by $\vec{\lambda}_n$ in the max/min operations. In our numerical tests below we typically choose $\bar{c} = c$.

The algorithms in sections 3 and 4 are stated in terms of exact system solutions. Our numerical implementation utilizes inexact Newton techniques to underscore the feasibility of the proposed methods for large scale problems. In order to describe our approach let r_k denote the residual of the respective system, i.e., (4.2) for Algorithm A and (4.3) for Algorithm B. We resolve the respective system with the preconditioned conjugate gradient method (CG-method). The preconditioner involves the (vector)

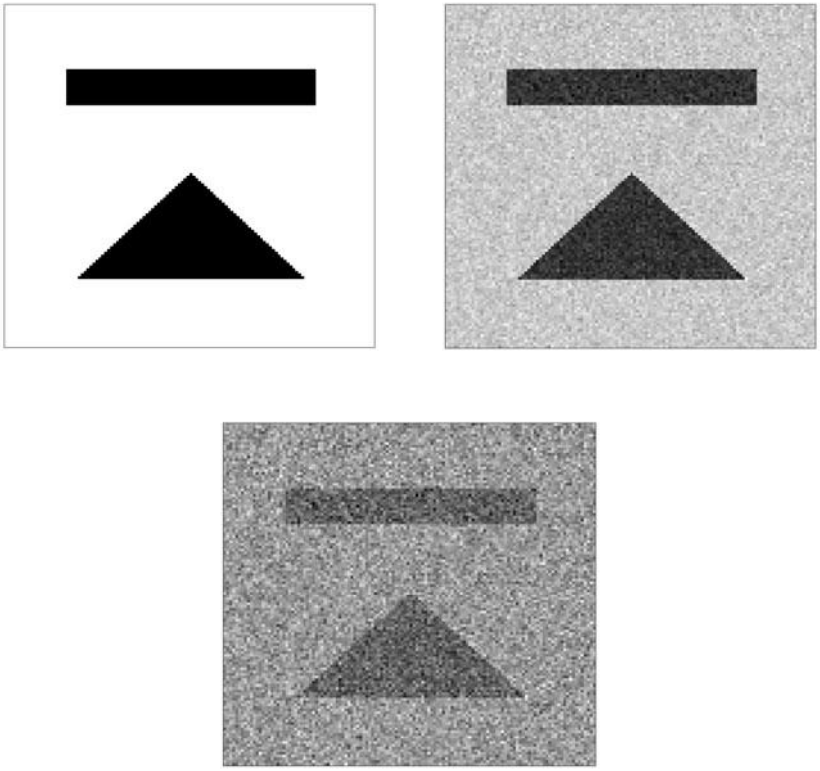


FIG. 1. *Upper left: original image (*128×128 *pixel). Upper right: noisy image with* 10% *noise. Lower: noisy image with* 50% *noise.*

Laplacian and, for Algorithm B, the terms involving the indicator functions of $\mathcal{A}^{\pm}_{k+1}$. The stopping tolerance for the CG-method in iteration $k+1$ is given by

$$\text{tol}_{k+1} = 0.1 \min(r_k^{1.25}, r_k).$$

This choice is motivated by the locally superlinear convergence rate of our algorithms.

Example 1. The test images for our first image denoising example are displayed in Figure 1. The upper left image is the original image, which is similar to the one in [8]. It has a dynamic range of $[0, 255]$. The other two images contain Gaussian white noise. The upper right one has 10% noise, and the remaining image contains 50% noise; i.e., we add Gaussian noise with standard deviation of 25.5 and 127.5, respectively. In the subsequent tables we denote by #as the total number of active set iterations, by #cg the total number of CG-iterations, and by #alm the total number of iterations updating $\vec{\lambda}_n$ for ALM. We stopped each algorithm as soon as the discrete L_2-norm of the residual dropped below `tol`$= \sqrt{\epsilon_M}$, with ϵ_M the machine precision, or when the difference between two successive residuals was smaller than `tol`, i.e., no further progress was observed.

Let us first report on the results obtained for denoising the image with 10% noise. For all algorithms we choose $c = 1E4$. However, let us note that Algorithm A does not require large c since c is not linked to a regularization term. Rather it is a parameter associated with the reformulation of the complementarity system induced by the box

TABLE 5.1
Results for 10% *noise.*

Algorithm	#as	#cg	#alm
Algorithm A	16	64	-
Algorithm B	9	23	-
ALM	14	27	3

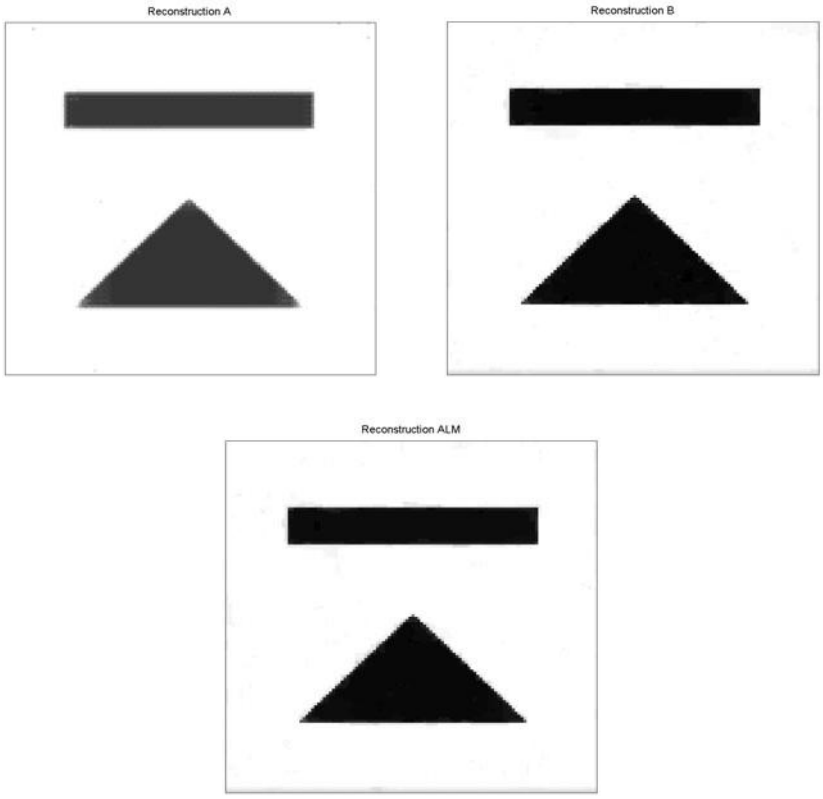


FIG. 2. *Upper left: result of Algorithm A. Upper right: result of Algorithm B. Lower: result of ALM.*

constraints. Further, for all three algorithms we chose $\beta = 0.2$, $\gamma = 0$ for ALM and Algorithm B, and $\gamma = 1E\text{-}3$ for Algorithm A. In general, for ALM and Algorithm B, γ had no noticeable effect on the results attained. However, Algorithm A is more sensitive to γ. This can be attributed to the fact that the system matrix in Algorithm A is singular for $\gamma = 0$. In Table 5.1 we report on the iteration numbers for the respective algorithms.

We note that Algorithm B requires the least number of AS-iterations. For ALM we point out that we initialized it with $\vec{\lambda}_0 \equiv 0$; then, typically, 8–10 AS-iterations were required in the first ALM-iteration. The subsequent ALM-iterations needed 2–3 AS-iterations.

In Figure 2 we display the reconstructions. The upper left and right correspond to Algorithms A and B. The lower image is the result obtained by Algorithm ALM. The quality of the reconstructions is equally good for all algorithms.

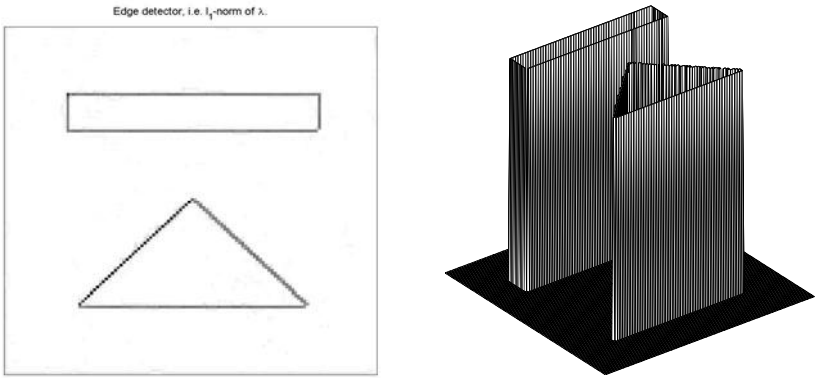


FIG. 3. *Left: Lagrange multiplier of Algorithm B. Right: corresponding edge detector.*

TABLE 5.2
Results for 50% *noise.*

Algorithm	#as	#cg	#alm
Algorithm A	15	59	-
Algorithm B	7	18	-
ALM	13	33	3

In the introduction we mentioned that the Lagrange multiplier associated with the box constraints serves as an edge detector. Figure 3 shows the ℓ_1-norm of the multiplier attained by Algorithm B and a resulting edge detector. The edge detector is obtained from a simple thresholding technique. In fact, as a threshold we took c and computed the edge detector λ_e as

$$\lambda_e(x_i) = \begin{cases} 1 & \text{if } |\vec{\lambda}^*(x_i)|_{\ell_1} \geq c, \\ 0 & \text{otherwise.} \end{cases}$$

In the above, x_i denotes the ith pixel of the image, and $\vec{\lambda}^*$ the multiplier upon termination of B. For the multipliers resulting from Algorithms A and ALM a similar observation holds true.

Now we turn to the results for the image containing 50% noise. The parameters had the values $c = 1E4$, $\beta = 0.9$, and $\gamma = 0$ for ALM and Algorithm B, and $c = 1E4$, $\beta = 0.75$, $\gamma = 1E$-3 for Algorithm A. Figure 4 shows the reconstructions obtained from our algorithms. As in the previous test case, the quality of the results for Algorithms B and ALM is comparable. Algorithm A appears to be slightly more sensitive to noise. This behavior could not be ruled out by tuning the parameters c, γ, and β. The iteration numbers are reported on in Table 5.2. As can be seen from these results, the number of iterations of the respective algorithm is rather stable with respect to the noise level.

In Figure 5 we display the ℓ_1-norm of the multiplier $\vec{\lambda}^*$ upon termination of Algorithm B. The related edge detector, which is obtained in the same way as explained previously, is given in the right image of Figure 5. We conclude that—without any thresholding—the Lagrange multiplier may act as an edge detector.

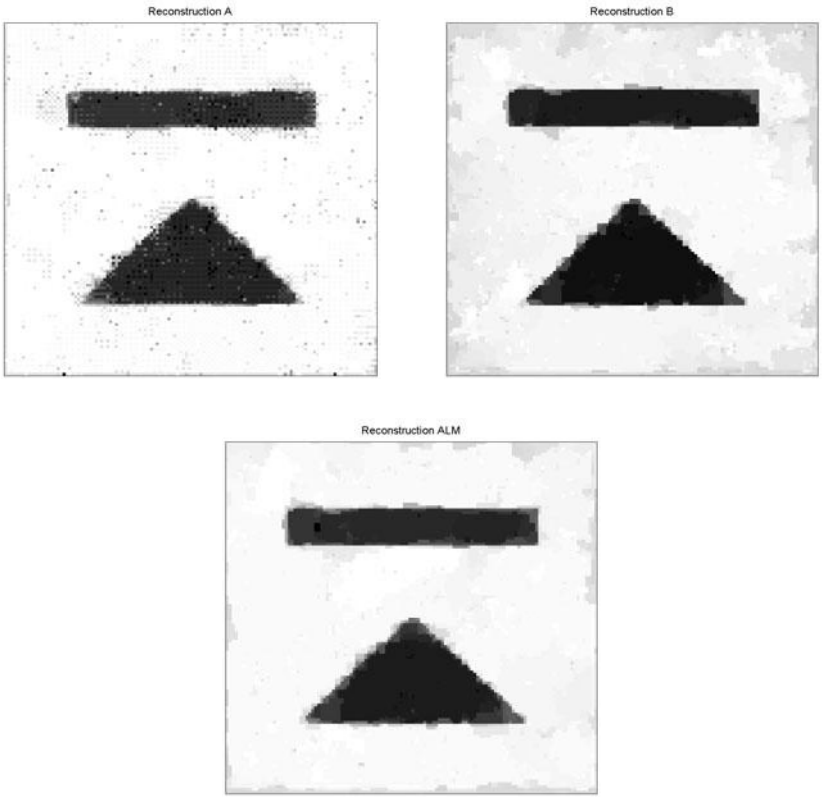


FIG. 4. *Upper left: result of Algorithm A. Upper right: result of Algorithm B. Lower: result of ALM.*

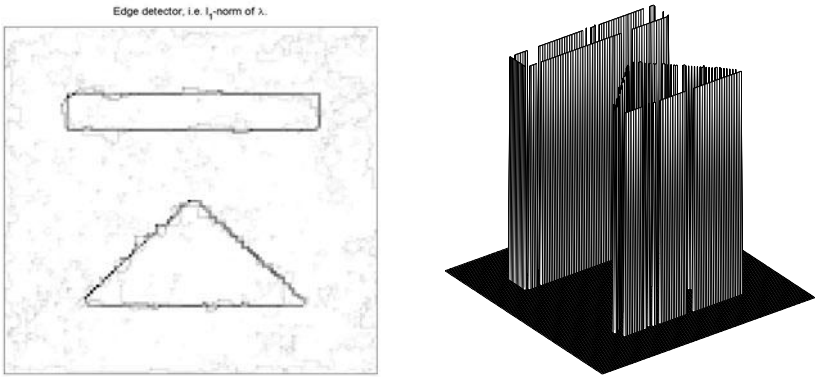


FIG. 5. *Left: Lagrange multiplier of Algorithm B. Right: corresponding edge detector.*

Let us briefly comment on the difference of stability with respect to β of Algorithms A and B compared to ALM. In general the choice of β influences the quality of the reconstruction. A large value for β decreases the number of active pixels, i.e., pixels at which $\vec{p}^*$ hits either the upper or the lower bound. As a consequence, details of the image are missed in the reconstruction. The right image in Figure 6 corresponds to the result attained by Algorithm B with $\beta = 1.5$ (compared to $\beta = 1.25$ in the

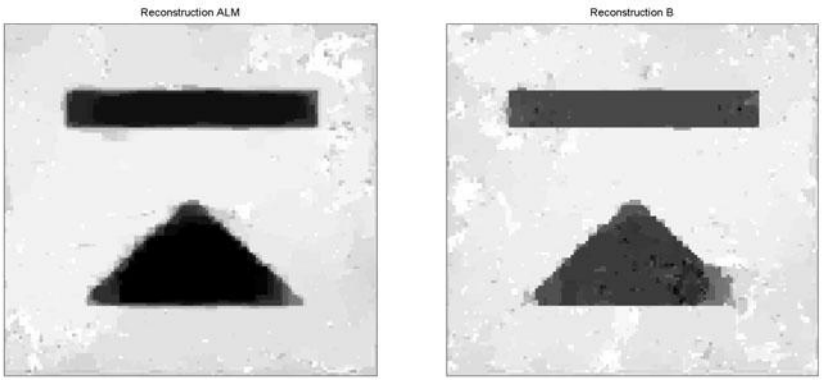


FIG. 6. *Reconstruction for $\beta = 1.5$. Left: ALM. Right: Algorithm B.*

TABLE 5.3
Convergence behavior of the residual.

#as	4.67E6	1.5E-1	9.6E-4	6.2E-7	1.8E-8	9.15E-9

previous run). Due to the larger β-value, the quality of the reconstruction degrades in the sense that details are missed, e.g., at the corners of the triangle. On the other hand, the left image in Figure 6 shows the result for ALM with $\beta = 1.5$.

Obviously the reconstruction is superior to the one obtained from Algorithm B. This reflects a general observation from our test runs, i.e., ALM is more stable with respect to the choice of β. The behavior of Algorithm A with respect to changes in β is comparable to that of Algorithm B.

Let us discuss the convergence behavior of our algorithms in terms of reductions of the residuals. From the results reported in Tables 5.1 and 5.2 we find that our algorithms require a rather small number of iterations which are even stable with respect to different noise levels. In Table 5.3 we show the behavior of the residual for Algorithm B for 50% noise, indicating a fast convergence. This fast convergence is also true for the numerical resolution of the auxiliary problem of ALM. A similar convergence behavior is obtained for Algorithm A. For smaller values of β the iterates converge superlinearly. Small values of β, however, imply a deterioration of the reconstruction. Here the ill-posedness in the problem becomes evident.

In [8] an inexact Newton method for solving a primal-dual formulation of the Euler–Lagrange equations associated with a regularized TV-based image reconstruction problem is proposed. The test problem in [8] involves the same geometry as in our test example. In Figure 7 (upper left plot) we show the noisy image containing Gaussian white noise with variance $\sigma^2 \approx 1200$, which gives a signal-to-noise ratio of approximately 1. This parallels the test setting in [8]. We also made an effort to adjust the stopping rule of Algorithm B for the comparison with the algorithm in [8]. Algorithm B requires nine iterations for obtaining the denoised image in the upper right plot of Figure 7. The algorithm in [8] with a line search and a continuation strategy with respect to δ in the regularization of the TV-seminorm of the type (1.2) is reported to need 12 iterations. The size of the systems which have to be solved per iteration in both algorithms is comparable. The edged detector, based on the ℓ_1-norm of $\vec{\lambda}$ upon termination of Algorithm B, is given in the last subplot of Figure 7.

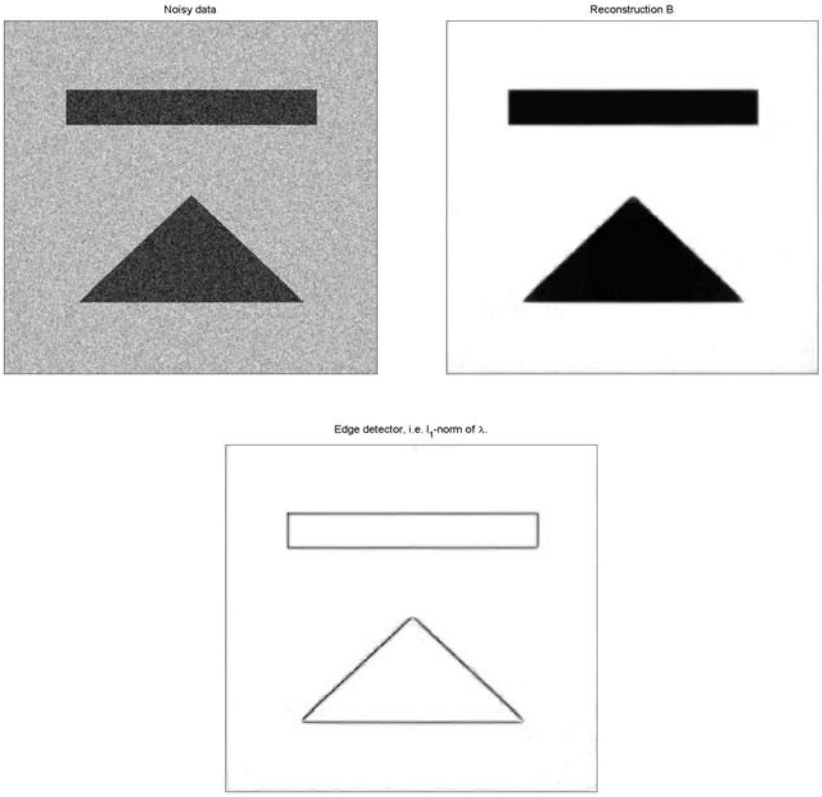


FIG. 7. *Upper left: noisy image (*256×256 *pixel). Upper right: result of Algorithm B. Lower:* ℓ_1*-norm of* $\vec{\lambda}$.

Example 2. Now we report on the behavior of Algorithm B for the benchmark problem in Figure 8. The upper left plot shows the original image. The upper right image contains 7.5% Gaussian white noise. The parameters had the values $c = 1E4$, $\gamma = 0$, and $\beta = 0.15$. The algorithm stopped after nine AS-iterations (31 CG-iterations total) with a residual of 6.2E-9. The corresponding reconstruction is given in the lower left plot of Figure 8. The lower right plot displays the ℓ_1-norm of the Lagrange multiplier associated with the box constraints. As in the previous examples, it behaves like an edge detector.

Example 3. We conclude our numerical section with the results obtained by Algorithm B for an image zooming/resizing problem. In this case, we have $K \neq \mathrm{I}$. The data f correspond to a *coarse* version of the original image satisfying $f_{2i-1,2j-1} = f_{2i,2j-1} = f_{2i-1,2j} = f_{2i,2j}$. For an arbitrary 256×256-pixel image u the application $v = Ku$ is related to a 128×128-pixel version $\tilde{v}$ of the image with $\tilde{v}_{i,j} = u_{2i-1,2j-1}$ and $v_{2i-1,2j-1} = v_{2i,2j-1} = v_{2i-1,2j} = v_{2i,2j} = \tilde{v}_{i,j}$. For more details on image zooming involving more advance operators K, we refer to [20]. Our aim is to use Algorithm B for reconstructing the fine image u from the given coarse image f. Since K has a nontrivial kernel, we choose $\alpha = 1\text{E-}10$. Further, we pick the parameter values $c = 1\text{E}5$, $\beta = 0.35$, and $\gamma = 0$. In Figure 9 we display the original image in the upper left plot. The result after 16 iterations of Algorithm B is shown in the upper right plot. The lower left plot shows the 128×128-pixel version expanded by a factor of 2, and

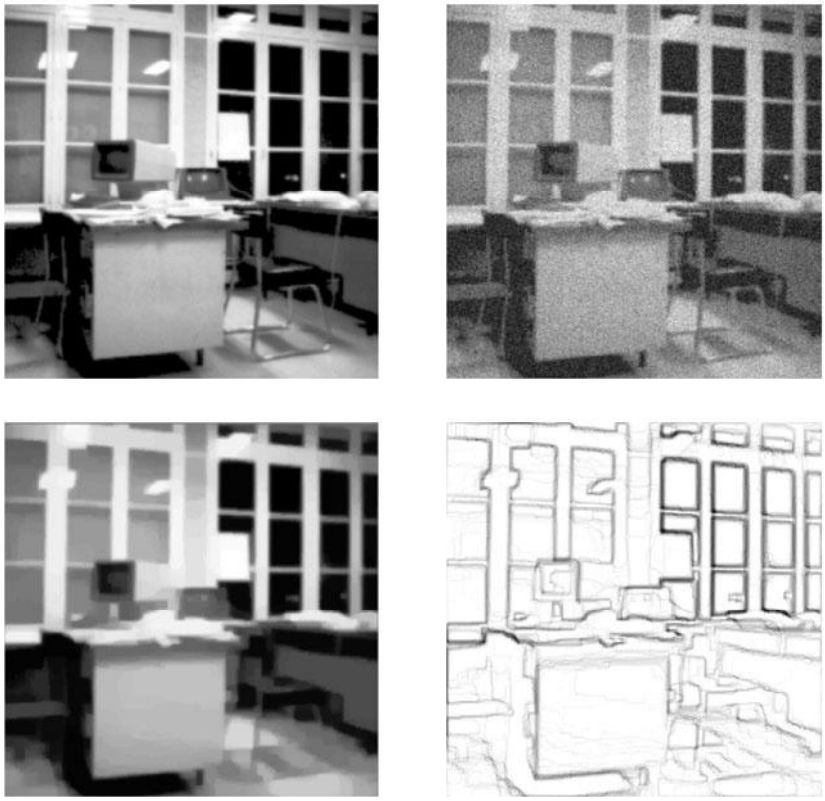


FIG. 8. *Upper left: exact data (*256×256 *pixel). Upper right: noisy data. Lower left: reconstruction obtained by Algorithm B. Lower right:* ℓ_1*-norm of the Lagrange multiplier.*

the lower right plot provides the result obtained by a nearest neighbor interpolation. Observe that the reconstructions differ quite noticeably along the boundaries of the person's left arm, for example.

6. Conclusions. The efficient numerical treatment of BV-regularization–based image restoration poses many challenges in theory as well as in the design of algorithms. In this paper we first establish the relationship between the primal problem in the nonreflexive Banach space BV and its predual which is posed in the Hilbert space $H_0(\mathrm{div})$. This analytical result appears to be of interest in its own right. We then introduce and study two semismooth Newton methods for solving the Fenchel predual problem of the underlying BV-regularized minimization problem. By predualization we obtain a box constrained minimization problem which—from the numerical optimization point of view—has the advantage that we can rely on sophisticated minimization algorithms. The convergence analysis of our semismooth Newton methods in function spaces relies on a smoothing procedure. The regularizing effect of our smoothing results in a two-norm property which is required for arguing locally superlinear convergence of our semismooth Newton method in an L^2-setting. Without smoothing we obtain a locally superlinearly convergent method on the discrete level.

Acknowledgment. We would like to thank Prof. O. Scherzer, University of Innsbruck, Austria, for making us aware of the taut string algorithm discussed in Remark 1.

FIG. 9. *Upper left: exact data (*256×256 *pixel). Upper right: reconstruction obtained by Algorithm B. Lower left:* 128×128*-pixel image expanded by a factor of* 2*. Lower right: result using a nearest neighbor interpolation technique.*

REFERENCES

[1] R. ACAR AND C. VOGEL, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems, 10 (1994), pp. 1217–1229.

[2] G. AUBERT AND L. VESE, *A variational method in image recovery*, SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[3] V. BARBU AND T. PRECUPANU, *Convexity and Optimization in Banach Spaces*, 2nd rev. and extended ed., Mathematics and Its Applications (East European Series) 10, translated from the Romanian, D. Reidel, Dordrecht/Boston/Lancaster, Editura Academiei, Bucuresti, 1986.

[4] P. CARBONNIER, L. BLAUD-FÉRAND, G. AUBERT, AND M. BARLAUD, *Deterministic edge-preserving regularisation in computed imaging*, IEEE Trans. Image Process., 6 (1997), pp. 298–311.

[5] E. CASAS, K. KUNISCH, AND C. POLA, *Regularization by functions of bounded variation and applications to image enhancement*, Appl. Math. Optim., 40 (1999), pp. 229–257.

[6] F. CATTÉ, P.-L. LIONS, J.-M. MOREL, AND T. COLL, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM J. Numer. Anal., 29 (1992), pp. 182–193.

[7] A. CHAMBOLLE AND P.-L. LIONS, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.

[8] T. F. Chan, G. H. Golub, and P. Mulet, *A nonlinear primal-dual method for total variation-based image restoration*, SIAM J. Sci. Comput., 20 (1999), pp. 1964–1977.
[9] G. Chavent and K. Kunisch, *Regularization of linear least squares problems by total bounded variation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 359–376.
[10] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol.* 3*: Spectral Theory and Applications*, with the collaboration of Michel Artola and Michel Cessenat, translated from the French by John C. Amson, Springer-Verlag, Berlin, 2000.
[11] P. L. Davies and A. Kovac, *Local extremes, runs, strings and multiresolution*, Ann. Statist., 29 (2001), pp. 1–65.
[12] D. C. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1198.
[13] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, (corrected republication of the 1976 English original) Classics in Appl. Math. 28, SIAM, Philadelphia, 1999.
[14] D. Geman and C. Yang, *Nonlinear image recovery with half-quadratic regularization*, IEEE Trans. Image Process., 4 (1995), pp. 932–945.
[15] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations. Theory and Algorithms*, Springer Ser. Comput. Math. 5, Springer-Verlag, Berlin, 1986.
[16] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Monogr. Math. 80, Birkhäuser, Boston, Basel, Stuttgart, 1984.
[17] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.
[18] K. Ito and K. Kunisch, *An active set strategy based on the augmented Lagrangian formulation for image restoration*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1–21.
[19] K. Ito and K. Kunisch, *BV-type regularization methods for convoluted objects with edge, flat and grey scales*, Inverse Problems, 16 (2000), pp. 909–928.
[20] F. Malgouyres and F. Guichard, *Edge direction preserving image zooming: A mathematical and numerical analysis*, SIAM J. Numer. Anal., 39 (2001), pp. 1–37.
[21] E. Mammen and S. van de Geer, *Locally adaptive regression splines*, Ann. Statist., 25 (1997), pp. 387–413.
[22] L. Q. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
[23] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[24] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–841.
[25] C. R. Vogel, *Computational Methods for Inverse Problems*, Frontiers Appl. Math. 23, SIAM, Philadelphia, 2002.
[26] C. R. Vogel and M. E. Oman, *Fast, robust total variation-based reconstruction of noisy, blurred images*, IEEE Trans. Image Process., 7 (1998), pp. 813–824.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1334–1359

CONSTRUCTING MULTIPLY CONNECTED QUADRATURE DOMAINS*

DARREN CROWDY† AND JONATHAN MARSHALL†

Abstract. Multiply connected bounded quadrature domains, with finite connectivity, are reconstructed from their quadrature data using conformal mappings that are ratios of products of Schottky–Klein prime functions. This method provides the natural generalization of the conformal maps to simply and doubly connected quadrature domains constructed by the first author in a number of physical applications. The efficacy of the method is demonstrated by the explicit construction of a range of examples as well as by comparison with alternative constructive methods recently introduced by Crowdy [*R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.*, 457 (2001), pp. 2337–2359] and Richardson [*European J. Appl. Math.*, 12 (2001), pp. 571–599].

Key words. quadrature domains, multiply connected, conformal mapping

AMS subject classification. 30Dxx

DOI. 10.1137/S0036139903438545

1. Introduction.

The mathematical theory of quadrature domains is well developed (e.g., [1], [2], [3], [4], [5], [6]). The simplest example of a quadrature domain is a circular disc. Let $z = x + iy$ and suppose that the disc is centered at the origin $z = 0$ with radius r. The well-known "mean value theorem" says that, if $h(z)$ is any function analytic in the disc D and integrable over it, then

$$\int\int_D h(z)dxdy = \pi r^2 h(0). \tag{1}$$

Equation (1) is a simple example of a *quadrature identity*. The idea of quadrature domain theory is to consider more complicated domains satisfying more complicated quadrature identities. Consider a planar domain D and let $h(z)$ be any function that is analytic in D and integrable over it. Suppose that

$$\int\int_D h(z)dxdy = \sum_{k=1}^{N}\sum_{j=0}^{n_k-1} c_{jk}h^{(j)}(z_k), \tag{2}$$

where $\{z_k \in \mathbb{C}\}$ is a set of points strictly inside D, $\{c_{jk} \in \mathbb{C}\}$, and $h^{(j)}(z)$ denotes the jth derivative of h. Here, N and $\{n_k \geq 1\}$ are integers. Then D is known as a *quadrature domain*. The quadrature identity (2) generalizes (1).

While quadrature domains are mathematically interesting in their own right, perhaps more remarkable is the fact that they are relevant to the mathematical study of a wide range of physical problems. An important early paper of Richardson [7] was the first to illustrate the connection with the study of the free boundary problem involving singularity-driven flows in a Hele–Shaw cell. Richardson's paper involved

*Received by the editors February 26, 2003; accepted for publication (in revised form) December 4, 2003; published electronically May 20, 2004. This work was supported by EPSRC. The research of the first author was supported by the National Science Foundation (grants DMS-9803167 and DMS-9803358) and the Nuffield Foundation. The research of the second author was supported by the Engineering and Physical Sciences Research Council.

http://www.siam.org/journals/siap/64-4/43854.html

†Department of Mathematics, Imperial College, 180 Queen's Gate, London, SW7 2AZ United Kingdom (d.crowdy@imperial.ac.uk, jonathan.marshall@imperial.ac.uk).

simply connected fluid domains. The essential result is that quadrature domains are preserved by the dynamics of the physical problem. Since then, quadrature domains have been found to be useful in a variety of different problems. For example, Entov, Etingof, and Kleinbock [8] have discussed a number of generalizations of the singularity-driven Hele–Shaw problem, including the dynamics of flows in a rotating Hele–Shaw cell and the problem of "squeeze flow" in a Hele–Shaw cell. Both problems also preserve quadrature domains [9], [10]. Crowdy [11] has pointed out the relevance of quadrature domains to a biharmonic-governed (as opposed to the harmonic-governed Hele–Shaw model) free boundary problem involving slow viscous flows driven by surface tension. Here, in certain circumstances, the dynamics is also such as to preserve quadrature domains. Outside the realm of free boundary problems, it has also been shown [12], [13] that quadrature domains have relevance to the study of multipolar vortical equilibria of the two-dimensional Euler equations governing the inviscid flow of an ideal fluid. The problem of finding equilibrium shapes of free boundaries involving irrotational Euler flows with surface tension (see, e.g., [14]) can also be interpreted in terms of quadrature domain theory.

This compendium of different physical applications suggests a need to be able to construct quadrature domains of various finite connectivities. Gustafsson [5] has shown that construction of an N-connected quadrature domain is equivalent to the construction of a conformal mapping, which is a meromorphic function on a Riemann surface of genus $N-1$. For $N=1$ and $N=2$, this is possible using the theory of rational functions and elliptic functions (or, equivalently, loxodromic functions, which are naturally related to elliptic functions [15]). Indeed, these two cases constitute most of the existing literature. Richardson used rational function conformal mappings in his original paper [7] and, more recently, elliptic function conformal mappings for singularity-driven Hele–Shaw flows of doubly connected fluid regions [16]. Crowdy [9] has used loxodromic functions to derive exact solutions for the evolution of doubly connected domains in a rotating Hele–Shaw cell, providing a mathematical model for some recent experimental results involving a fluid annulus [17]. Elliptic/loxodromic function theory has also been used to construct exact solutions to the problem of the surface tension–driven Stokes flow of doubly connected fluid regions [18], [19], [20].

For higher connectivities, the situation is much more challenging. The subject of constructing multiply connected quadrature domains (of connectivity greater than 2) has been the focus of much recent activity. Two new methods of construction have recently been proposed in the context of specific applications. The first author [13] has implemented a construction based on the fact that the boundaries of quadrature domains are algebraic curves. This method has been successfully applied, for example, to the construction of vortical equilibria of the Euler equations [13] and to the squeeze flow problem in a Hele–Shaw cell [10]. Meanwhile, in considering the related problem of singularity-driven flow of multiply connected fluid domains (with zero surface tension) in a Hele–Shaw cell, Richardson [23] has proposed a different method based on conformal mapping. In this paper we present a new method which, like Richardson's, is based on conformal maps. However, our approach is different. In light of all the recent work on this problem, we also discuss in detail how the new construction differs from other methods, and how it compares to them in terms of practical application.

To motivate the present work, we recall that it is a standard result [1] that simply connected quadrature domains can be constructed by rational function mappings from a unit ζ-disc to the domain. Any rational function with given zeros and poles can be written as a ratio of products of the fundamental function $P(\zeta)=1-\zeta$. For example,

if the conformal map has poles $\{\alpha_k | k = 1, \ldots, N\}$ and zeros $\{\beta_k | k = 1, \ldots, N\}$, it can be written

$$z(\zeta) = R \frac{\prod_{k=1}^{N} P(\zeta \beta_k^{-1})}{\prod_{k=1}^{N} P(\zeta \alpha_k^{-1})}, \tag{3}$$

where R is a constant. Furthermore, it is also known that a representation of a general conformal mapping (again with poles $\{\alpha_k | k = 1, \ldots, N\}$ and zeros $\{\beta_k | k = 1, \ldots, N\}$) from the annulus $\rho < |\zeta| < 1$ to a doubly connected quadrature domain can be written exactly as in (3) but with the fundamental function $P(\zeta, \rho)$ defined differently as

$$P(\zeta, \rho) = (1 - \zeta) \prod_{k=1}^{\infty} (1 - \rho^{2k} \zeta)(1 - \rho^{2k} \zeta^{-1}). \tag{4}$$

Note that when $\rho = 0$, (4) reduces to the function $P(\zeta) = (1 - \zeta)$ relevant for the construction of the rational functions in the simply connected case. In light of this, it is natural to ask whether the representation (3) can *also* be used for quadrature domains with connectivity greater than two but with suitably generalized "fundamental functions." This is the question addressed in this paper. The generalized "fundamental functions" needed are known as the Schottky–Klein prime functions [21].

The treatment in this paper is based on the presentation in Chapter 12 of a monograph by Baker [21]. Our aim here is to show how to apply these general results for the specific purpose of constructing multiply connected quadrature domains. For clarity, any general results needed are stated without proof and in modified form suited to present purposes. The interested reader is referred to [21] for more details.

Richardson's constructive method also employs the Schottky model and mappings from the circular domains used here, but his representation of the conformal maps is different. Richardson does not use, or define, the Schottky–Klein prime function. Instead, his conformal maps are constructed as ratios of Poincaré series—a method of constructing meromorphic functions on compact Riemann surfaces described, for example, by Beardon [22]. The present authors believe the new construction based on the Schottky–Klein prime function presented herein to be an important alternative to Richardson's method for two reasons. First, it is the natural generalization of the representation (4) used by Crowdy [20] and, moreover, it is closely related to a representation in terms of ratios of products of generalized theta functions defined on Riemann surfaces [21], [25]. Note that (4) *is* the Schottky–Klein prime function in the genus-1 case. Second, we have found the present method to be easier to implement than Richardson's method both conceptually and practically. The majority of the example domains in this paper have been constructed using both methods. In all cases considered, the boundaries of the domains are indistinguishable even at very low orders of truncation.

2. Quadrature domains. Let D denote a bounded $g+1$-connected quadrature domain. It is known [5] that the conformal mapping from a conformally equivalent region (in, say, a parametric ζ-plane) to D is given by a meromorphic function on a Riemann surface of genus g. This Riemann surface can be identified with the *Schottky double* of the region D [5]. These conformal mapping functions will be explicitly constructed here as ratios of products of Schottky–Klein prime functions [21]. Such functions are defined in section 4. In order to define them, it is necessary to introduce Schottky groups [22], [24]; these are discussed in section 3. In what follows, we first show how Schottky groups are relevant to multiply connected quadrature domains.

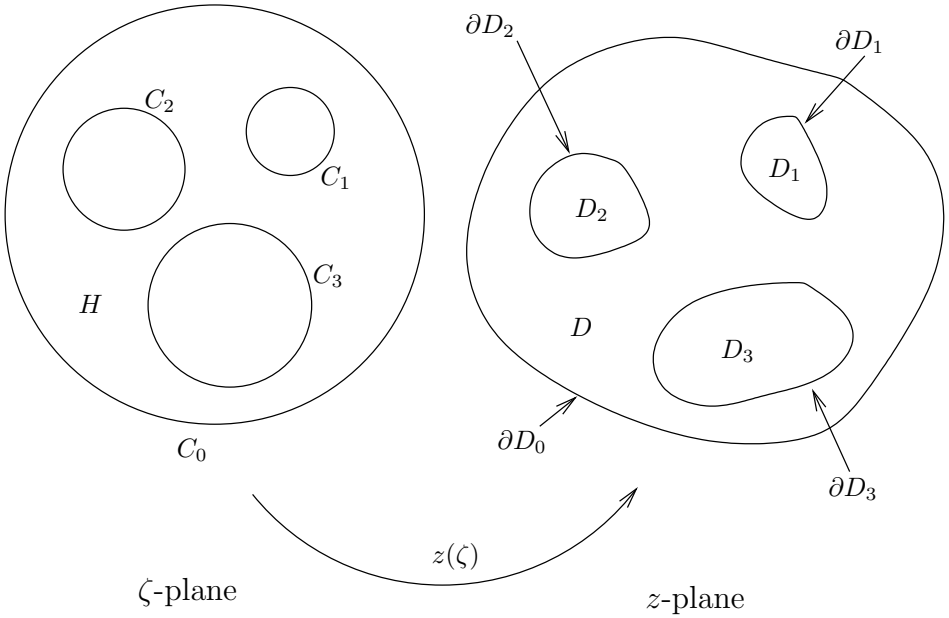


FIG. 1. *Schematic of conformal mapping from region H in the ζ-plane to region D in the physical z-plane.*

Consider the integral

$$\int\int_D h(z)d\bar{z}\wedge dz, \tag{5}$$

where $h(z)$ is analytic in D and $d\bar{z}\wedge dz = 2idxdy$ if $z = x + iy$. If D is a quadrature domain, then

$$\int\int_D h(z)d\bar{z}\wedge dz = \sum_{k=1}^{N}\sum_{j=0}^{N_k-1} c_{jk}h^{(j)}(z_k) \tag{6}$$

for some set of complex numbers $\{c_{jk}, z_k\}$, where the points $\{z_k\}$ are strictly inside D. $\{N_k \geq 1 | k = 1,\ldots,N\}$ are a set of integers, and $\sum_{k=1}^{N} N_k$ is known as the *order* of the quadrature identity. Using Green's theorem,

$$\int\int_D h(z)d\bar{z}\wedge dz = \oint_{\partial D_0} h(z)\bar{z}dz - \sum_{j=1}^{g}\oint_{\partial D_j} h(z)\bar{z}dz, \tag{7}$$

where ∂D_0 denotes the outer boundary of the bounded quadrature domain and $\partial D_i,\ i = 1,\ldots,g$, denotes the boundaries of the g enclosed holes.

Now let us introduce a conformal mapping $z(\zeta)$ to the domain D from a region H in a parametric ζ-plane bounded by the unit ζ-circle and a set of g smaller nonoverlapping circles totally contained inside $|\zeta| = 1$. Such a region shall be referred to as a *circular region*. Let the unit circle be denoted C_0, and let the g enclosed circles be labeled $C_i,\ i = 1,\ldots,g$, with centers δ_i and radii ρ_i. The circle C_0 will map to the outer boundary ∂D_0, while the circle C_i maps to the boundary ∂D_i. A schematic is shown in Figure 1. Note that, by the assumed reflectional symmetry about the

real axis of the domains considered here, the conjugate conformal map, defined by $\bar{z}(\zeta) \equiv \overline{z(\bar{\zeta})}$, satisfies

$$\bar{z}(\zeta) = z(\zeta). \tag{8}$$

Using this conformal mapping function, the integral in (7) becomes

$$\int\int_D h(z)d\bar{z} \wedge dz = \oint_{C_0} h(z(\zeta))\bar{z}(\bar{\zeta})z_\zeta(\zeta)d\zeta - \sum_{j=1}^{g} \oint_{C_j} h(z(\zeta))\bar{z}(\bar{\zeta})z_\zeta(\zeta)d\zeta. \tag{9}$$

On C_0,

$$\bar{\zeta} = \zeta^{-1}, \tag{10}$$

so that, written as a function of ζ, the first integrand on the right-hand side of (9) is

$$h(z(\zeta))\bar{z}(\zeta^{-1})z_\zeta(\zeta). \tag{11}$$

Let us now consider what is necessary in order that the sum of *all* the integrals on the right-hand side of (9) reduces to a *single* integral of the *same* integrand (11) around the entire boundary of H. For this to happen, it is necessary that

$$\bar{z}(\phi_j(\zeta)) = \bar{z}(\zeta^{-1}), \qquad j = 1, \ldots, g, \tag{12}$$

where, on C_j,

$$\bar{\zeta} = \phi_j(\zeta) \equiv \bar{\delta}_j + \frac{\rho_j^2}{\zeta - \delta_j}. \tag{13}$$

Therefore, defining

$$\theta_j(\zeta) \equiv \overline{\phi}_j(\zeta^{-1}) = \delta_j + \frac{\rho_j^2 \zeta}{1 - \bar{\delta}_j \zeta}, \tag{14}$$

we require that the conformal mapping $z(\zeta)$ satisfy

$$z(\zeta) = z(\theta_j(\zeta)), \qquad j = 1, \ldots, g. \tag{15}$$

The g maps $\{\theta_j\}$ are Mobius maps and generate a free group of transformations known as a Schottky group [21], [24]. See section 3 to follow. The mapping $z(\zeta)$ must be invariant with respect to the substitutions of this group.

For $j = 1, 2, \ldots, g$, let C_j' be the circle obtained by reflection of the circle C_j in the unit circle $|\zeta| = 1$ (i.e., the circle obtained by the transformation $\zeta \mapsto 1/\bar{\zeta}$). C_j' lies in the region exterior to the unit ζ-circle. The image of the circle C_j' under the transformation θ_j is the circle C_j. Since the g circles $\{C_j\}$ are nonoverlapping, so are the g circles $\{C_j'\}$. Consider the region of the plane exterior to the $2g$ circles $\{C_j\}$ and $\{C_j'\}$; a schematic is shown in Figure 2. This region turns out to have a special significance. (It is known as the *fundamental region* associated with the Schottky group generated by the Mobius maps $\{\theta_j | j = 1, \ldots, g\}$ and their inverses—see the next section.) This is because the functional relations (15) allow the function $z(\zeta)$ to be analytically continued outside this fundamental region to any point of the plane which can be reached by a finite number of applications of the transformations $\{\theta_j\}$

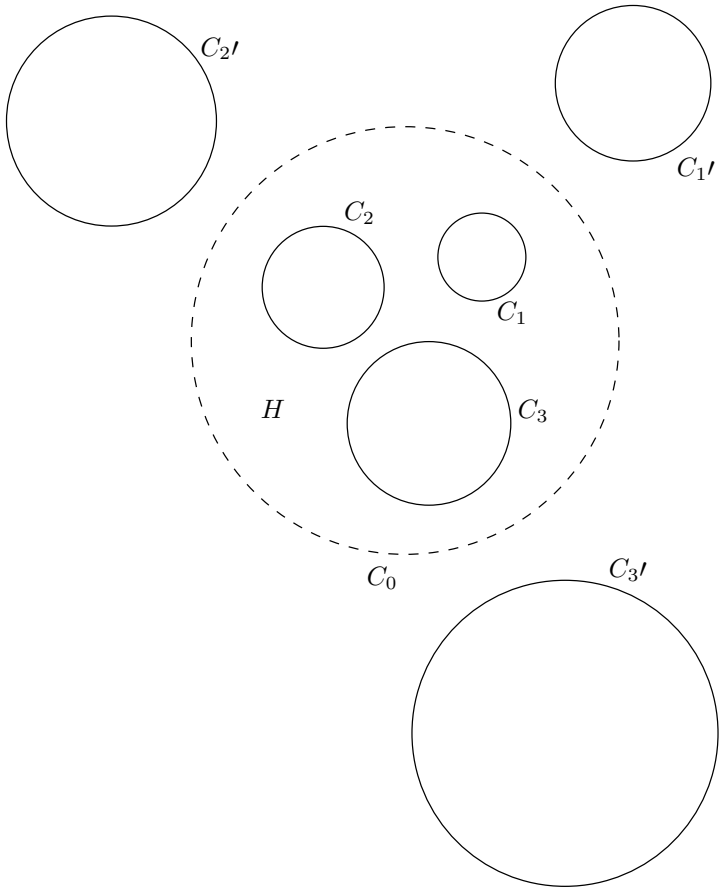


FIG. 2. *The fundamental region is the unbounded region exterior to all six Schottky circles* $C_1, C_1', C_2, C_2', C_3, C_3'$.

and their inverses to a point in this fundamental region. It is therefore enough to establish the singularity structure of $z(\zeta)$ within just this fundamental region.

Now, if the functional relations (15) hold, we have

$$\int\int_D h(z)d\bar{z}\wedge dz = \oint_{\partial H} h(z(\zeta))\bar{z}(\zeta^{-1})z_\zeta(\zeta)d\zeta, \tag{16}$$

where ∂H denotes the whole boundary of H. Now let

$$z_k = z(\bar{\alpha}_k^{-1}), \qquad k = 1,\ldots,N, \tag{17}$$

for some points $\{\bar{\alpha}_k^{-1} | k = 1,\ldots,N\}$ contained in H. Now if $\bar{z}(\zeta^{-1})$ has poles in H only at the points $\{\bar{\alpha}_k^{-1}\}$, then the integral on the right-hand side of (16) produces the pure sum of residues (6). This means that $z(\zeta)$ will have poles in the fundamental region only at the points $\{\alpha_k\}$. That is, $z(\zeta)$ is meromorphic in the fundamental region.

3. Schottky groups. Consider the $2g$ Mobius maps given by

$$\theta_1, \theta_1^{-1}, \theta_2, \theta_2^{-1}, \ldots, \theta_g, \theta_g^{-1}. \tag{18}$$

The Schottky group of transformations will be denoted Θ and is the infinite free group formed by all possible compositions of the Mobius maps (18). These maps are also referred to as the primary substitutions. Associated with this group is a fundamental region mentioned in the previous section. Sometimes we shall refer to the *ordinary* and *singular* points of the group [22], [24]. If a point in the plane can be reached by a finite number of applications of any of the $2g$ primary substitutions to a point in the fundamental region, then it is called an ordinary point of the group. If it can be reached only by an infinite number of applications, then it is called a singular point. A very accessible discussion of Schottky groups and their applications can be found in a recent monograph by Mumford, Series, and Wright [24].

Some special infinite subsets of transformations in a given Schottky group will be needed in the construction of the conformal mapping functions taking the circular regions in the ζ-plane to the multiply connected quadrature domains. A special notation is now introduced. This notation is not standard but is introduced here to clarify the presentation.

Notation. The full Schottky group is denoted Θ. The notation ${}_i\Theta_j$ is used to denote all transformations of the full group which do not have a power of θ_i or θ_i^{-1} on the left-hand end or a power of θ_j or θ_j^{-1} on the right-hand end. As a special case of this, the notation Θ_j simply means all substitutions of the group which do not have any positive or negative power of θ_j at the right-hand end (but with no stipulation about what appears on the left-hand end). Similarly, ${}_j\Theta$ means all substitutions which do not have any positive or negative power of θ_j at the left-hand end (but with no stipulation about what appears on the right-hand end). In addition, the single prime notation will be used to denote a subset where the identity is excluded from the set; thus Θ_1' denotes all substitutions, excluding the identity, and all transformations with a positive or negative power of θ_1 at the right-hand end. The double prime notation will be used to denote a subset where the identity *and* all inverse substitutions are excluded from the set. This means, for example, that if $\theta_1\theta_2$ is included in the set, the transformation $\theta_2^{-1}\theta_1^{-1}$ must be excluded. Thus, Θ'' means all substitutions of the group excluding the identity and all inverses. Similarly, the notation ${}_1\Theta_2''$ denotes all substitutions of the group, excluding inverses and the identity, which do not have any power of θ_1 or θ_1^{-1} on the left-hand end or any power of θ_2 or θ_2^{-1} on the right-hand end. In the same way, Θ_j'' denotes all substitutions of the group, excluding the identity and all inverses, which do not have any positive or negative power of θ_j at the right-hand end.

3.1. The loxodromic group. Consider a mapping to a doubly connected quadrature domain from an annular region $\rho_1 < |\zeta| < 1$ in the ζ-plane. In this case, the mapping must satisfy

$$z(\zeta) = z(\rho_1^2\zeta). \tag{19}$$

Meromorphic functions satisfying (19) are known as loxodromic functions [15]. They are automorphic with respect to the transformations of the *loxodromic group* generated by a single map of the form $\theta_1(\zeta) = \rho_1^2\zeta$. It should be noted that the fundamental region in this case can be taken to be the annulus $\rho_1 < |\zeta| < \rho_1^{-1}$, which does not include the point at infinity. The usual definition of the fundamental region associated with a classical Schottky group [22] *does* include the point at infinity. However, we adopt the convention of considering the loxodromic group to be a special case of a general Schottky group. Richardson [23] adopts the same convention.

4. The Schottky–Klein prime function. Following Baker [21], if the ith substitution of the group Θ acts on some point ζ, then the image point will be denoted ζ_i for brevity. Using this notation, the Schottky–Klein prime function is defined as

$$\omega(\zeta,\gamma) = (\zeta-\gamma)\prod_{i\in\Theta''}\left\{\zeta,\frac{\gamma}{\gamma_i},\zeta_i\right\}, \tag{20}$$

where the product is over all substitutions in the set Θ''. The curly bracket notation denotes the cross-ratio defined in the standard way as

$$\left\{\zeta,\frac{\gamma}{\gamma_i},\zeta_i\right\} \equiv \frac{(\zeta_i-\gamma)(\gamma_i-\zeta)}{(\zeta_i-\zeta)(\gamma_i-\gamma)}. \tag{21}$$

The function $\omega(\zeta,\gamma)$ is single-valued on the whole ζ-plane, has a zero at γ and all points equivalent to γ under the substitutions of the group Θ, and, excepting the singular points of the group, is infinite only at $\zeta=\infty$.

The Schottky–Klein prime function can be regarded as fundamental and is the generalization to Riemann surfaces of genus g of the irreducible factor $(\zeta-\gamma)$ used in the construction of meromorphic functions on a genus-0 Riemann surface (i.e., the rational functions) and the function $P(\zeta/\gamma,\rho)$ (see (4)) used in the construction of meromorphic functions on a genus-1 Riemann surface (i.e., the loxodromic functions).

4.1. Trivial group. When the Schottky group is just the trivial group, the definition (20) reduces to $\omega(\zeta,\gamma)=(\zeta-\gamma)$. It is well known that any rational function with poles at the N points $\{\alpha_k|k=1,\dots,N\}$ and zeros at $\{\beta_k|k=1,\dots,N\}$ admits the representation

$$R\frac{(\zeta-\beta_1)(\zeta-\beta_2)\cdots(\zeta-\beta_N)}{(\zeta-\alpha_1)(\zeta-\alpha_2)\cdots(\zeta-\alpha_N)}, \tag{22}$$

where R is a multiplicative constant. Note that in this case there is no restriction on the locations of the poles and zeros of the function.

4.2. Loxodromic group. When the relevant Schottky group is the loxodromic group generated by the single substitution $\theta_1(\zeta)=\rho_1^2\zeta$, the definition (20) reduces to

$$\begin{aligned}\omega(\zeta,\gamma) &= (\zeta-\gamma)\prod_{k=1}^{\infty}\frac{(\rho_1^{2k}\zeta-\gamma)(\rho_1^{2k}\gamma-\zeta)}{(\rho_1^{2k}\zeta-\zeta)(\rho_1^{2k}\gamma-\gamma)}\\ &= (\zeta-\gamma)\prod_{k=1}^{\infty}\frac{(\rho_1^{2k}\zeta/\gamma-1)(\rho_1^{2k}\gamma/\zeta-1)}{(\rho_1^{2k}-1)(\rho_1^{2k}-1)}\\ &= \left(\frac{-\gamma}{\prod_{k=1}^{\infty}(\rho_1^{2k}-1)^2}\right)P(\zeta/\gamma,\rho_1),\end{aligned}$$

so that the relevant Schottky–Klein prime function $\omega(\zeta,\gamma)$ in this case is simply proportional to the function $P(\zeta/\gamma,\rho_1)$ given in the introduction. It is well known [15] that one representation for a loxodromic function with poles at $\{\alpha_k|k=1,\dots,N\}$ and zeros at $\{\beta_k|k=1,\dots,N\}$ is

$$R\frac{P(\zeta/\beta_1,\rho_1)P(\zeta/\beta_2,\rho_1)\cdots P(\zeta/\beta_N,\rho_1)}{P(\zeta/\alpha_1,\rho_1)P(\zeta/\alpha_2,\rho_1)\cdots P(\zeta/\alpha_N,\rho_1)}, \tag{23}$$

provided the poles and zeros satisfy the condition

$$\prod_{k=1}^{N} \alpha_k = \prod_{k=1}^{N} \beta_k, \tag{24}$$

i.e., there is a single condition on the poles and zeros of the function. It is important to point out that another representation of a loxodromic function with the same poles and zeros is given by

$$R\zeta \frac{P(\zeta/\beta_1,\rho_1)P(\zeta/\beta_2,\rho_1)\cdots P(\zeta/\beta_N,\rho_1)}{P(\zeta/\alpha_1,\rho_1)P(\zeta/\alpha_2,\rho_1)\cdots P(\zeta/\alpha_N,\rho_1)}, \tag{25}$$

where we emphasize the appearance of an additional prefactor of ζ in front of the ratio of products of the $P(\zeta,\rho)$-functions. In this case, the poles and zeros must satisfy the modified condition

$$\prod_{k=1}^{N} \alpha_k = \rho_1^2 \prod_{k=1}^{N} \beta_k. \tag{26}$$

Crowdy [20] has explicitly constructed quadrature domains corresponding to annular arrays of near-touching cylindrical particles using the second representation (25).

4.3. More general Schottky groups. By a natural extension of the familiar special cases of sections 4.1 and 4.2, it can be shown [21] that one representation of a meromorphic function on a Riemann surface of genus g with the poles $\{\alpha_k | k = 1,\ldots,N\}$ and zeros $\{\beta_k | k = 1,\ldots,N\}$ is

$$R\frac{\omega(\zeta,\beta_1)\omega(\zeta,\beta_2)\cdots\omega(\zeta,\beta_N)}{\omega(\zeta,\alpha_1)\omega(\zeta,\alpha_2)\cdots\omega(\zeta,\alpha_N)}. \tag{27}$$

It is natural that in the genus-g case there exist g conditions on the poles and zeros. These are the generalizations of the single condition (24) or (26) in the $g = 1$ case. To ascertain these conditions, introduce A_k and B_k as the two fixed points of the generating substitution θ_k defined as

$$A_k = \theta_k^{-\infty}\zeta, \quad B_k = \theta_k^{\infty}\zeta, \tag{28}$$

where ζ is any given point. Note that A_k and B_k are simply the roots of $\zeta = \theta_k(\zeta)$, which is just a quadratic because $\theta_k(\zeta)$ is a Mobius transformation. Letting $\theta_k = \zeta'$, it is possible to write

$$\frac{\zeta' - B_k}{\zeta' - A_k} = \mu_k e^{i\kappa_k}\frac{\zeta - B_k}{\zeta - A_k}, \tag{29}$$

where $\mu_k, \kappa_k \in \mathbb{R}$. The two roots A_k and B_k are then distinguished by the fact that $|\mu_k| < 1$ in (29). Now the function (27) is the required meromorphic function, provided the following g conditions hold:

$$\prod_{j=1}^{N}\prod_{\theta_i\in\Theta_k}\frac{(\alpha_j - \theta_i(B_k))}{(\alpha_j - \theta_i(A_k))}\Big/\frac{(\beta_j - \theta_i(B_k))}{(\beta_j - \theta_i(A_k))} = 1, \quad k = 1,\ldots,g. \tag{30}$$

Note that the substitutions in the second product are taken from the subset Θ_k. The g conditions (30) will be referred to henceforth as the *automorphicity conditions*.

In the same way that both (23) and (25) are two different representations of a loxodromic function with the same distribution of poles and zeros, there are a number of distinct representations of meromorphic functions on a Riemann surface of genus $g > 1$ as shown in Baker [21]. In constructing a particular quadrature domain, it is necessary to ascertain which representation is the appropriate one needed to construct the required conformal mapping. One alternative representation (used later in the case studies) is

$$\left(R\zeta \prod_{i\in\Theta_1'} \frac{\zeta-\theta_i(B_1)}{\zeta-\theta_i(A_1)} \right) \frac{\omega(\zeta,\beta_1)\omega(\zeta,\beta_2)\cdots\omega(\zeta,\beta_N)}{\omega(\zeta,\alpha_1)\omega(\zeta,\alpha_2)\cdots\omega(\zeta,\alpha_N)}, \tag{31}$$

where θ_1 denotes the loxodromic transformation given as

$$\theta_1(\zeta) = \rho_1^2\zeta. \tag{32}$$

The poles and zeros also satisfy g automorphicity conditions, one of which is given by

$$\prod_{i=1}^{N}\prod_{j\in\Theta_1}\left(\frac{\beta_i-\theta_j(B_1)}{\beta_i-\theta_j(A_1)}\Big/\frac{\alpha_i-\theta_j(B_1)}{\alpha_i-\theta_j(A_1)}\right) = \frac{1}{\mu_1 e^{i\kappa_1}}\prod_{s\in_1\Theta_1''}\left(\frac{B_1-\theta_s(A_1)}{A_1-\theta_s(A_1)}\Big/\frac{B_1-\theta_s(B_1)}{A_1-\theta_s(B_1)}\right)^2, \tag{33}$$

while the remaining $g-1$ conditions are given by

$$\prod_{i=1}^{N}\prod_{j\in\Theta_b}\left(\frac{\beta_i-\theta_j(B_b)}{\beta_i-\theta_j(A_b)}\Big/\frac{\alpha_i-\theta_j(B_b)}{\alpha_i-\theta_j(A_b)}\right) = \prod_{s\in_1\Theta_b}\left(\frac{\theta_s^{-1}(B_1)-A_b}{\theta_s^{-1}(A_1)-A_b}\Big/\frac{\theta_s^{-1}(B_1)-B_b}{\theta_s^{-1}(A_1)-B_b}\right) \tag{34}$$

for $b = 2,\ldots,g$, where A_b and B_b denote the fixed points of the mapping θ_b.

Finally, it is instructive to see how the general condition (30) reduces to (24) in the $g = 1$ case, where the Schottky group is the loxodromic group. In this case the group is generated by the single substitution,

$$\theta_1(\zeta) = \rho_1^2\zeta. \tag{35}$$

The subset Θ_1 then contains only the identity. It is also clear that

$$A_1 = \infty, \quad B_1 = 0. \tag{36}$$

With these identifications, it is easy to show that (30) is precisely equivalent to (24). Indeed, it is also straightforward to show that the factor

$$\zeta\prod_{i\in\Theta_1'}\frac{\zeta-\theta_i(B_1)}{\zeta-\theta_i(A_1)} \tag{37}$$

in (31) reduces simply to ζ in the case where the Schottky group is precisely the loxodromic group, so that (31) reduces to (25). At the same time, the automorphicity condition (33) reduces to (26).

5. Equations for the mapping parameters. Only quadrature domains satisfying quadrature identities of the form

$$\int\int_D h(z)d\bar{z}\wedge dz = 2i\sum_{k=1}^{N} a_k h(z_k) \tag{38}$$

will be considered here. The equations to be satisfied by the conformal mapping parameters come from the specified quadrature identity together with any assumptions made regarding the areas of the enclosed holes. Intuitively, it is useful to think of specifying the real parameter ρ_i as equivalent to specifying the area of the ith hole.

From (17) recall that we require

$$z(\bar{\alpha}_k^{-1}) = z_k, \quad k = 1, \ldots, N. \tag{39}$$

Also, recall that we require $\bar{z}(\zeta^{-1})$ to have poles in H only at the points $\bar{\alpha}_k^{-1}$. In this case, where the quadrature identity is of the form (38), these poles are simple. Thus near $\zeta = \bar{\alpha}_k^{-1}$, $\bar{z}(\zeta^{-1})$ has the form

$$\bar{z}(\zeta^{-1}) = \frac{P_k}{\zeta - \bar{\alpha}_k^{-1}} + \text{regular}, \tag{40}$$

where $P_k \in \mathbb{C}$. We therefore require that

$$a_k = \pi P_k z_\zeta(\bar{\alpha}_k^{-1}), \quad k = 1, \ldots, N. \tag{41}$$

It is useful to think of the N conditions (39) as being equations for the N poles $\{\alpha_k | k = 1, \ldots, N\}$, while (41) provides equations for the N zeros $\{\beta_k | k = 1, \ldots, N\}$.

This leaves only the set $\{\delta_k | k = 1, \ldots, g\}$ to be determined. However, equations for these can be understood as being given by the g automorphicity conditions (30). The equation count is therefore very natural, as indicated by the following schematic encapsulating the correspondence between parameters:

$$\begin{aligned} \{z_k \in \mathbb{C} | k = 1, \ldots, N\} &\to \{\alpha_k \in \mathbb{C} | k = 1, \ldots, N\}, \\ \{a_k \in \mathbb{C} | k = 1, \ldots, N\} &\to \{\beta_k \in \mathbb{C} | k = 1, \ldots, N\}, \\ \{\text{specifying the area of g holes}\} &\to \{\rho_k \in \mathbb{R} | k = 1, \ldots, g\}, \\ \{\text{g automorphicity conditions}\} &\to \{\delta_k \in \mathbb{C} | k = 1, \ldots, g\}. \end{aligned} \tag{42}$$

A minor modification of the prime function (20) is needed when the mapping is required to have a zero or pole at the point at infinity. In this case formula (20) must be replaced by

$$\omega(\zeta, \infty) = \prod_{i \in \Theta''} \frac{(\infty_i - \zeta)}{(\zeta_i - \zeta)}, \tag{43}$$

where ∞_i denotes the images of the point at infinity under the ith substitution of the set Θ''.

Many of the examples to follow possess various rotational symmetries in the distribution of the poles and zeros of the relevant conformal mapping function. It is therefore convenient to define $\omega_n(\zeta, \gamma)$ as

$$\omega_n(\zeta, \gamma) \equiv \prod_{k=0}^{n-1} \omega(\zeta, e^{2\pi i k/n}\gamma). \tag{44}$$

It should be noted that the Schottky–Klein prime function depends implicitly on the parameters $\{\delta_k, \rho_k | k = 1, \ldots, g\}$ from which the primary substitutions are constructed; however, the notation $\omega(\zeta, \gamma)$ suppresses this dependence.

6. Examples. In computing explicit cases it is necessary to truncate the number of maps used from any of the relevant infinite sets. This is done in a natural way by picking a *level* of composition of the primary substitutions up to which all composed substitutions are included. (Mumford, Series, and Wright [24] discuss various other methods of truncation.) For example, if a Schottky group has two primary substitutions θ_1 and θ_2 and all maps up to and including level 2 are used, the following maps would be included in the definition of $\omega(\zeta,\gamma)$:

$$\text{level } 1: \ \theta_1,\theta_2; \quad \text{level } 2: \ \theta_1^2,\theta_2^2,\theta_1\theta_2,\theta_2\theta_1,\theta_1^{-1}\theta_2,\theta_1\theta_2^{-1}. \tag{45}$$

Note that the identity is the only map at level 0, and this is excluded in defining the Schottky–Klein prime function.

Since the zeros and poles in the Schottky–Klein prime function representation are explicit, construction of a given quadrature domain requires only consideration of the distribution of the poles and zeros of the conformal mapping in the ζ preimage plane. Often, the quadrature identity combined with symmetry considerations can be used to deduce the positions of these poles and zeros. The functional form of the relevant conformal mapping can then be written down immediately.

6.1. A triply connected quadrature domain. Consider four circular discs of equal radius r initially less than 1, with centers at $\pm\sqrt{3}$ and $\pm i$. For $r<1$ the circular discs are disconnected. If we increase r to 1, then the discs touch. If $r\le 1$, such a configuration is a disconnected quadrature domain satisfying the quadrature identity (38) with $N=4$, $a_1=a_2=a_3=a_4=\pi r^2$, and $z_1=\sqrt{3}=-z_3$, $z_2=i=-z_4$. When r increases above 1, the domain satisfying the quadrature identity (38) with quadrature data given by $a_1=a_2=a_3=a_4=\pi r^2$ and $z_1=\sqrt{3}=-z_3, z_2=i=-z_4$ can be expected to form a triply connected quadrature domain.

We shall construct a triply connected domain which is close to the case of touching circular discs. In particular, we take $a_1=a_2=a_3=a_4=1.0010\pi$ and $z_1=1.6966, z_2=0.9969i$.

Note that the quadrature domain is symmetric with respect to reflection in both the real and imaginary axes, and its two holes have their centers on the real axis. It is natural to expect the same structure in the associated circular region H in the ζ-plane. If C_1 and C_2 are the circles mapping to the boundaries of these two holes, we expect them to have equal radii with centers at $\delta_1,\delta_2\in\mathbb{R}$, where $\delta_1=-\delta_2$. The conformal mapping will have four poles corresponding to each of the z_k for $k=1,2,3,4$. We label these α_k for $k=1,2,3,4$. It will also have four zeros, which we label β_k $(k=1,2,3,4)$. Again, it is natural to expect the distribution of the poles of the conformal map in the ζ-plane to mirror the distribution of the points z_k $(k=1,2,3,4)$ in the physical plane. We therefore expect $\alpha_1=-\alpha_3$ purely real and $\alpha_2=-\alpha_4$ purely imaginary. Thus, the combination $\omega_2(\zeta,\alpha_1)\omega_2(\zeta,\alpha_2)$ will appear in the denominator of the conformal map. Note that the compact notation $\omega_2(\zeta,\alpha_1)$ (defined in (44)) automatically captures both the pole at α_1 and that at $-\alpha_1$. As for the zeros, because we choose $\zeta=0$ to map to $z=0$, one of the zeros (say β_3) is at the origin. Thus $\omega(\zeta,0)$ appears in the numerator of the conformal map. By symmetry, one of the remaining three zeros (say β_4) must be at ∞, while the other two, β_1 and β_2, should be either purely real or purely imaginary with $\beta_1=-\beta_2$. Thus, the combination $\omega(\zeta,\infty)\omega_2(\zeta,\beta_1)$ also appears in the numerator. In fact, it is found that β_1 is purely real. Figure 3 shows a schematic illustrating the ζ preimage plane and the distribution of poles and zeros in the fundamental region.

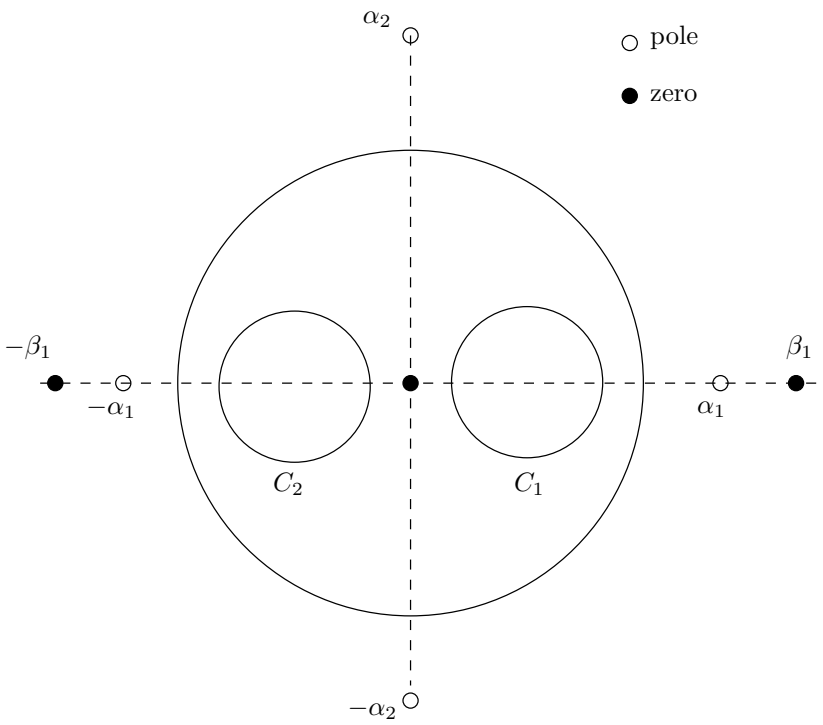


FIG. 3. *Schematic illustrating the ζ preimage plane with distribution of poles and zeros of the conformal mapping to the triply connected quadrature domain in Figure* 4.

Combining the above considerations, the form of the conformal map is deduced to be

$$z(\zeta) = R\frac{\omega(\zeta,0)\omega(\zeta,\infty)\omega_2(\zeta,\beta_1)}{\omega_2(\zeta,\alpha_1)\omega_2(\zeta,\alpha_2)}. \tag{46}$$

The map contains six parameters: $R, \beta_1, \alpha_1, \alpha_2, \delta_1, \rho_1$. We can specify ρ_1, which corresponds to fixing the area of each of the two holes. Then the equations to solve for the remaining five unknowns come from (30), (39), (41). Note that, due to symmetry, the two equations given by (30) are actually the same, and (39), (41) each give only two independent equations. Thus we have five equations for five unknowns. Explicitly, these are

$$\prod_{j=1}^{4}\prod_{\theta_i\in\Theta_1}\frac{(\alpha_j-\theta_i(B_1))}{(\alpha_j-\theta_i(A_1))}\Big/\frac{(\beta_j-\theta_i(B_1))}{(\beta_j-\theta_i(A_1))} = 1, \tag{47}$$

$$z_1 = z(\bar{\alpha}_1^{-1}), \tag{48}$$

$$z_2 = z(\bar{\alpha}_2^{-1}), \tag{49}$$

$$a_1 = \pi P_1 z_\zeta(\bar{\alpha}_1^{-1}), \tag{50}$$

$$a_2 = \pi P_2 z_\zeta(\bar{\alpha}_2^{-1}), \tag{51}$$

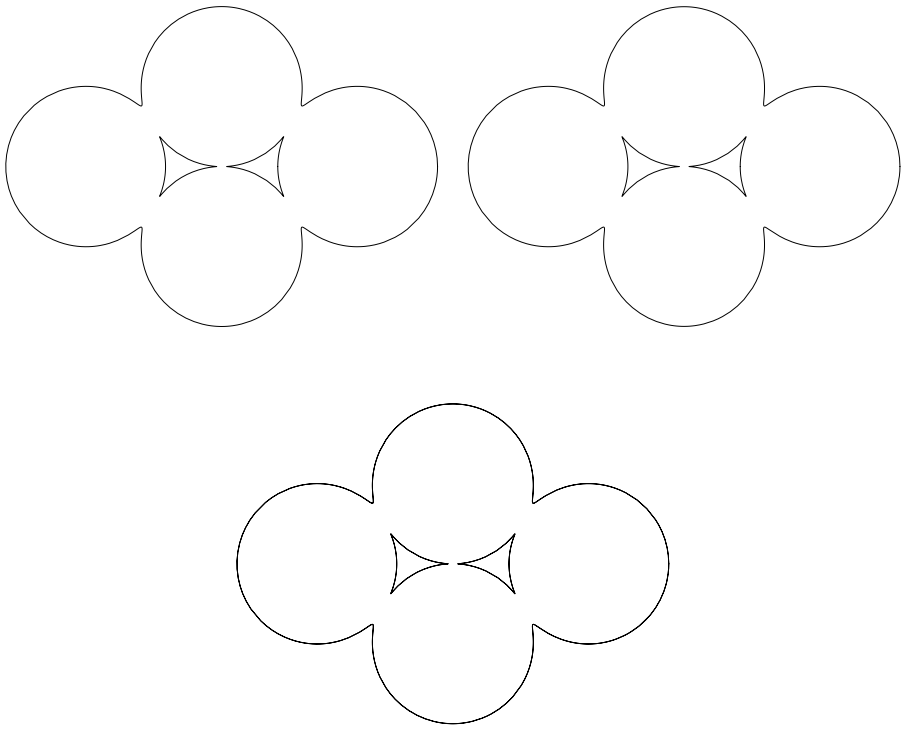


FIG. 4. *Triply connected domain constructed using the Schottky–Klein prime function (top left) and Poincaré series (top right) both at level* 3. *Here* $R = 2.5215,\ \beta_1 = 1.4776,\ \alpha_1 = 1.1520,\ \alpha_2 = 1.5969i,\ \delta_1 = 0.3160,\ \rho_1 = 0.2000$. *For comparison, the lower diagram shows a superposition of the upper two diagrams.*

where analytical formulae for P_1 and P_2 can easily be deduced. For example,

$$P_1 = -\frac{R}{\alpha_1^2}\left(\frac{\omega(\alpha_1,0)\omega(\alpha_1,\infty)\omega_2(\alpha_1,\beta_1)}{\hat{\omega}(\alpha_1,\alpha_1)\omega(\alpha_1,-\alpha_1)\omega_2(\alpha_1,\alpha_2)}\right), \tag{52}$$

where $\hat{\omega}(\zeta,\gamma)$ is defined as

$$\hat{\omega}(\zeta,\gamma) \equiv \prod_{i\in\Theta''}\left\{\zeta, \frac{\gamma}{\gamma_i}, \zeta_i\right\}. \tag{53}$$

These five equations are solved for the unknown parameters using Newton's method. The image of the conformal map is shown in the left-most diagram in Figure 4.

For purposes of comparison with the approach to constructing quadrature domains expounded recently by Richardson [23] we constructed the *same* quadrature domain using a conformal map based on the use of Poincaré series as opposed to the Schottky–Klein prime function. The image of this map is shown in the right-most diagram in Figure 4. The images of the respective conformal maps are indistinguishable, as can be seen from their superposition in the lower diagram in Figure 4. A brief overview of Richardson's general method, and details of how it was used to construct the above triply connected domain, are given in the appendix.

6.2. A quadruply connected quadrature domain. A second example is to consider three circular discs in an annular array surrounding a smaller circular disc.

The quadrature identity associated with such a domain is of the form (38) with $N = 4$ and z_1 purely real, $z_2 = z_1 e^{2\pi i/3}, z_3 = z_1 e^{4\pi i/3}$, and $z_4 = 0$. In the case where the circular discs are touching, we have $a_1 = a_2 = a_3 = \pi$, $a_4 = \pi(2/\sqrt{3} - 1)^2$ and $z_1 = 2/\sqrt{3}, z_4 = 0$. We shall construct a quadruply connected domain which is close to the case of touching discs. In particular, we have taken the quadrature data to be $a_1 = a_2 = a_3 = 1.0010\pi, a_4 = 0.0250\pi$ and $z_1 = 1.1488, z_4 = 0$.

The domain has three holes. Since the holes in the physical plane are rotations of each other through $\frac{2\pi}{3}$, we expect the circles in the preimage plane to share these symmetries. Let C_1, C_2, and C_3 be the circles inside the unit ζ-circle mapping to the boundaries of the holes. Then C_1 will be centered on the ray $\arg[\zeta] = \frac{\pi}{3}$, and C_2 and C_3 will be rotations of this circle through $\frac{2\pi}{3}$.

If we fix the physical origin to be the image of $\zeta = 0$, then we require the factor $\omega(\zeta, 0)$ to appear in the numerator of the conformal map. Note that one of the z_k is zero, namely z_4. Thus from (17), we see that we require the pole α_4 (corresponding to the point z_4) to be at ∞. Thus we must include the factor $\omega(\zeta, \infty)$ in the denominator of the map. There will also be three other poles, symmetrically disposed about $\zeta = 0$, corresponding to the symmetrically disposed points z_1, z_2, and z_3. One of these, denoted α_1, is on the real ζ-axis. There will also be three additional zeros of the conformal map in the fundamental region, which are also expected to be symmetrically disposed about $\zeta = 0$. One of these, β_1 say, is found to be real.

Using these considerations, the conformal map is deduced to have the form

$$z(\zeta) = R\frac{\omega(\zeta, 0)\omega_3(\zeta, \beta_1)}{\omega(\zeta, \infty)\omega_3(\zeta, \alpha_1)}. \tag{54}$$

The image of this map, constructed to level-3 accuracy, is shown in the left-most diagram in Figure 5 along with the image of the conformal map constructed using the Poincaré series method of Richardson [23] to its right (again, to level-3 accuracy). Their superposition is also shown in Figure 5. The boundaries are indistinguishable.

6.3. A quintuply connected quadrature domain. It is straightforward to generalize the previous example to a quadrature domain which is close to the case of four circular discs in an annular array surrounding a smaller circular disc. The quadrature identity associated with such a domain is of the form (38) with $N = 5$ and z_1 purely real, $z_k = z_1 e^{(k-1)\pi i/2}, k = 2, 3, 4$, and $z_5 = 0$. In the case where the discs are touching, we have $a_1 = a_2 = a_3 = a_4 = \pi$, $a_5 = \pi(\sqrt{2} - 1)^2$ and $z_1 = \sqrt{2}, z_5 = 0$. We shall construct a quintuply connected domain which is close to the case of touching circular discs. In particular, we choose quadrature data given by $a_1 = a_2 = a_3 = a_4 = 0.9980\pi, a_5 = 0.1716\pi$ and $z_1 = 1.4029, z_5 = 0$.

Considerations similar to the previous example can be used to deduce that the associated map has the form

$$z(\zeta) = R\frac{\omega(\zeta, 0)\omega_4(\zeta, \beta_1)}{\omega(\zeta, \infty)\omega_4(\zeta, \alpha_1)}. \tag{55}$$

The image of the conformal map is shown in the left-most diagram in Figure 6 along with the image of the conformal map constructed using Poincaré series. Both are constructed to level-3 accuracy. Their superposition is also shown in Figure 6 and, again, the quadrature domain boundaries are indistinguishable. Crowdy [13] has considered this class of domains from the point of view of algebraic curves in the context of constructing multipolar equilibria of the Euler equations, and this will be considered again later in section 8.

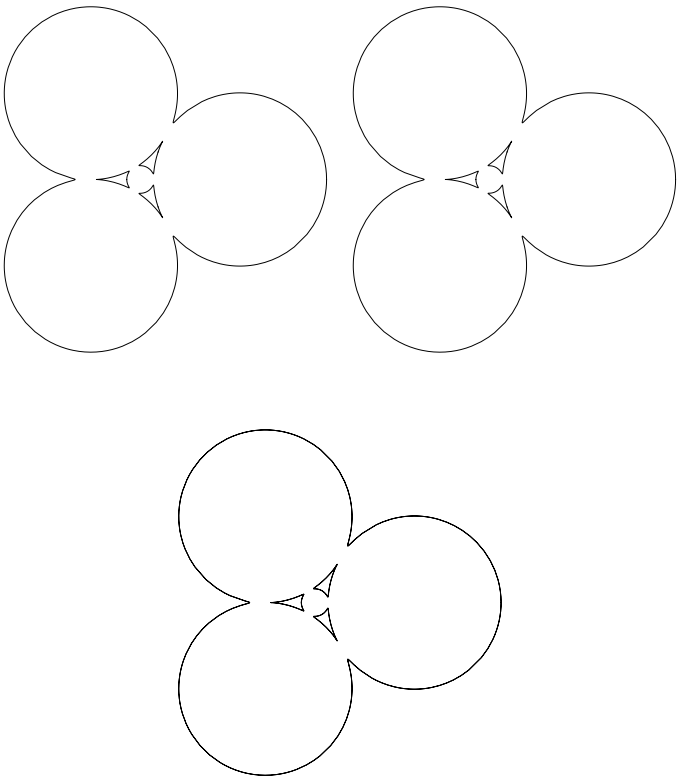


FIG. 5. *Quadruply connected domain constructed using the Schottky–Klein prime function (top left) and Poincaré series (top right) at level* 3*, with superposition (lower). Here* $R = 0.0536,\ \beta_1 = 3.4038,\ \alpha_1 = 1.2500,\ \delta_1 = 0.2608e^{i\pi/3},\ \rho_1 = 0.1275.$

6.4. A septuply connected quadrature domain. Richardson [23] has considered the case of *six* circular discs in an annular array containing a disc of equal radius in the center. Such a case is a trivial extension of the examples in sections 6.2 and 6.3. The preceding two examples have conformal mappings of the general functional form

$$z(\zeta) = R\frac{\omega(\zeta,0)\omega_n(\zeta,\beta_1)}{\omega(\zeta,\infty)\omega_n(\zeta,\alpha_1)}, \tag{56}$$

where section 6.2 deals with $n = 3$ while section 6.3 treats the case $n = 4$. The case of six circular discs surrounding a central one will have a conformal map of the form

$$z(\zeta) = R\frac{\omega(\zeta,0)\omega_6(\zeta,\beta_1)}{\omega(\zeta,\infty)\omega_6(\zeta,\alpha_1)}, \tag{57}$$

i.e., it is given by a mapping of the form (56) with $n = 6$. The associated quadrature identity is of the form (38) with $N = 7$ and z_1 purely real, $z_k = z_1 e^{(k-1)\pi i/3}, k = 2,\ldots,6$, and $z_7 = 0$. In the case of touching circular discs, we have $a_1 = \cdots = a_7 = \pi$ and $z_1 = 2, z_7 = 0$. For illustration, we construct a septuply connected domain which is close to the case of touching circular discs with $a_1 = \cdots = a_6 = 1.0266\pi, a_7 = 1.0010\pi$ and $z_1 = 2.0002, z_7 = 0$. Figure 7 shows the results constructed using both methods to level-2 accuracy.

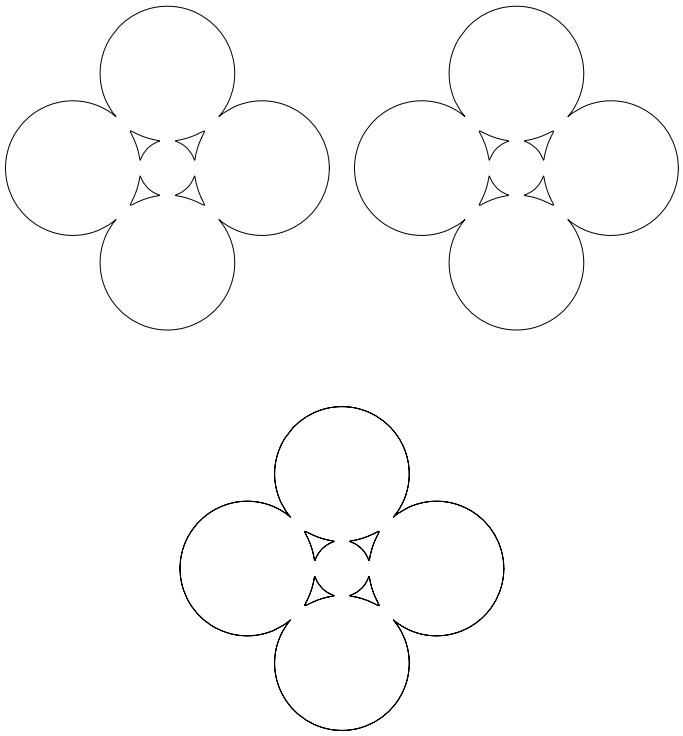


FIG. 6. *Quintuply connected domain constructed using the Schottky–Klein prime function (top left) and Poincaré series (top right) at level* 3, *with superposition (lower). Here* $R = 0.1849,\ \beta_1 = 2.0789,\ \alpha_1 = 1.2329,\ \delta_1 = 0.3591e^{i\pi/4},\ \rho_1 = 0.1290.$

6.5. 3-by-3 square array. More complicated domains are also easy to construct using the Schottky–Klein prime function representation. Consider, for example, nine circles in a 3-by-3 square array. The associated quadrature domain is of the form (38) with $N = 9$ and z_1 real, $z_2 = z_1 i, z_3 = -z_1, z_4 = -z_1 i$, and z_5 on the $\pi/4$ ray, with $z_6 = z_5 i, z_7 = -z_5, z_8 = -z_5 i$, and $z_9 = 0$. In the case where the circular discs are touching, we have $z_1 = 2, z_5 = 2\sqrt{2}e^{\pi i/4}$, and $a_1 = \cdots = a_9 = \pi$. We shall construct a quintuply connected domain which is close to the case of touching circular discs with $a_1 = \cdots = a_9 = 1.0010\pi$ and $z_1 = 1.9533, z_5 = 2.7696e^{\pi i/4}, z_9 = 0$.

There will be four circles $C_1, \ldots, C_4$ inside the unit circle in the preimage ζ-plane. Since the holes in the physical plane are centered on the rays $\arg[z] = \frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4}$, we also expect the centers of the circles $C_1, \ldots, C_4$ to be on these rays in the ζ-plane. Let $\zeta = 0$ map to $z = 0$. This means that $\omega(\zeta, 0)$ must appear in the numerator of the conformal map. Furthermore, because $z_9 = 0$, there must be a corresponding pole of the conformal map at infinity in the ζ-plane. Therefore, $\omega(\zeta, \infty)$ must appear in the denominator. We expect four symmetrically disposed poles in the ζ-plane corresponding to $z_1, \ldots, z_4$. Let one of these be α_1 on the real axis. Similarly, let α_2 (taken on the ray $\arg[\zeta] = \frac{\pi}{4}$) and its rotations through $\frac{\pi}{2}$ correspond to $z_5, \ldots, z_8$. Thus, the combination $\omega_4(\zeta, \alpha_1)\omega_4(\zeta, \alpha_2)$ will also appear in the denominator. The zeros are expected to be similarly distributed in the ζ-plane. Therefore we include the combination $\omega_4(\zeta, \beta_1)\omega_4(\zeta, \beta_2)$ in the numerator so that the zeros of the map are β_1 (and its three rotations through $\frac{\pi}{2}$) and β_2 (along with its three rotations through $\frac{\pi}{2}$). It is found that β_1 is real while β_2 is on the ray $\arg[\zeta] = \frac{\pi}{4}$.

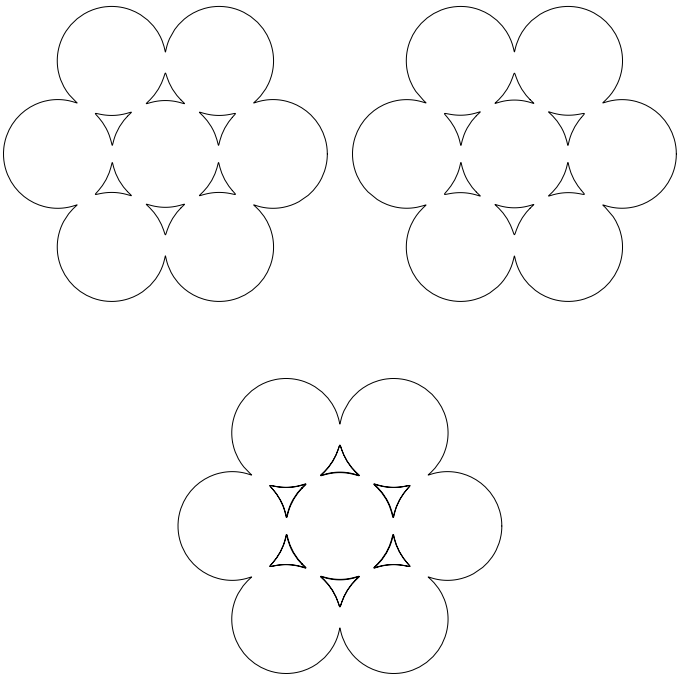


FIG. 7. *Septuply connected domain constructed using the Schottky–Klein prime function (top left) and Poincaré series (top right) at level* 2, *with superposition (lower). Here* $R = 0.6089,\ \beta_1 = 1.5358,\ \alpha_1 = 1.2195,\ \delta_1 = 0.4900e^{i\pi/6},\ \rho_1 = 0.1260.$

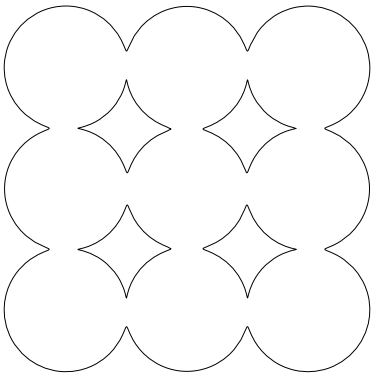


FIG. 8. *3-by-3 square array constructed using the Schottky–Klein prime function at level* 2. *Here* $R = 0.7887,\ \beta_1 = 1.8298,\ \beta_2 = 1.1828e^{i\pi/4},\ \alpha_1 = 1.3899,\ \alpha_2 = 1.0989e^{i\pi/4},\ \delta_1 = 0.5093e^{i\pi/4},\ \rho_1 = 0.2100.$

The conformal map therefore has the form

$$z(\zeta) = R\frac{\omega(\zeta,0)\omega_4(\zeta,\beta_1)\omega_4(\zeta,\beta_2)}{\omega(\zeta,\infty)\omega_4(\zeta,\alpha_1)\omega_4(\zeta,\alpha_2)}. \tag{58}$$

The image of the conformal map is shown in Figure 8.

6.6. Examples with a loxodromic subgroup. If we map H onto a multiply connected quadrature domain which is rotationally symmetric about the origin and which has the origin inside one of its holes, then the boundary ∂D_1 of this hole will be centered at the origin, and thus the circle C_1 in the associated circular region in the ζ-plane will be centered at $\zeta = 0$. So, referring to (14), we see that the associated Schottky group will contain a loxodromic subgroup.

If the quadrature domain is in fact just doubly connected and ∂D_1 is its only inner boundary, then the associated Schottky group will be precisely the loxodromic group. In this case, the form (23) does not map the circular region to the required image. However, Crowdy [20] has shown that the appropriate loxodromic function is given by (25) where the poles and zeros satisfy (26). It is similarly found that if a more general Schottky group contains the loxodromic group as a subgroup, it is necessary to use a suitably generalized representation of the required conformal mapping.

We now present an example where the Schottky group has a loxodromic subgroup. The example chosen is one suggested by Richardson [23]. Consider six circular discs arranged in a triangular array. The quadrature identity associated with such a domain is of the form (38) with $N = 6$ and z_1 purely real, $z_2 = z_1 e^{2\pi i/3}$, $z_3 = z_1 e^{4\pi i/3}$, and z_4 on the $\pi/3$ ray, $z_5 = z_4 e^{2\pi i/3}$, $z_6 = z_4 e^{4\pi i/3}$. In the case where the circular discs are touching we have $a_1 = \cdots = a_6 = \pi$ and $z_1 = \frac{2}{\sqrt{3}}$, $z_4 = \frac{4}{\sqrt{3}} e^{\pi i/3}$. We shall construct a quintuply connected domain which is close to the case of touching circular discs with $a_1 = \cdots = a_6 = 1.0500\pi$ and $z_1 = 1.1737$, $z_4 = 2.3536 e^{\pi i/3}$.

In this case, there will be a total of four enclosed holes: one centered at the origin and three others at symmetrically disposed positions about the origin. Let C_1 be the circle in the ζ-plane mapping to the central hole, and let C_2, C_3, C_4 map to the other three holes. C_2 is a circle centered at some point δ_2 on the ray $\arg[\zeta] = \frac{\pi}{3}$, while C_3 and C_4 are the rotations of this circle through $\frac{2\pi}{3}$ and $\frac{4\pi}{3}$, respectively. Corresponding to z_1, z_2, and z_3 we expect three symmetrically disposed poles in the ζ-plane. Let one of these be α_1 on the real axis. Similarly, let α_2 (on the ray $\arg[\zeta] = \frac{\pi}{3}$) and its two rotations through $\frac{2\pi}{3}$ correspond to z_4, z_5, and z_6. The combination $\omega_3(\zeta, \alpha_1)\omega_3(\zeta, \alpha_2)$ will therefore appear in the denominator of the conformal map. Again, the distribution of zeros is expected to be similar. Thus, we put $\omega_3(\zeta, \beta_1)\omega_3(\zeta, \beta_2)$ in the numerator so that β_1 and β_2 (along with their respective rotations through $\frac{2\pi}{3}$) will be the zeros of the conformal map in the fundamental region. It is found that β_1 is real while β_2 is on the ray $\arg[\zeta] = \frac{\pi}{3}$.

A natural choice to make for the mapping is therefore

$$z(\zeta) = R \frac{\omega_3(\zeta, \beta_1)\omega_3(\zeta, \beta_2)}{\omega_3(\zeta, \alpha_1)\omega_3(\zeta, \alpha_2)}. \tag{59}$$

However, no univalent conformal maps to a quadrature domain with the given quadrature data could be found for a map of this form. Therefore, a modified representation of a meromorphic function on the same Riemann surface (and with the same poles and zeros) is required. Such a representation is given by (31). Thus, it is natural to propose that the conformal mapping has the generalized form

$$z(\zeta) = \left(R\zeta \prod_{i \in \Theta_1'} \frac{\zeta - \theta_i(B_1)}{\zeta - \theta_i(A_1)} \right) \frac{\omega_3(\zeta, \beta_1)\omega_3(\zeta, \beta_2)}{\omega_3(\zeta, \alpha_1)\omega_3(\zeta, \alpha_2)}, \tag{60}$$

where θ_1 denotes the loxodromic transformation

$$\theta_1(\zeta) = \rho_1^2 \zeta \tag{61}$$

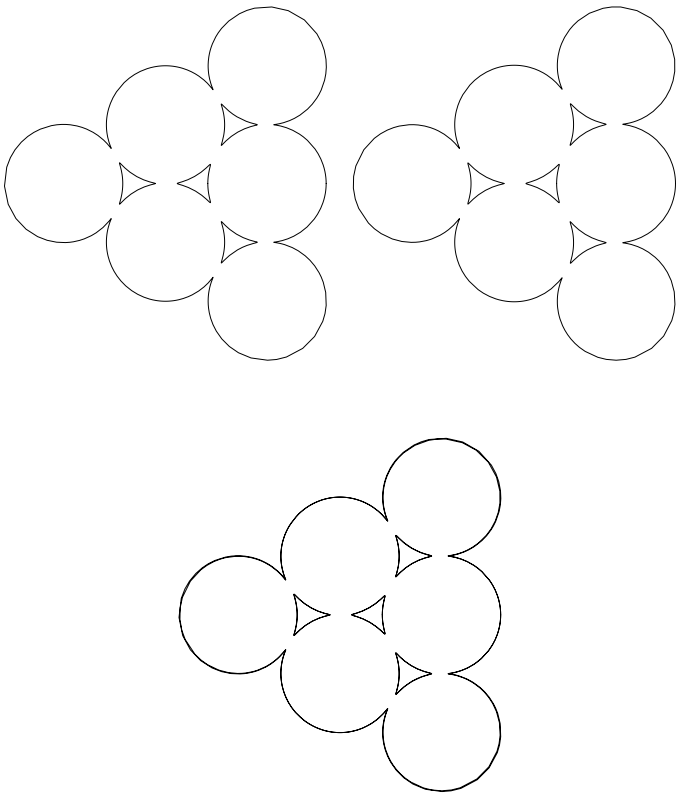


FIG. 9. *Quintuply connected domain constructed using the Schottky–Klein prime function (top left) and Poincaré series (top right) at level* 2, *with superposition (lower). Here* $R = 0.0302$, $\beta_1 = 5.3203$, $\beta_2 = 1.1228e^{i\pi/3}$, $\alpha_1 = 1.6393$, $\alpha_2 = 1.0526e^{i\pi/3}$, $\delta_1 = 0$, $\delta_2 = 0.6054e^{i\pi/3}$, $\rho_1 = 0.1450$, $\rho_2 = 0.1450$.

associated with the circle $C_1 = \{|\zeta| = \rho_1\}$. Note that, accordingly, the poles and zeros must now satisfy g modified automorphicity conditions given (in the general case) by (33) and (34). The map (60) is indeed found to provide the required univalent map to a quadrature domain satisfying the given quadrature identity. It is emphasized that the additional prefactor in (60) relative to (59) is precisely the generalization of the additional ζ-prefactor in (25) relative to (23).

The image of the conformal map constructed using both conformal mapping methods is shown in Figure 9 along with their superposition. Although this complicated domain is only constructed to level-2 accuracy (in both methods), the plots are again virtually indistinguishable.

7. Nonsymmetric domains. All the examples considered so far have certain degrees of spatial symmetry. However, the general method also applies to domains devoid of any such symmetry. Figure 10 shows two typical quadrature domains, plotted using conformal maps based on Schottky–Klein prime functions, possessing less symmetry than those in Figure 4. The constructive method is essentially the same, with only minor differences. For example, with no symmetry, there are now *two* independent automorphicity conditions, whereas in the symmetric case there was just one.

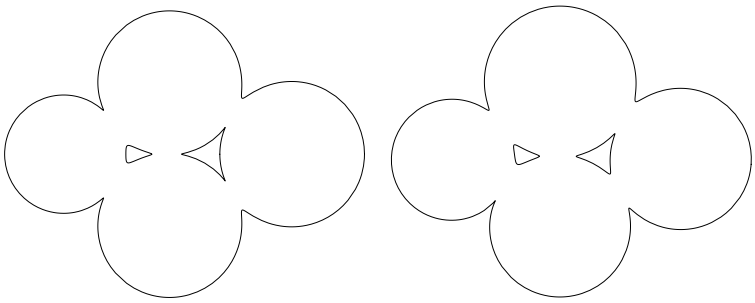


FIG. 10. *More general triply connected quadrature domains.*

8. Algebraic curves and uniformization. In the context of steady vortical equilibria of the Euler equation, Crowdy [13] has recently presented an alternative construction of multiply connected quadrature domains from their quadrature data. The method makes use of the result that the boundaries of quadrature domains are algebraic curves [2]. For completeness, and purposes of comparison, we now use conformal maps to reconstruct one of the domains of [13].

The quintuply connected quadrature domains constructed in [13] satisfy the identity

$$\int\int_D h(z)dxdy = \pi r^2 h(z_1) + \pi r^2 h(z_2) + \pi r^2 h(z_3) + \pi r^2 h(z_4) + \pi p^2 h(0). \tag{62}$$

To within a finite set of *special points* [6] (which turn out to be useful in the construction; see [13]), the boundaries of the domains corresponding to (62) are given by the algebraic curve

$$\mathcal{P}(z, \bar{z}) = 0, \tag{63}$$

where

$$\mathcal{P}(z, w) = \sum_{k,j=0}^{5} a_{kj} z^k w^j. \tag{64}$$

The set of coefficients $\{a_{kj}\}$ form a Hermitian matrix $\mathbf{A}$, where $\mathbf{A}_{kj} = a_{kj}$ and

$$\mathbf{A} = \begin{pmatrix} k & 0 & 0 & 0 & 4p^2 & 0 \\ 0 & g & 0 & 0 & 0 & -4 \\ 0 & 0 & f & 0 & 0 & 0 \\ 0 & 0 & 0 & e & 0 & 0 \\ 4p^2 & 0 & 0 & 0 & -(4r^2+p^2) & 0 \\ 0 & -4 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{65}$$

The top-left diagram of Figure 11 features a reproduction of the quadrature domain in Figure 8 of Crowdy [13] (this reference contains all the information required to derive the matrix $\mathbf{A}$).

The relevant conformal map will have the form (55). In addition to the quadrature data, to determine this map we also need to specify the area of the holes. This can be computed using the algebraic curve, but we employ an alternative (equivalent)

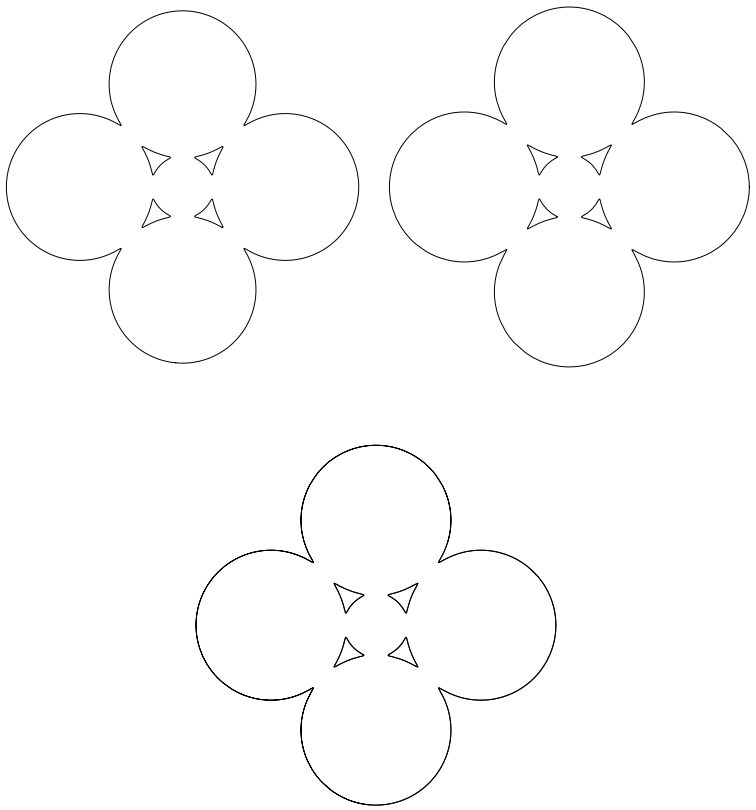


FIG. 11. *A quintuply connected quadrature domain from Crowdy (see Figure* 8 *of* [13]*) constructed using algebraic curves (top left); the same domain constructed using conformal maps based on the Schottky–Klein prime function (top right). The superposition is shown in the lower diagram.*

method. Following Crowdy [13], it is known that there exists a so-called special point at some point $z_s = se^{i\pi/4}$. At such a point it is known [13] that

$$\bar{z}_s = S(z_s), \tag{66}$$

where $S(z)$ is known as the Schwarz function of the quadrature domain boundary [13]. It is related to the conformal mapping function by the relation

$$S(z(\zeta)) = \bar{z}(\zeta^{-1}). \tag{67}$$

Crowdy [13] gives the explicit value $s = 1.008$ for the domain shown in Figure 11 (on the left). In terms of the conformal map, this point must correspond to the image of some point $\hat{\delta}e^{i\pi/4}$ in the ζ preimage plane, i.e.,

$$z_s = se^{i\pi/4} = z(\hat{\delta}e^{i\pi/4}). \tag{68}$$

At the same time, by the property (66), we must have

$$\bar{z}_s = se^{-i\pi/4} = \bar{z}(\hat{\delta}^{-1}e^{-i\pi/4}). \tag{69}$$

Therefore, instead of considering the area of the holes, (68) and (69) provide two equations relating the conformal mapping parameters, one determining the newly

introduced $\hat{\delta}$ and the other effectively specifying the area of the hole. The top-right diagram in Figure 11 shows the domain constructed using the conformal map. The lower figure shows a superposition with the domain constructed using algebraic curves. Again, the boundaries are indistinguishable.

From a theoretical viewpoint, the conformal map just constructed essentially provides the *uniformization* of the algebraic curve (63). That is, given the matrix $\mathbf{A}$, the conformal map is such that

$$\mathcal{P}(z(\zeta), \bar{z}(\zeta^{-1})) = 0. \tag{70}$$

This relation holds everywhere on the boundary of the quadrature domain, but it also holds globally by analytic continuation. In this sense, the conformal map has uniformized the algebraic curve.

9. Discussion. There are a variety of ways in which multiply connected quadrature domains can be constructed from their quadrature data (and information regarding the area of any holes). The algebraic curve method of Crowdy [13] has many conceptual advantages and requires the least analytical overhead. Using this method, an implicit description of the boundary is obtained. The idea is to iterate on the algebraic curve coefficients until equations deriving from the quadrature identity are satisfied. When the domains have symmetry, the consideration of the special points of the domain can greatly facilitate the construction by providing explicit sets of equations to be satisfied by the coefficients of the curve. The special points can also have physical significance; in Crowdy [13] they corresponded to stagnation points of the vortical flow.

In this paper, a conceptually different method has been used based on conformal mapping from a canonical region in a parametric plane. This leads to an explicit representation of the boundary curve. The Schottky model has been employed and the mappings written as ratios of products of Schottky–Klein prime functions. These functions are the natural generalizations of the well-known prime functions in a simply and doubly connected case, as discussed in the introduction. The conformal mappings are essentially "uniformizing functions" of the algebraic curves considered in [13]. Richardson [23] has presented an alternative conformal mapping method based on the use of Poincaré series to represent the mapping functions.

From a mathematical point of view, it is natural to ask questions about the convergence properties of the infinite products used in defining the Schottky–Klein prime functions. We have not studied such questions in detail. However, the boundaries of the quadrature domains obtained in the explicit examples of this paper have been found to be indistinguishable from those obtained using either the algebraic curve method of Crowdy [13] or the conformal mapping method based on Poincaré series introduced by Richardson [23]. We consider this to be direct evidence that convergence issues do not necessarily constitute an impediment to the practical use of the Schottky–Klein prime function in the reconstruction of quadrature domains from their quadrature data.

Appendix. The method of Richardson [23]. We shall now briefly describe an alternative construction of the quadrature domains via an approach using Poincaré series as expounded recently by Richardson [23]. This method also produces maps from circular regions of a parametric ζ-plane and requires the machinery of the Schottky groups associated with these circular regions. The method differs in the functional form, and representation, of the conformal mapping functions; Richardson constructs his maps as a ratio of two automorphic forms which are each constructed as Poincaré series.

DEFINITION A.1. *A Poincaré series associated with a given Schottky group is of the form*

$$T(\zeta) = \sum_{i=0}^{\infty} \frac{H(\theta_i(\zeta))}{(c_i\zeta + d_i)^{2m}}, \tag{71}$$

where

$$\theta_i(\zeta) \equiv \frac{a_i\zeta + b_i}{c_i\zeta + d_i}, \quad a_i d_i - b_i c_i = 1 \tag{72}$$

denotes the ith Mobius map of the Schottky group, $H(\zeta)$ is some rational function of which none of the poles is at a singular point of the Schottky group, and m is an integer. Provided $\zeta = \infty$ is not a singular point of the Schottky group, this series converges for all $m \geq 2$.

DEFINITION A.2. *A form $\phi(\zeta)$ is called an automorphic form with respect to the Schottky group if it has the property*

$$\phi(\theta_i(\zeta)) = (c_i\zeta + d_i)^{2m}\phi(\zeta) \tag{73}$$

for all maps θ_i of the Schottky group, where m is some integer.

If the Schottky group Θ is generated by g basic maps, and $\phi(\zeta)$ is an automorphic form with Z zeros and P poles in the fundamental region, then it is known that

$$Z - P = 2mg. \tag{74}$$

If $\phi(\zeta)$ is in fact an automorphic function, then we see $Z = P$.

Richardson's construction is to use two different choices of the rational functions $H_n(\zeta), H_d(\zeta)$ to form the respective Poincaré series for two automorphic forms $T_n(\zeta), T_d(\zeta)$ corresponding to the same value of $m \geq 2$. Then the ratio

$$\frac{T_n(\zeta)}{T_d(\zeta)} \tag{75}$$

and any constant multiple of this give the required automorphic function. Richardson's strategy is precisely the one described by Beardon [22] for the construction of meromorphic functions on compact Riemann surfaces.

Given a quadrature domain, there are a number of constraints on the relevant choices for $H_n(\zeta)$ and $H_d(\zeta)$. These are discussed in the context of a number of specific examples in Richardson [23]. Here we give very brief details of the construction for the triply connected example of section 6.1. Following Richardson, we choose $m = 2$. Recall that, in this example, there are poles at α_1 and α_3, where $\alpha_1 = -\alpha_3$ (purely real), and two at α_2 and α_4, where $\alpha_2 = -\alpha_4$ (purely imaginary). Also, $g = 2$. Following Richardson [23], we take $H_d(\zeta)$ to be 1. From (74) it follows that $T_d(\zeta)$ has eight zeros in the fundamental region. Due to the symmetry of the quadrature domain, we expect these zeros to be arranged in a pattern that is symmetric with respect to reflection in both axes. Thus we include in $H_n(\zeta)$ the polynomial factor $(\zeta^8 + a\zeta^6 + b\zeta^4 + c\zeta^2 + d)$, where the four real parameters a, b, c, d are to be chosen so that $T_n(\zeta)$ has the same zeros as $T_d(\zeta)$ in the fundamental region. Also due to the symmetry of the quadrature domain, we include a factor of ζ in the numerator of $H_n(\zeta)$. Finally, because none of the z_k in the associated quadrature identity are zero,

the map must be bounded as $\zeta \to \infty$; in fact, we require it to behave like $1/\zeta$ at ∞. So the denominator of $H_n(\zeta)$ must be of degree 10. Since we require simple poles at $\alpha_1, \alpha_2, \alpha_3$, and α_4, we include the factors $(\zeta - \alpha_j)$ for $j = 1, \ldots, 4$ in the denominator of $H_n(\zeta)$. We must then choose the remaining factors such that $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are still the only poles of the map in the fundamental region. This can be done by choosing to include $(\zeta - \theta_i(\alpha_j))$ for $i = 1, 2$ and $j = 1, 2, 3$ as the extra factors. Finally, the form for the map is the ratio (75) multiplied by some constant R to be determined.

In this representation there are *nine* unknowns, namely $R, \alpha_1, \alpha_2, \delta_1, \rho_1$ as well as a, b, c, d. We can specify a value for ρ_1, thus leaving eight unknowns. The eight equations for the remaining eight unknowns are (39) and (41) plus the four from the requirement that $T_n(\zeta)$ be zero at the zeros of $T_d(\zeta)$ in the fundamental region. Note that these zeros are not known explicitly and must therefore be found (numerically) as part of the solution.

There are a number of comments to be made concerning the two methods:

(i) The zeros of the map are not explicit in the Poincaré series representation, but are explicit in the Schottky–Klein prime function representation. The explicitness of the poles and zeros means that the general form of the required mapping can be written down immediately.

(ii) Once ρ is specified, the Poincaré series representation depends on *eight* parameters compared to only *five* parameters when the Schottky–Klein prime function representation is used. Moreover, the determination of the eight parameters in the Poincaré series representation in fact requires the solution of *twelve* nonlinear equations, owing to the fact that the four (distinct) zeros of $T_d(\zeta)$ (in the fundamental region) must be found numerically during the solution process. In the prime function representation, exactly five equations are solved for exactly five unknowns.

(iii) Two of the equations to be solved in either method are the residue equations (41). With the prime function representation, explicit formulae for the residues P_1, P_2 are straightforward to compute (cf. the formula for P_1 in (52)). However, care has to be taken when finding the analogous equations with the Poincaré series representation because the inclusion of factors such as $(\zeta - \theta_i(\alpha_j))$ in the denominator of $H_n(\zeta)$ can mean that more than one term in the sum $T_n(\zeta)$ contributes to the residue at each of the poles.

(iv) As discussed in detail by Richardson [23], the most convenient choice is $H_d(\zeta) = 1$. However, if the Schottky group has a loxodromic subgroup, then the Poincaré series $T_d(\zeta)$ with $H_d(\zeta) = 1$ does not converge. Richardson therefore proposes three possible remedial measures in this case, two of which are not implemented for various reasons. Such complications do not arise when using the Schottky–Klein prime function representation. In the latter case, it is simply necessary to pick the appropriate representation for the mapping, which can involve additional prefactors of the ratio of products of prime functions, as illustrated explicitly in the context of the example in section 6.6.

(v) A particular advantage of using the Schottky–Klein prime function representations concerns changes of topology, particularly in cases where the connectivity of the domain decreases. In the conformal mappings constructed in this paper, the functional form of the mappings as ratios of products of prime functions is the same; the only change is the definition of the relevant Schottky group.

REFERENCES

[1] P. Davis, *The Schwarz Function and Its Applications*, Carus Math. Monogr. 17, Math. Assoc. America, Washington, DC, 1974.

[2] D. Aharonov and H. Shapiro, *Domains on which analytic functions satisfy quadrature identities*, J. Anal. Math., 30 (1976), pp. 39–73.

[3] M. Sakai, *Quadrature Domains*, Lecture Notes in Math. 934, Springer-Verlag, New York, 1982.

[4] H. S. Shapiro, *Unbounded quadrature domains*, in Complex Analysis I (University of Maryland 1985–86), C. A. Bernstein, ed., Lecture Notes in Math. 1275, Springer-Verlag, Berlin, 1987, pp. 287–331.

[5] B. Gustafsson, *Quadrature identities and the Schottky double*, Acta. Appl. Math., 1 (1983), pp. 209–240.

[6] B. Gustafsson, *Singular and special points on quadrature domains from an algebro-geometric point of view*, J. Anal. Math., 51 (1988), pp. 91–117.

[7] S. Richardson, *Hele–Shaw flows with a free boundary produced by the injection of fluid into a narrow channel*, J. Fluid Mech., 56 (1972), pp. 609–618.

[8] V. M. Entov, P. I. Etingof, and D. Ya Kleinbock, *On nonlinear interface dynamics in Hele–Shaw flows*, European J. Appl. Math., 6 (1995), pp. 399–420.

[9] D. G. Crowdy, *Theory of exact solutions for the evolution of a fluid annulus in a rotating Hele–Shaw cell*, Quart. Appl. Math., 60 (2002), pp. 11–36.

[10] D. G. Crowdy and H. Kang, *Squeeze flow of multiply connected fluid domains in a Hele–Shaw cell*, J. Nonlinear Sci., 11 (2001), pp. 279–304.

[11] D. G. Crowdy, *A note on viscous sintering and quadrature identities*, European J. Appl. Math., 10 (1999), pp. 623–634.

[12] D. G. Crowdy, *A class of exact multipolar vortices*, Phys. Fluids, 11 (1999), pp. 2556–2564.

[13] D. G. Crowdy, *Multipolar vortices and algebraic curves*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 2337–2359.

[14] D. G. Crowdy, *Circulation-induced deformations of drops and bubbles: Exact two-dimensional models*, Phys. Fluids, 11 (1999), pp. 2836–2845.

[15] G. Valiron, *Cours d'Analyse Mathematique: Theorie des Fonctions*, Masson, Paris, 1947.

[16] S. Richardson, *Hele–Shaw flows with time-dependent free boundaries in which the fluid occupies a multiply connected region*, European J. Appl. Math., 5 (1994), pp. 97–122.

[17] L. Carrillo, J. Soriano, and J. Ortin, *Radial displacement of a fluid annulus in a rotating Hele–Shaw cell*, Phys. Fluids, 11 (1999), pp. 778–785.

[18] D. G. Crowdy and S. Tanveer, *A theory of exact solutions for annular viscous blobs*, J. Nonlinear Sci., 8 (1998), pp. 375–400; *Erratum*, J. Nonlinear Sci., 11 (2001), p. 237.

[19] S. Richardson, *Plane Stokes flow with time-dependent free boundaries in which the fluid occupies a doubly connected region*, European J. Appl. Math., 11 (2000), pp. 249–269.

[20] D. G. Crowdy, *Viscous sintering of unimodal and bimodal cylindrical packings with shrinking pores*, European J. Appl. Math., 14 (2003), pp. 421–445.

[21] H. F. Baker, *Abelian Functions: Abel's Theorem and the Allied Theory of Theta Functions*, Cambridge University Press, Cambridge, UK, 1995.

[22] A. F. Beardon, *A Primer on Riemann Surfaces*, London Math. Soc. Lecture Note Ser. 78, Cambridge University Press, Cambridge, UK, 1984.

[23] S. Richardson, *Hele–Shaw flows with time-dependent free boundaries involving a multiply connected fluid region*, European J. Appl. Math., 12 (2001), pp. 571–599.

[24] D. Mumford, C. Series, and D. Wright, *Indra's Pearls: The Vision of Felix Klein*, Cambridge University Press, Cambridge, UK, 2002.

[25] D. Mumford, *Tata Lectures on Theta*, Birkhäuser Boston, Boston, MA, 1983.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1360–1382

PARAMETRIC RESONANCE AND RADIATIVE DECAY OF DISPERSION-MANAGED SOLITONS*

DMITRY E. PELINOVSKY† AND JIANKE YANG‡

Abstract. We study propagation of dispersion-managed solitons in optical fibers which are modeled by the nonlinear Schrödinger equation with a periodic dispersion coefficient. When the dispersion variations are weak compared to the average dispersion, we develop perturbation series expansions and construct asymptotic solutions at the first and second orders of approximation. Due to a parametric resonance between the dispersion map and the dispersion-managed soliton, the soliton generates continuous-wave radiation leading to its radiative decay. The nonlinear Fermi golden rule for radiative decay of dispersion-managed solitons is derived from the solvability condition for the perturbation series expansions. Analytical results are compared to direct numerical simulations, and good agreement is obtained.

Key words. dispersion management, optical solitons, perturbation series, parametric resonance, radiative decay, Fermi golden rule

AMS subject classifications. 35Q55, 78M30, 78M35

DOI. 10.1137/S0036139903422358

1. Introduction.

This paper addresses the dispersion-periodic nonlinear Schrödinger (NLS) equation,

$$i\frac{\partial u}{\partial z} + \frac{m}{2\epsilon} D_\epsilon(z) \frac{\partial^2 u}{\partial t^2} + \frac{1}{2} D_0 \frac{\partial^2 u}{\partial t^2} + |u|^2 u = 0, \tag{1.1}$$

which models optical pulse propagation in dispersion-managed communication systems. Here $u \in \mathbb{C}$ is the wave envelope of the electromagnetic field, z (≥ 0) is the distance along the optical fiber, $t \in \mathbb{R}$ is the retarded time of the optical pulse, D_0 is the average dispersion, $D_\epsilon(z)$ is an ϵ-periodic mean-zero dispersion map, and m is the strength of the map variations. Lump amplification and losses are not included in the model (1.1) for the sake of simplicity.

Special solutions of the dispersion-periodic NLS equation (1.1) are called dispersion-managed (DM) solitons. They have been the subject of growing interest in recent literature [1, 2, 3]. DM solitons are periodic solutions of (1.1) in the form

$$u(z,t) = \Phi(z,t)\, e^{i\mu z}, \tag{1.2}$$

where $\Phi(z+\epsilon,t) = \Phi(z,t)$ and $\mu \in \mathbb{R}$. Existence of periodic solutions of (1.1) is studied with the normal-form transformations in the limit $\epsilon \to 0$ [4]. The normal-form transformations average the fast periodic variations of $\frac{1}{\epsilon} D_\epsilon(z)$ and reduce the dispersion-periodic NLS equation (1.1) to an integral NLS equation [5, 6]. Bound states of the integral NLS equation exist in the case of $D_0 > 0$ [7] and in the case

*Received by the editors February 6, 2003; accepted for publication (in revised form) November 1, 2003; published electronically May 20, 2004.
http://www.siam.org/journals/siap/64-4/42235.html

†Department of Mathematics, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada, L8S 4K1 (dmpeli@math.mcmaster.ca). The work of this author was supported by NSERC grant 5-36694.

‡Department of Mathematics, University of Vermont, Burlington, VT 05401 (jyang@emba.uvm.edu). The work of this author was supported by NSF grant DMS-9971712 and by a NASA EPSCoR minigrant.

$D_0 = 0$ [8]. Numerical results indicate nonexistence of bound states of the integral NLS equation in the case $D_0 < 0$ [9].

In what follows we consider the case $D_0 > 0$ only. Early papers by Nijhof et al. [11] reported numerically the existence of "exactly" periodic bound states in the dispersion-periodic NLS equation (1.1), which do not radiate any energy. Later, more careful numerics [3] showed that such bound states actually had nonvanishing radiation tails. Recent results of Yang and Kath [10] showed that exactly-periodic DM solitons do not exist in the dispersion-periodic NLS equation (1.1) because resonances in the perturbation series generate nonvanishing radiation tails. These tails can be extremely small in certain parameter regimes [10], but they do not vanish when $D_0 > 0$.

Radiation tails of DM solitons occur due to parametric resonance between the DM soliton and the periodic variation of the dispersion. This parametric resonance drains energy out of the DM soliton and leads to its radiative damping. Parametric resonances can be predicted by viewing the periodic term of (1.1) as an external forcing term:

$$i\frac{\partial u}{\partial z} + \frac{1}{2}D_0\frac{\partial^2 u}{\partial t^2} + |u|^2 u = -\frac{m}{2\epsilon}D_\epsilon(z)\frac{\partial^2 u}{\partial t^2}. \tag{1.3}$$

We expand $D_\epsilon(z)$ into a Fourier series,

$$D_\epsilon(z) = \sum_{n=-\infty}^{\infty} d_n\, e^{\frac{2\pi i n z}{\epsilon}}, \quad d_0 = 0, \quad d_{-n} = \bar{d}_n, \tag{1.4}$$

where $\bar{d}_n$ is the complex conjugate of d_n. When the nonlinear term in (1.3) is neglected and the averaged DM soliton $u(z,t) = \Phi(t)\, e^{i\mu z}$ is substituted into the right-hand side of (1.3), we find a solution of the linear inhomogeneous problem in the form of the Fourier series in z,

$$u(z,t) = \left(\sum_{n=-\infty}^{\infty} u_n(t)\, e^{\frac{2\pi i n z}{\epsilon}}\right)\, e^{i\mu z}. \tag{1.5}$$

The correction terms $u_n(t)$ take the form of Fourier integrals in t,

$$u_n(t) = -\frac{m d_n}{4\pi\epsilon}\int_{-\infty}^{\infty} \frac{\omega^2 \hat{\Phi}(\omega) e^{i\omega t} d\omega}{\frac{1}{2}D_0\omega^2 + \frac{2\pi n}{\epsilon} + \mu}, \tag{1.6}$$

where $\hat{\Phi}(\omega)$ is the Fourier transform of $\Phi(t)$. The inhomogeneous solution has resonant denominators at

$$\omega^2 = \omega_n^2 = -\frac{2}{D_0}\left(\mu + \frac{2\pi n}{\epsilon}\right) > 0. \tag{1.7}$$

Resonances are absent if $D_0 = 0$ and $\mu \neq -2\pi n/\epsilon$ for any integer n. This is the only case when DM soliton solutions (1.2) may exist in the dispersion-periodic NLS equation (1.1). In this case, the asymptotic representation of $\Phi(z,t)$ in (1.2) was found recently in [12] in the limit $\epsilon = \mathrm{O}(m) \gg 1$ with the use of the inverse scattering transform methods.

If $D_0 > 0$, sufficiently large negative terms of the Fourier series (1.5)–(1.6) are in resonance (1.7) for $n \leq -N_\mu$, where $N_\mu = \left\lceil \frac{\epsilon\mu}{2\pi} \right\rceil$ is the integer ceiling of $\frac{\epsilon\mu}{2\pi} > 0$. The periodic variations of the dispersion map $D_\epsilon(z)$ lead to a coupling of a bound state

and linear waves of the averaged dispersion map and to the energy transfer from the bound state to radiative waves. As a result, the pulse solution has resonant peaks in the spectrum $\hat{u}(z,\omega)$ at $\omega = \pm\omega_n$, and nonzero values of $u(z,t)$ in the far-field $|t| \gg 1$, as reported numerically in [3, 10].

Radiation damping of solitons in the presence of a weak sinusoidal dispersion variation was considered analytically in [13]. The radiative wave amplitudes and decay rates of solitons were computed by means of the soliton perturbation theory for the standard NLS equation. Dynamics of DM solitons was studied in [14, 15, 16] by variational and numerical methods. Recently, analytical and numerical studies of the same problem were undertaken in [10] by asymptotic beyond-all-orders methods in the limit $\epsilon = \mathrm{O}(m) \ll 1$. Radiation-tail amplitudes and decay rates of DM solitons were found to be exponentially small in this limit. It was also shown in [10] that radiation-tail amplitudes drop to near-zero values in certain windows on the m-axis.

We study here nonlinear parametric resonance of DM solitons for average-anomalous dispersion ($D_0 > 0$) in the limit $m \ll 1$, while we keep $\epsilon = \mathrm{O}(1)$. This is a different limit from the one studied in [10]. In this limit, the DM soliton decays much faster because radiation-tail amplitudes are only algebraically small in terms of $\mathrm{O}(m)$. The new feature of our analysis is that the periodic dispersion map $D_\epsilon(z)$ is allowed to be arbitrary in (1.4) as compared to a single sine function in [13]. Thus, our dispersion maps include the piecewise-constant dispersion map which is widely used in fiber communication systems.

Our analysis starts with the standard NLS equation (1.3) for $m = 0$, such that the right-hand side of (1.3) is treated as a small perturbation. The first-order perturbation theory describes generation of linear waves due to parametric resonances (1.7), and the second-order perturbation theory leads to the decay rate of DM solitons. Methods of our analysis are similar to the soliton perturbation theory in [13], but our calculations are more systematic. We find that the DM soliton decays according to a nonlinear Fermi golden rule, which generalizes the Fermi golden rule for radiative decay of bound states in the linear Schrödinger equation with a time-periodic potential. Rigorous analysis of decay rates in the linear Schrödinger equation was recently considered in [17, 18], where the bound states were supported by a time-dependent periodic potential in [17] and by a time-independent potential in [18].

This paper is structured as follows. Section 2 contains perturbation series expansions and derivations of the Fermi golden rule for DM solitons. Section 3 is devoted to analytical approximations of radiative decay of DM solitons. Section 4 describes a comparison between the analytical and numerical results. Section 5 concludes the paper. Appendices A and B describe technical details of the first-order solution in the perturbation series expansions.

2. Perturbation series expansions. We start with the dispersion-periodic NLS equation in the form (1.3), where ϵ is finite and m is small. If $D_0 > 0$, we employ the following rescaling of variables:

$$z = \epsilon\hat{z}, \quad u = \frac{\hat{u}}{\sqrt{\epsilon}}, \quad t = \sqrt{\epsilon D_0}\,\hat{t}, \quad m = \epsilon D_0 \hat{m}. \tag{2.1}$$

When the hats are dropped, (1.3) becomes

$$iu_z + \frac{1}{2}u_{tt} + |u|^2 u = -\frac{m}{2}D_1(z)u_{tt}, \tag{2.2}$$

where the dispersion map $D_1(z)$ has unit period. In other words, we have normalized ϵ and D_0 in (1.3) so that $\epsilon = 1$ and $D_0 = 1$.

When $m = 0$, the standard NLS equation (2.2) has a bound state:

$$u(z,t) = \Phi(t;\mu)e^{i\mu z}, \tag{2.3}$$

where $\mu > 0$ and $\Phi(t;\mu) = \sqrt{2\mu}\ \mathrm{sech}\left(\sqrt{2\mu}\, t\right)$. When $m \neq 0$, the NLS soliton (2.3) would generate radiative tails and decay accordingly. Parameter μ of the NLS soliton (2.3) changes in z, such that the z-dependence of $\mu(z)$ serves as a condition for Poincaré continuation of the perturbation series for $u(z,t)$ in powers of m. The Fermi golden rule of radiative decay of NLS solitons follows from the dynamical equation for $\mu = \mu(z)$. In order to formalize this qualitative picture, we employ the transformation

$$u(z,t) = U(z,t;\mu(z))e^{i\int_0^z \mu(z')dz'}, \tag{2.4}$$

where $U(z,t;\mu)$ solves the problem

$$i\frac{\partial U}{\partial z} + i\dot{\mu}\frac{\partial U}{\partial \mu} - \mu U + \frac{1}{2}\frac{\partial^2 U}{\partial t^2} + |U|^2 U = -\frac{m}{2}D_1(z)\frac{\partial^2 U}{\partial t^2} \tag{2.5}$$

with the initial data $U(0,t;\mu_0) = \Phi(t;\mu_0)$ and $\mu(0) = \mu_0$. The transformation (2.4) describes the adiabatically varying orbit of the NLS soliton (2.3). We present the asymptotic solution of (2.5) as a perturbation series for $U(z,t;\mu)$ and $\mu(z)$ in powers of m:

$$U(z,t;\mu) = \sum_{k=0}^{\infty} m^k U^{(k)}(z,t;\mu) \tag{2.6}$$

and

$$\dot{\mu} = \sum_{k=1}^{\infty} m^{2k}\Gamma^{(2k)}(\mu), \tag{2.7}$$

where $\Gamma^{(2k)}(\mu)$ are corrections of the Fermi golden rule for radiative decay of NLS solitons. Substitution of (2.6)–(2.7) into (2.5) produces a chain of equations for corrections of the perturbation series. At the leading, first and second orders, the chain of perturbative equations takes the form

$$i\frac{\partial U^{(0)}}{\partial z} - \mu U^{(0)} + \frac{1}{2}\frac{\partial^2 U^{(0)}}{\partial t^2} + |U^{(0)}|^2 U^{(0)} = 0, \tag{2.8}$$

$$i\frac{\partial U^{(1)}}{\partial z} - \mu U^{(1)} + \frac{1}{2}\frac{\partial^2 U^{(1)}}{\partial t^2} + 2|U^{(0)}|^2 U^{(1)} + U^{(0)2}\bar{U}^{(1)} = -\frac{1}{2}D_1(z)\frac{\partial^2 U^{(0)}}{\partial t^2}, \tag{2.9}$$

and

$$\begin{aligned} i\frac{\partial U^{(2)}}{\partial z} - \mu U^{(2)} + \frac{1}{2}\frac{\partial^2 U^{(2)}}{\partial t^2} + 2|U^{(0)}|^2 U^{(2)} + U^{(0)2}\bar{U}^{(2)} \\ = -i\Gamma^{(2)}(\mu)\frac{\partial U^{(0)}}{\partial \mu} - \frac{1}{2}D_1(z)\frac{\partial^2 U^{(1)}}{\partial t^2} - 2|U^{(1)}|^2 U^{(0)} - U^{(1)2}\bar{U}^{(0)}. \end{aligned} \tag{2.10}$$

Initial conditions for these equations are

$$U^{(0)}(0,t;\mu_0) = \Phi(t;\mu_0), \qquad \mu(0) = \mu_0, \tag{2.11}$$

and

$$U^{(k)}(0,t;\mu_0) = 0, \quad k \geq 1. \tag{2.12}$$

Order O(1). The nonlinear equation (2.8) at order O(1) with initial data (2.11) has a unique solution, $U^{(0)}(z,t;\mu) = \Phi(t;\mu)$, which is the NLS soliton with the adiabatic change of $\mu = \mu(z)$.

Order O(m). The linear inhomogeneous equation (2.9) at order O(m) has the Fourier series solution

$$U^{(1)}(z,t;\mu) = \sum_{n=-\infty}^{\infty} U_n^{(1)}(z,t;\mu)\, e^{2\pi i n z}, \tag{2.13}$$

where $U_0^{(1)} = 0$ and $(U_n^{(1)}, \bar{U}_{-n}^{(1)})$ at $n \geq 1$ solve the coupled equations

$$\begin{aligned} & i\frac{\partial U_n^{(1)}}{\partial z} - (\mu + 2\pi n)\, U_n^{(1)} + \frac{1}{2}\frac{\partial^2 U_n^{(1)}}{\partial t^2} + \Phi^2(t;\mu)\, \left(2U_n^{(1)} + \bar{U}_{-n}^{(1)}\right) \\ & \quad = -\frac{d_n}{2}\Phi''(t;\mu), \end{aligned} \tag{2.14}$$

$$\begin{aligned} & -i\frac{\partial \bar{U}_{-n}^{(1)}}{\partial z} - (\mu - 2\pi n)\, \bar{U}_{-n}^{(1)} + \frac{1}{2}\frac{\partial^2 \bar{U}_{-n}^{(1)}}{\partial t^2} + \Phi^2(t;\mu)\, \left(2\bar{U}_{-n}^{(1)} + U_n^{(1)}\right) \\ & \quad = -\frac{d_n}{2}\Phi''(t;\mu). \end{aligned} \tag{2.15}$$

It follows from (2.12) that the system (2.14)–(2.15) is supplemented with zero initial conditions: $U_n^{(1)}(0,t;\mu_0) = 0$ for any $|n| \geq 1$. Solutions of the system (2.14)–(2.15) are constructed in Appendix A with the use of the spectral decomposition for a linearized NLS operator [19, 20]. Asymptotic limits of the correction terms $U_n^{(1)}(z,t;\mu)$ are obtained in Appendix B with the use of generalized functions. These calculations show that the continuous-wave radiation in the solution $U_n^{(1)}(z,t;\mu)$ at large distance z and time t is given by the following expression [see (A.1) and (B.9)]:

$$\lim_{|t|\to\infty, z\to\infty} U_{-n}^{(1)} = -\frac{\pi i\sqrt{2\mu}\, d_{-n}(k_n+i)^2}{4k_n}\, \operatorname{sech}\frac{\pi k_n}{2}\, e^{i\sqrt{2\mu}\, k_n |t|}, \quad n \geq N_\mu, \tag{2.16}$$

and

$$\lim_{|t|\to\infty, z\to\infty} U_{-n}^{(1)} = 0, \quad n < N_\mu, \tag{2.17}$$

where

$$k_n = \sqrt{\frac{2\pi n}{\mu} - 1} > 0, \qquad N_\mu = \left[\frac{\mu}{2\pi}\right]. \tag{2.18}$$

This result will be used at order O(m^2) to calculate the decay rate $\Gamma^{(2)}(\mu)$ of DM solitons.

Order O(m^2). Solution of the linear inhomogeneous equation (2.11) at order O(m^2) can also be represented by the Fourier series:

$$U^{(2)}(z,t;\mu) = \sum_{n=-\infty}^{\infty} U_n^{(2)}(z,t;\mu)\, e^{2\pi i n z}. \tag{2.19}$$

Since the right-hand side of (2.11) has a nonzero mean term in z, the nonzero mean term $U_0^{(2)}(z,t;\mu)$ satisfies the inhomogeneous equation

$$i\frac{\partial U_0^{(2)}}{\partial z} - \mu U_0^{(2)} + \frac{1}{2}\frac{\partial^2 U_0^{(2)}}{\partial t^2} + \Phi^2(t;\mu)\left(2U_0^{(2)} + \bar{U}_0^{(2)}\right) = -i\Gamma^{(2)}(\mu)\frac{\partial \Phi(t;\mu)}{\partial \mu}$$
$$- \sum_{n=-\infty}^{\infty}\left(\frac{1}{2}d_{-n}\frac{\partial^2 U_n^{(1)}}{\partial t^2} + 2\Phi(t;\mu)U_n^{(1)}\bar{U}_n^{(1)} + \Phi(t;\mu)U_n^{(1)}U_{-n}^{(1)}\right). \tag{2.20}$$

The mean term in the right-hand side of (2.20) leads to a secular growth of $U_0^{(2)}(z,t;\mu)$ in z unless the right-hand side of (2.11) is orthogonalized with respect to eigenfunctions of the kernel of the linearized operator (the Fredholm alternative theorem). The correction $\Gamma^{(2)}(\mu)$ is found from the orthogonalization constraint as follows. Projecting (2.20) onto $\Phi(t;\mu)$ and subtracting a complex conjugate equation, we obtain a single equation under the condition that $U_0^{(2)}(z,t;\mu)$ is bounded in t:

$$i\frac{\partial}{\partial z}\langle \Phi, U_0^{(2)} + \bar{U}_0^{(2)}\rangle = -i\Gamma^{(2)}(\mu)\frac{\partial}{\partial \mu}\langle \Phi, \Phi\rangle - \frac{1}{2}\sum_{n=-\infty}^{\infty}\langle \Phi'', d_{-n}U_n^{(1)} - \bar{d}_{-n}\bar{U}_n^{(1)}\rangle$$
$$- \sum_{n=-\infty}^{\infty}\langle \Phi^2, U_n^{(1)}U_{-n}^{(1)} - \bar{U}_n^{(1)}\bar{U}_{-n}^{(1)}\rangle, \tag{2.21}$$

where $\langle f, g\rangle$ is the standard inner product in $L^2(\mathbb{R})$:

$$\langle f, g\rangle = \int_{-\infty}^{\infty}\bar{f}(t)g(t)dt.$$

The right-hand side of (2.21) can be simplified with the use of the system (2.14)–(2.15) as follows:

$$i\frac{\partial}{\partial z}|U_n^{(1)}|^2 + \frac{1}{2}\frac{\partial}{\partial t}\left(\bar{U}_n^{(1)}\frac{\partial U_n^{(1)}}{\partial t} - U_n^{(1)}\frac{\partial \bar{U}_n^{(1)}}{\partial t}\right) = -\frac{1}{2}\Phi''(t;\mu)\left(\bar{d}_{-n}\bar{U}_n^{(1)} - d_{-n}U_n^{(1)}\right)$$
$$-\Phi^2(t;\mu)\left(\bar{U}_n^{(1)}\bar{U}_{-n}^{(1)} - U_n^{(1)}U_{-n}^{(1)}\right). \tag{2.22}$$

As a result, the projection formula (2.21) takes the form

$$i\frac{\partial}{\partial z}\left[\langle \Phi, U_0^{(2)} + \bar{U}_0^{(2)}\rangle + \sum_{n=-\infty}^{\infty}\langle U_n^{(1)}, U_n^{(1)}\rangle\right]$$
$$= -i\Gamma^{(2)}(\mu)\frac{\partial}{\partial \mu}\langle \Phi, \Phi\rangle - \frac{1}{2}\sum_{n=-\infty}^{\infty}\left(\bar{U}_n^{(1)}\frac{\partial U_n^{(1)}}{\partial t} - U_n^{(1)}\frac{\partial \bar{U}_n^{(1)}}{\partial t}\right)\Bigg|_{t=-\infty}^{t=\infty}. \tag{2.23}$$

It follows from (B.1) and (B.5) of Appendix B for finite t that $\langle U_n^{(1)}, U_n^{(1)}\rangle$ becomes z-independent in the limit $z \to \infty$. It also follows from (2.16) that the limiting values of $U_n^{(1)}$ at $|t| \gg 1$ are nonzero and constant in the limit $z \to \infty$ for large negative $n \le -N_\mu$, where $N_\mu = [\frac{\mu}{2\pi}]$ is the integer ceiling of $\frac{\mu}{2\pi} > 0$. Therefore, we conclude that the correction term $U_0^{(2)}(z,t;\mu)$ is free of secular terms in z in the limit $z \to \infty$ only if $\Gamma^{(2)}(\mu)$ is defined by the nonlinear Fermi golden rule,

$$\Gamma^{(2)}(\mu) = -\frac{\sqrt{2\mu}}{4i}\sum_{n=-\infty}^{-N_\mu}\lim_{z\to\infty}\left(\bar{U}_n^{(1)}\frac{\partial U_n^{(1)}}{\partial t} - U_n^{(1)}\frac{\partial \bar{U}_n^{(1)}}{\partial t}\right)\Bigg|_{t=-\infty}^{t=\infty}, \tag{2.24}$$

where we use the formula

$$\frac{\partial}{\partial \mu}\langle \Phi, \Phi \rangle = \frac{\sqrt{2}}{\sqrt{\mu}}. \tag{2.25}$$

Using (2.16), we transform (2.24) to the explicit form

$$\Gamma^{(2)}(\mu) = -\frac{\pi^2 \mu^2}{4} \sum_{n=N_\mu}^{\infty} \frac{|d_n|^2 (1+k_n^2)^2}{k_n} \operatorname{sech}^2\left(\frac{\pi k_n}{2}\right). \tag{2.26}$$

Assuming $\lim_{n\to\infty} |d_n|^2 = 0$, the infinite series in (2.26) converges when $\mu \neq \mu_n \equiv 2\pi n$, where n is any positive integer. Critical resonances occur at $\mu = \mu_n$, when $k_n = 0$. This case will be studied in more detail in section 3.

The correction term $U_0^{(2)}(z,t;\mu)$ solves the linear inhomogeneous equation (2.20) under the constraint (2.26). The right-hand side of (2.20) is bounded but nondecaying in the limits $|t| \to \infty$ and $z \to \infty$ because of the asymptotic limit (2.16). The nondecaying terms in (2.16) are not in resonance with the left-hand side of (2.20) since $k_n^2 + 1 = \frac{2\pi n}{\mu} \neq 0$ for $n \neq 0$. As a result, we conclude from (2.20) that a solution $U_0^{(2)}(z,t;\mu)$ exists and is bounded in the limit $z \to \infty$ under the condition (2.26). Similarly, one can show that a bounded solution exists for any $U_n^{(2)}(z,t;\mu)$ where n is an integer; i.e., the bounded right-hand side term $D_1(z)U_{tt}^{(1)}$ in (2.11) is not in resonance with the left-hand side of (2.11). This completes consideration of the order $\mathrm{O}(m^2)$ of the perturbation series expansions.

3. Decay rates of DM solitons. Formula (2.26) generalizes the Fermi golden rule for radiative decay of bound states in a linear Schrödinger equation with time-periodic potentials [17, 18]. The correction term $\Gamma^{(2)}(\mu)$ is always negative, such that the dynamical system (2.7) exhibits a simple behavior of a monotonic decay of $\mu(z)$ to zero, starting with any initial value $\mu(0) = \mu_0 > 0$. Therefore, the DM soliton decays due to parametric resonances and radiative losses. The decay rate of $\mu(z)$ depends on the nonlinear function $\Gamma^{(2)}(\mu)$ in (2.26). Here we study solutions of the truncated equations (2.7) and (2.26) at the order of $\mathrm{O}(m^2)$:

$$\frac{d\mu}{dz} = -m^2 \pi^4 \sum_{n=N_\mu}^{\infty} \frac{|d_n|^2 n^2}{k_n} \operatorname{sech}^2\left(\frac{\pi k_n}{2}\right). \tag{3.1}$$

We choose the dispersion coefficient $D_1(z)$ as a two-step symmetric function,

$$D_1(z) = \begin{cases} 1, & \mathrm{mod}(z,1) \in \left(0, \frac{1}{4}\right) \cup \left(\frac{3}{4}, 1\right), \\ -1, & \mathrm{mod}(z,1) \in \left(\frac{1}{4}, \frac{3}{4}\right). \end{cases} \tag{3.2}$$

For this dispersion map, the DM soliton is chirp-free at $\mathrm{mod}(z,1) = 0$ and $\mathrm{mod}(z,1) = \frac{1}{2}$ (see [21], for instance). The Fourier coefficients d_n for this dispersion map are

$$d_n = \frac{2(-1)^{n+1}}{\pi n} \sin\left(\frac{\pi n}{2}\right). \tag{3.3}$$

As a result, the dynamical system (3.1) takes an explicit form,

$$\frac{d\mu}{dz} = -4\pi^2 m^2 \sum_{\substack{n=N_\mu \\ n \text{ odd}}}^{\infty} \frac{1}{k_n} \operatorname{sech}^2\left(\frac{\pi k_n}{2}\right), \qquad k_n = \sqrt{\frac{2\pi n}{\mu} - 1}. \tag{3.4}$$

This equation is the main result of this paper. It describes the radiation damping of DM solitons in the normalized dispersion-periodic NLS equation (2.2) with piecewise-constant dispersion maps. It is asymptotically accurate when $m \ll 1$ and μ is not close to critical values $\mu_n = 2\pi n$, where n is a positive odd integer. If $\mu \approx \mu_n$, critical resonances occur and radiation tails become large, such that the perturbation series breaks down in a strict mathematical sense. The decay-rate function $\Gamma^{(2)}(\mu)$ in the right-hand side of (3.4) for $m = 1$ is plotted in Figure 1.

A similar equation for the radiative decay of DM solitons in the presence of weak sinusoidal dispersion variation has been derived in [13]. In that paper, only one term appears in the right-hand side of (3.1) since the Fourier series for $D_\epsilon(z)$ in (1.4) contains only a single term in that case.

Below, we analyze the dynamical equation (3.4) under three different limits: (i) $\mu \ll 1$; (ii) $\mu = \mathrm{O}(1)$; (iii) $\mu \gg 1$.

1. *Limit of small values of* μ. When $\mu \ll 1$, all terms in the series in (3.4) are present since $N_\mu = 1$. But only the first term with $n = 1$ dominates, since the higher terms are exponentially smaller in μ compared to the (exponentially small) first term. Therefore, the dynamical equation (3.4) can be truncated at the first term and simplified as

$$\frac{d\mu}{dz} = -\frac{16\pi^2 m^2 e^{-\pi k_1}}{k_1}, \qquad k_1 = \sqrt{\frac{2\pi}{\mu} - 1}. \tag{3.5}$$

Comparison between numerical solutions of the simplified equation (3.5) and the original equation (3.4) indicates that the simplified equation (3.5) gives a very good approximation to the original equation (3.4) not only for $\mu \ll 1$, but also for $\mu < 2\pi$ (see Figures 2 and 3).

In the limit $\mu \to 0$, methods of exponential asymptotics can be developed after further simplification of the dynamical equation (3.5):

$$\frac{d\mu}{dz} = -\alpha m^2 \mu^{1/2} \exp\left(-\frac{\beta}{\mu^{1/2}}\right), \tag{3.6}$$

where $\alpha = 4(2\pi)^{3/2}$ and $\beta = \pi(2\pi)^{1/2}$. In this limit, the radiation damping of DM solitons and the continuous-tail radiation emitted by the DM soliton are exponentially weak. This agrees with the asymptotic beyond-all-orders calculations by Yang and Kath [10]. A similar situation occurs in the dynamics of embedded solitons in the perturbed integrable fifth-order KdV equation in the small velocity limit [22].

Results of [10] are valid when $\mu \ll 1$ and m is arbitrary, while our results are valid when $m \ll 1$ and μ arbitrary. In the regime of common validity, i.e., $m \ll 1$ and $\mu \ll 1$, the two results match each other, as shown next. When $\mu \ll 1$, the radiation field is dominated by the $n = -1$ term in the Fourier-series solution (2.13) for $U^{(1)}(z, t; \mu)$. The amplitude of this radiation field is thus given asymptotically from (2.16) as

$$u_{\mathrm{rad}} = 2m\pi^{\frac{1}{2}} \exp\left(-\frac{\pi^{3/2}}{\sqrt{2\mu}}\right). \tag{3.7}$$

Due to the rescaling of variables (2.1) and different notations, results of [10] need to be reformulated. In the present notations, the amplitude of the radiation field obtained in [10] is in the form

$$u_{\mathrm{rad}} = \frac{1}{2} C\pi^{\frac{1}{2}} \exp\left(-\frac{\pi^{3/2}}{\sqrt{2\mu}}\right), \tag{3.8}$$

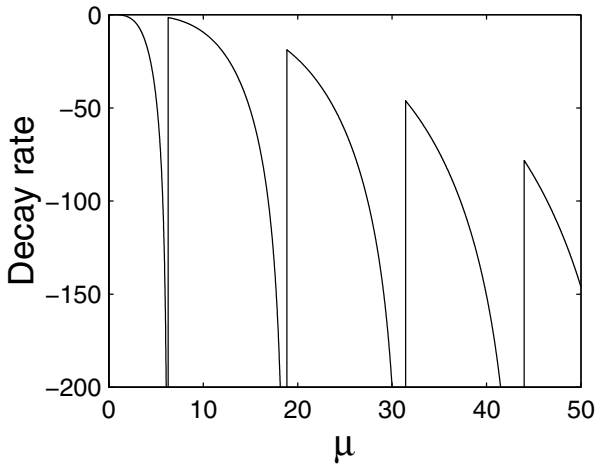


FIG. 1. *Decay rate* $\Gamma^{(2)}(\mu)$ *of DM solitons versus the parameter* μ *as in* (3.4) *for* $m = 1$.

where C is a dispersion map-dependent constant given in Figure 1 of [10]. Our parameter m is equal to $\frac{1}{2}(\sigma_2 - \sigma_1)$ of [10]. When $m \ll 1$, inspection of Figure 1 in [10] shows that $C \approx 2(\sigma_2 - \sigma_1) = 4m$. Thus, in the limits $m, \mu \ll 1$, the radiation field (3.7) from our analysis agrees perfectly with (3.8) from [10]. We note that Yang and Kath [10] also found windows of low radiation field at large values of m. Since our results are valid only in the limit $m \ll 1$ and up to the order of $\mathrm{O}(m^2)$, the low-radiation windows cannot be recovered in our analysis unless the perturbation series (2.6) and (2.7) are extended to at least $\mathrm{O}(m^4)$.

The dynamical equation (3.6) can be integrated with the help of the Laplace method as follows:

$$\frac{1}{2}\alpha m^2(z + z_0) = \exp\left(\frac{\beta}{\mu^{1/2}}\right)\left[\frac{\mu}{\beta} + \mathrm{O}(\mu^{3/2})\right], \tag{3.9}$$

where z_0 is a constant of integration. The leading-order asymptotic solution for $\mu(z)$ in the limit $\mu \to 0$ is derived from (3.9) in the form

$$\mu(z) = \left[\frac{\beta}{\log\left[\frac{\alpha m^2}{2\beta}(z + z_0)\log^2(z + z_0)\right] + \mathrm{O}\left(\frac{1}{\log(z+z_0)}\right)}\right]^2. \tag{3.10}$$

As $z \to \infty$, the parameter $\mu(z)$ decays logarithmically as

$$\mu(z) \sim \frac{\beta^2}{\log^2 z}\left[1 - 4\frac{\log\log z}{\log z}\right]. \tag{3.11}$$

This logarithmic decay of bound states has been reported previously for internal modes of envelope solitons in [23]. Logarithmic decay is associated with an exponentially small Fermi golden rule for exponentially small radiative waves.

2. *Solutions near critical values* μ_n. If $\mu(0) > 2\pi$, the decay of DM solitons always leads to the point where the parameter μ has to pass through a critical value $\mu_n = 2\pi n$, where n is a positive odd integer. When this happens, the radiation field becomes large, and the perturbation-series solution formally breaks down. Consequently, the

solution of the dynamical equation (3.4) may no longer give a good approximation to the true solution. However, numerical results indicate that the solution of (3.4) still agrees qualitatively with the solution of the full equation (2.2) (see Figure 4). Here we derive the solution of the dynamical equation (3.4) when it passes through a single critical value $\mu = \mu_N$ at $z = z_N$.

Since N_μ is an integer ceiling of $\frac{\mu}{2\pi}$, and N_μ is odd, the one-sided limit $z \to z_N^-$ is nonsingular, and the parameter $\mu(z)$ approaches μ_N with a linear slope:

$$\mu(z) = \mu_N - \mu_N'(z - z_N) + \mathrm{O}(z - z_N)^2, \qquad z < z_N, \tag{3.12}$$

where

$$\mu_N' = 4m^2\pi^2 \sum_{\substack{n=N_\mu \\ n \text{ odd}}}^{\infty} \frac{1}{k_n} \mathrm{sech}^2\left(\frac{\pi k_n}{2}\right), \tag{3.13}$$

and k_n are all computed at $\mu = \mu_N$. Once the parameter $\mu(z)$ passes below μ_N, a singular term with $n = N$ appears in the dynamical equation (3.4) because $k_N = 0$ at $\mu = \mu_N$. The leading-order asymptotic approximation for the solution $\mu(z)$ for $z > z_N$ takes the form

$$\mu(z) = \mu_N - [\alpha(z - z_N)]^{2/3} + \mathrm{O}(z - z_N), \qquad z > z_N, \tag{3.14}$$

where $\alpha = 6m^2\pi^2\sqrt{\mu_N}$. The slope of $\mu(z)$ is infinite in the limit $z \to z_N^+$, but the solution $\mu(z)$ is still continuous at $z = z_N$. The asymptotic solution (3.14) describes a sharp drop in the amplitude of the DM soliton after it passes through a critical resonance value μ_N.

3. *Limit of large values of* μ. When $\mu \gg 1$, the dynamical equation (3.4) can also be simplified. Using the formula

$$k_{n+2}^2 - k_n^2 = \frac{4\pi}{\mu} \tag{3.15}$$

and the Riemann sum approximation for the integral with areas of rectangles, we approximate the sum as

$$\begin{aligned} \frac{1}{\mu} \sum_{\substack{n=N_\mu \\ n \text{ odd}}}^{\infty} \frac{1}{k_n} \mathrm{sech}^2\left(\frac{\pi k_n}{2}\right) &= \frac{1}{2\pi} \sum_{\substack{n=N_\mu \\ n \text{ odd}}}^{\infty} \mathrm{sech}^2\left(\frac{\pi k_n}{2}\right) \frac{(k_{n+2} + k_n)}{2k_n}(k_{n+2} - k_n) \\ &\approx \frac{1}{2\pi} \int_{k_0(\mu)}^{\infty} \mathrm{sech}^2\left(\frac{\pi k}{2}\right) dk, \end{aligned} \tag{3.16}$$

where $k_0(\mu) = k_{N_\mu}$ such that $0 < k_0(\mu) < 1$. In this approximation, the dynamical system (3.4) simplifies to the form

$$\frac{d\mu}{dz} = -4m^2\mu\left[1 - \tanh(k_0(\mu))\right]. \tag{3.17}$$

Using the comparison principle for (3.17), we conclude that DM solitons decay with a linear decay rate when $\mu(z) \gg 1$:

$$\mu(0)\, \exp\left(-4m^2 z\right) \le \mu(z) \le \mu(0)\, \exp\left(-4m^2\alpha_0 z\right), \tag{3.18}$$

where $\alpha_0 = 1 - \tanh 1 > 0$.

We note that when $\mu(0) \gg 1$, the monotonic decay of $\mu(z)$ passes through many critical values, where radiation amplitudes are large. As a result, the asymptotic solution (3.18) may not give a good quantitative approximation to the true solution. Nevertheless, the solution (3.18) still describes qualitatively the decay of DM solitons for $\mu(0) \gg 1$ (see Figure 5).

4. Numerical simulations of DM solitons. Here we directly simulate the normalized dispersion-periodic NLS equation (2.2) and compare numerical solutions with the above analytical solutions. Our numerical method uses the fast Fourier transform (FFT) to compute the derivatives in t, and the fourth-order Runge–Kutta scheme to advance in z. At the values of z where the dispersion has a discontinuity (i.e., $\mathrm{mod}(z,1) = \frac{1}{4}$ and $\mathrm{mod}(z,1) = \frac{3}{4}$), the stepsize Δz is reduced so that the overall fourth-order accuracy in z is assured. To eliminate radiation reflection at the boundaries of the t-interval, damping boundary conditions are used. Our results are checked with longer t-intervals, more grid points in t, and smaller stepsize Δz, and the results are found to remain the same.

Our numerical simulation starts with the initial condition of a standard (unchirped) NLS soliton:

$$u(0,t) = \sqrt{2\mu_0}\,\mathrm{sech}\sqrt{2\mu_0}\,t. \tag{4.1}$$

It is known that DM solitons are unchirped in the middle point of each constant-dispersion segment, i.e., at $\mathrm{mod}(z,1) = 0$ and $\mathrm{mod}(z,1) = 1/2$ in the present case. Thus, when the unchirped NLS soliton (4.1) is launched at $z = 0$, the radiation emission is minimal compared to that of chirped solitons.

Below, we describe numerical computations with $m = 0.1$ and four different values of $\mu(0)$.

1. *Figure 2:* $\mu(0) = 1$. Figure 2(a) shows the soliton amplitude versus distance z. We see that this soliton's amplitude is oscillating (breathing) with unit period, which is the period of the dispersion map $D_1(z)$. This behavior is a signature of DM solitons. The evolution of the average soliton amplitude in z is plotted in Figure 2(b). This average amplitude is numerically calculated for each unit distance z as the average between the maximum and minimum amplitudes. It is clear from Figure 2(b) that the DM soliton slowly decays due to the parametric resonance between the soliton and the dispersion map, in accordance with the analytical prediction above. Also in Figure 2(b), the analytical values of the average soliton amplitude $\sqrt{2\mu}$ obtained from the dynamical equation (3.4) and its simplified version (3.5) are plotted as circles "o" and crosses "x," respectively. We see that both analytical equations (3.4) and (3.5) agree with numerical values and with each other extremely well. This comparison confirms that the dynamical equation (3.4) for radiation damping of DM solitons is asymptotically accurate in the case $m \ll 1$ and $\mu(0) < \mu_1 = 2\pi$, and that the simplified equation (3.5) is a very good approximation to (3.4) not only for $\mu \ll 1$, but also for $\mu = \mathrm{O}(1)$. The soliton profile at $z = 2000$ is shown in Figure 2(c) in a logarithmic scale. We clearly see the central DM-pulse is flanked by continuous-wave radiation. The radiation amplitude is nearly constant. This is because the radiation is excited mainly by the lowest-order resonance with $n = 1$ in (3.4), and the radiation field is dominated by the lowest-order radiative waves with $n = 1$ in (2.16). At $z = 2000$, the parameter μ can be inferred from Figure 2(b) as roughly 0.6731, and the radiation field should be dominated by waves

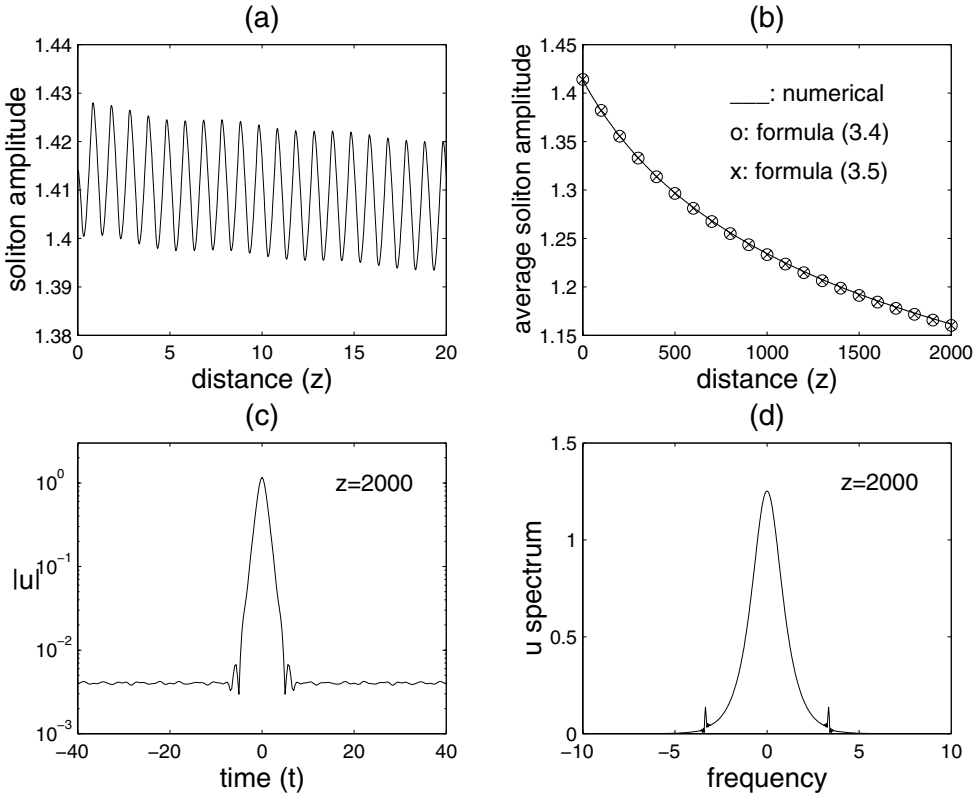


FIG. 2. *Numerical evolution of the DM soliton with* $m = 0.1$ *and* $\mu(0) = 1$. (a) *Soliton amplitude versus distance* z. (b) *Average soliton amplitude versus* z*: numerical results (solid curve); analytical average soliton amplitude* $\sqrt{2\mu}$ *from* (3.4) *(circles); analytical average soliton amplitude* $\sqrt{2\mu}$ *from* (3.5) *(crosses).* (c) *Solution profile at* $z = 2000$. (d) *Fourier spectrum of the solution at* $z = 2000$.

with frequencies $\pm\sqrt{2\mu}\, k_1 \approx \pm 3.35$, according to (2.16). This is confirmed in Figure 2(d), where the solution spectrum at $z = 2000$ is shown. This spectrum has two spikes at frequencies ± 3.33, which are due to the radiation field. The locations of these frequency spikes are in excellent agreement with the theoretical values ± 3.35.

2. *Figure* 3*:* $\mu(0) = 6$. In this case, the initial value of μ is close to but still below the lowest critical resonance value $\mu_1 = 2\pi$. Therefore, we expect that the radiation field would be larger, and the theoretical approximation (3.4) for the DM soliton less accurate. This is indeed the case. In Figure 3(a), the soliton amplitude versus distance z is plotted. We see that the amplitude oscillates irregularly, and the period of oscillations is not equal to the unit dispersion map period any-more. This is an indication that the central pulse has deviated from the DM soliton. However, our analytical solution for the average soliton amplitude $\sqrt{2\mu}$, which is calculated from the dynamical equation (3.4), still gives a very reasonable approximation to the true solution (see the dashed line in Figure 3(a)). We have also compared solutions from the dynamical equation (3.4) and its simplified form (3.5) for the present set of parameters, and found that the two solutions differ by only less than 6%. Thus, over a wide range of μ values below the critical resonance $\mu_1 = 2\pi$, the simplified equation (3.5) gives a very good approximation to the original dynamical equation

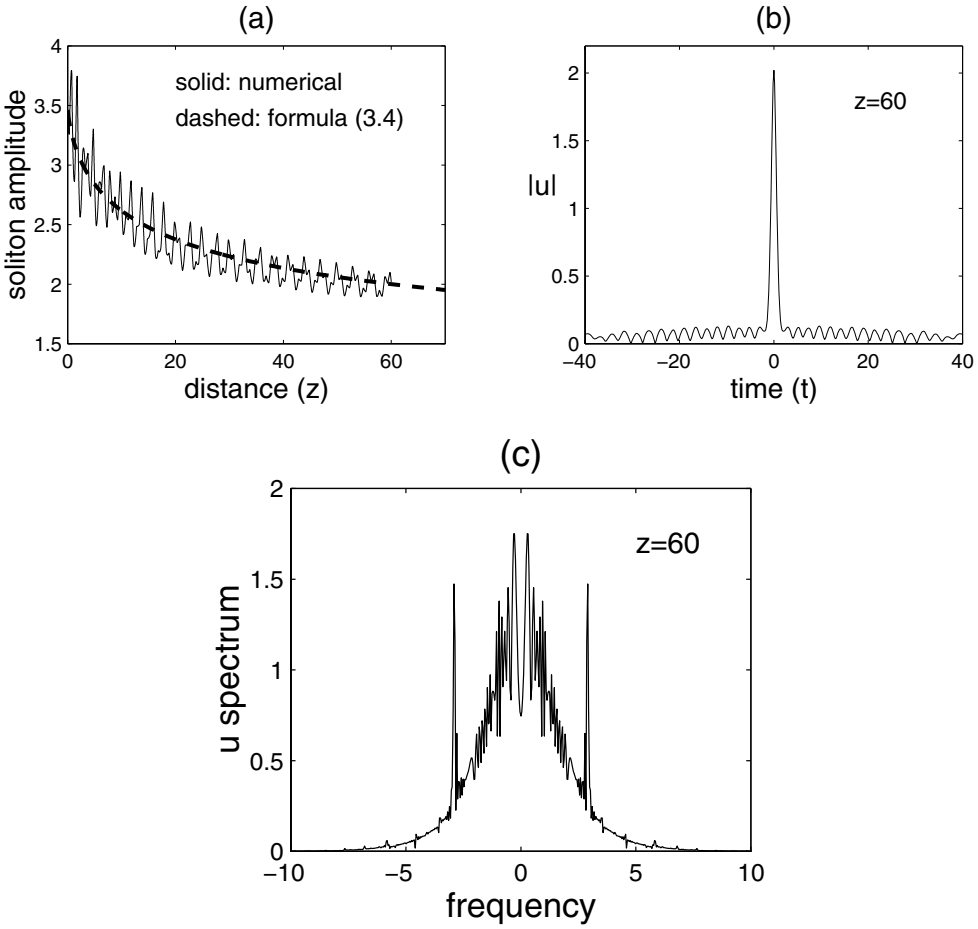


FIG. 3. *Numerical evolution of the DM soliton with* $m = 0.1$ *and* $\mu(0) = 6$. (a) *Soliton amplitude versus* z*: numerical results (solid curve); analytical average soliton amplitude* $\sqrt{2\mu}$ *from* (3.4) *(dashed curve).* (b) *Solution profile at* $z = 60$. (c) *Spectrum of the solution at* $z = 60$.

(3.4). In Figures 3(b) and (c), the numerically obtained field profile and its Fourier spectrum at distance $z = 60$ are plotted. Due to the quasi-critical resonance, the radiation field in Figure 3(b) is much larger than that in Figure 2(c). In addition, the Fourier spectrum in Figure 3(c) indicates that the solution can no longer be called a DM soliton. Nevertheless, the main resonant spikes on the two sides of Figure 3(c) are still well predicted by the resonance conditions at $k = \pm\sqrt{2\mu}\, k_1$.

3. *Figure* 4*:* $\mu(0) = 12$. In this case, the initial value of μ is above the lowest critical resonance value $\mu_1 = 2\pi$, and the monotonic decay of the DM soliton passes through this critical resonance. Here we focus on how this transition occurs. In Figure 4(a), the soliton amplitude versus distance z is plotted as the solid curve. We see that radiation damping is initially slow, as the average soliton amplitude decreases toward the critical value at $\sqrt{2\mu_1} \approx 3.54$. In this process, the DM soliton oscillates with the unit period of the dispersion map. A solution profile plotted in Figure 4(b) at $z = 50$ shows a weak radiation field, which is the reason for the slow decay of the DM soliton. The corresponding Fourier spectrum in Figure 4(d) shows that the radiation field consists of a discrete set of frequencies which are precisely the resonant frequencies.

When the average soliton amplitude passes through $\sqrt{2\mu_1}$, a critical resonance occurs. Consequently, the soliton decays much faster (see Figure 4(a)). Strong continuous-wave radiation is emitted in this process, and the DM soliton is strongly modified. After the average soliton amplitude passes below $\sqrt{2\mu_1}$, the pulse oscillates irregularly, and its oscillation period is no longer equal to the unit period of the dispersion map. A solution profile shown in Figure 4(c) at $z = 100$ confirms that the radiation field becomes much stronger past the critical-resonance stage. The Fourier spectrum in Figure 4(e) shows that the radiation field is no longer dominated by a discrete set of resonant frequencies. In addition, the Fourier spectrum appears to be quite noisy.

When a critical resonance is reached, the perturbation-series solution (2.6) and (2.7) formally breaks down, and the analytical results are not expected to provide quantitatively accurate approximations to the numerical solution. This is indeed the case. In Figure 4(a), the analytical average soliton amplitude $\sqrt{2\mu}$ obtained from (3.4) is also plotted (dashed line). We see that prior to the critical resonance, the analytical curve closely follows the numerical average soliton amplitude (not shown). However, when the numerical solution gets close to the critical resonance, it starts to deviate from the analytical curve considerably. In fact, the numerical solution passes through the critical resonance much earlier than what the theory predicts (see Figure 4(a)). Nevertheless, the analytical solution still agrees qualitatively with the numerical solution. For instance, the sharp (infinite-slope) drop of the soliton amplitude as predicted in (3.14) does occur past the critical value of the soliton amplitude at $\sqrt{2\mu_1} \approx 3.54$ (see Figure 4(a)).

4. *Figure* 5*:* $\mu(0) = 100$. When the initial value of μ is large, the asymptotic analysis predicts that the soliton decays exponentially according to the bounds in (3.18). However, the monotonic soliton-decay passes through many critical resonances in this case. Thus, the accuracy of the analytical prediction needs to be examined. To address this issue, the results from numerical simulations at $m = 0.1$ and $\mu(0) = 100$ are shown in Figures 5(a)–(e). When $\mu \gg 1$, the DM soliton spends most of the time inside individual constant-dispersion segments, where the DM soliton is governed by the standard NLS equation. This is reflected by the fast amplitude oscillations inside each constant-dispersion segment in Figure 5(a). Due to the radiative damping, the DM soliton passes through critical resonances at $n = 15, 13, 11, 9, 7, \ldots, 1$, when the average soliton amplitude matches the critical values at $\sqrt{4\pi n} = 13.73, 12.78, 11.76, 10.63, 9.38, \ldots, 3.54$, respectively. It follows from Figure 5(a) that, even though the DM soliton passes through a number of critical resonances here, it still holds up and maintains its DM soliton character and the unit periodicity up to the first four critical resonances. The solution profile and the Fourier spectrum of the DM soliton at $z = 15$ are shown in Figures 5(b) and (d). Further evolution of the DM soliton shows that the DM soliton character is lost after the fifth critical resonance at the average amplitude about 9.38. The solution profile and its Fourier spectrum at $z = 24$ are shown in Figures 5(c) and (e). A noisy spectrum past the critical resonances similar to that of Figure 4(e) is observed.

It follows from Figure 5(a) that higher-order critical resonances have a much weaker effect on the dynamics of the DM soliton than do lower-order resonances. As a result, the analytical dashed curve in Figure 5(a) for the average soliton amplitude agrees well with the numerical results until the fifth critical resonance is reached. We have also checked that the analytical curve in Figure 5(a) is indeed bounded between the two exponentially decaying functions in (3.18).

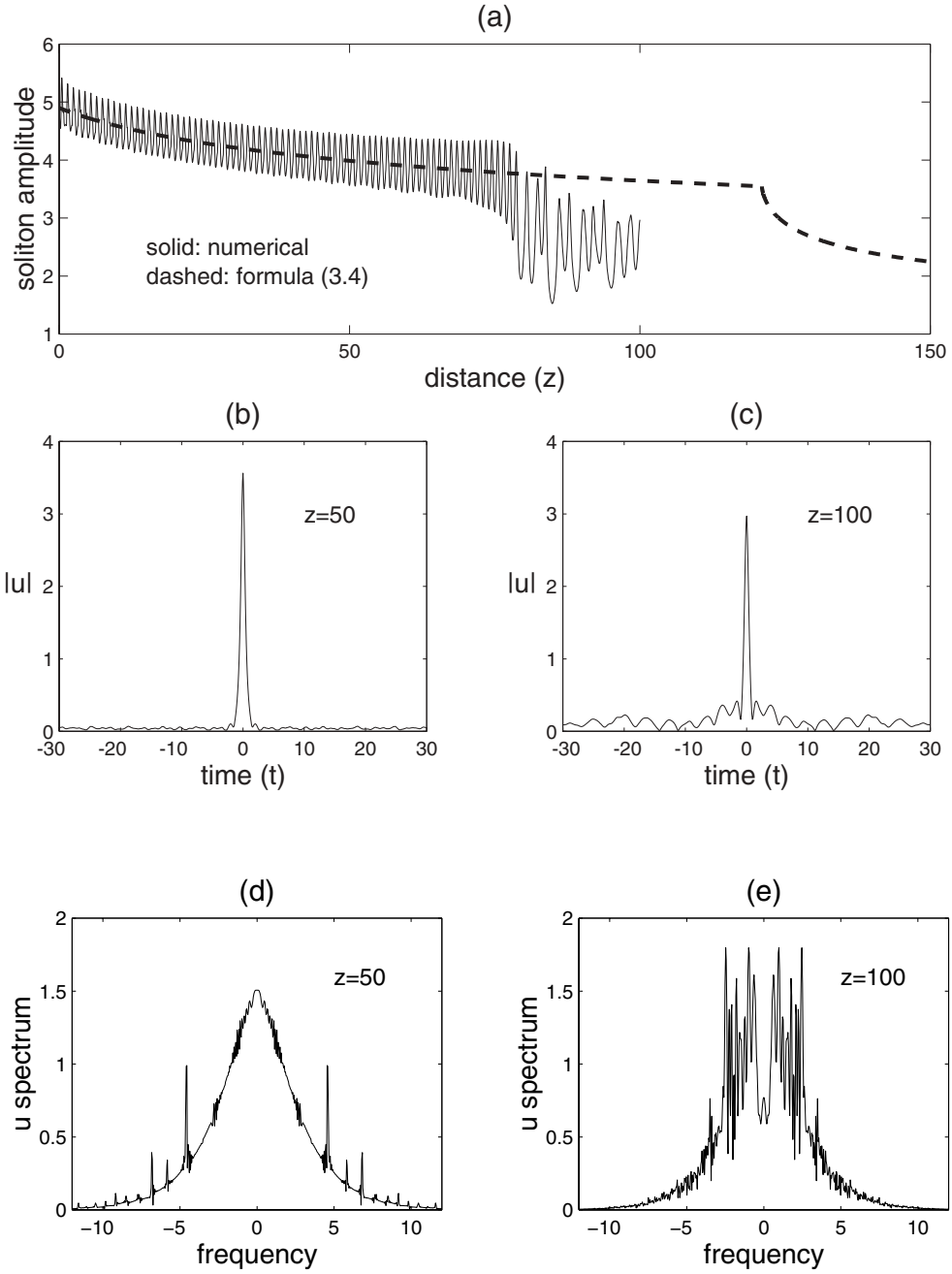


FIG. 4. *Numerical evolution of the DM soliton with* $m = 0.1$ *and* $\mu(0) = 12$. (a) *Soliton amplitude versus* z*: numerical results (solid curve); analytical average soliton amplitude* $\sqrt{2\mu}$ *from* (3.4) *(dashed curve).* (b), (c) *Solution profiles at* $z = 50$ *and* $z = 100$. (d), (e) *Spectra of the solutions at* $z = 50$ *and* $z = 100$.

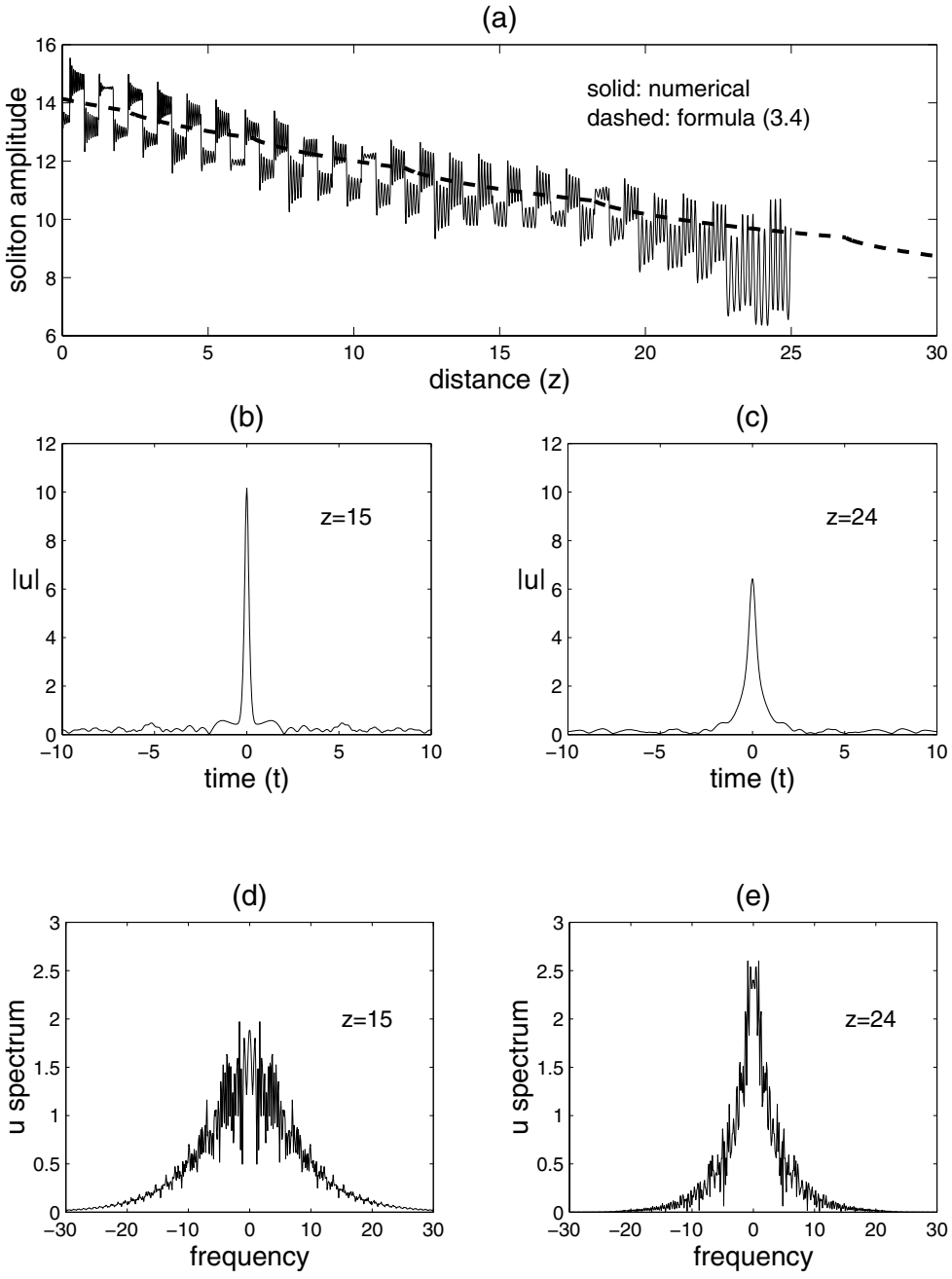


FIG. 5. *Numerical evolution of the DM soliton with* $m = 0.1$ *and* $\mu(0) = 100$. (a) *Soliton amplitude versus* z*: numerical results (solid curve); analytical average soliton amplitude* $\sqrt{2\mu}$ *from* (3.4) *(dashed curve).* (b), (c) *Solution profiles at* $z = 15$ *and* $z = 24$. (d), (e) *Spectra of the solutions at* $z = 15$ *and* $z = 24$.

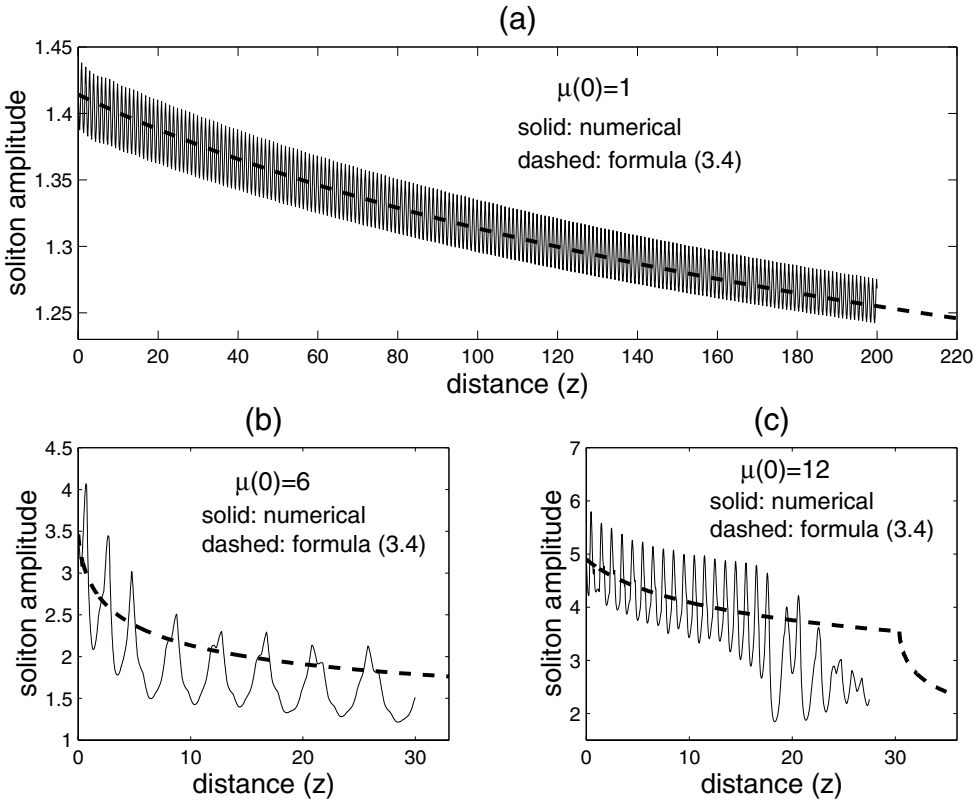


FIG. 6. *Numerical evolutions of the DM soliton amplitudes with* $m = 0.2$ *and* (a) $\mu(0) = 1$*;* (b) $\mu(0) = 6$*;* (c) $\mu(0) = 12$*. Numerical results are shown by solid curves. Analytical average soliton amplitudes* $\sqrt{2\mu}$ *from* (3.4) *are shown by dashed curves.*

In the end of this section, we discuss how the solution changes when the perturbation strength m gets larger. For this purpose, we choose $m = 0.2$, compared to $m = 0.1$ in Figures 2–5. The soliton amplitudes versus distance z for $\mu(0) = 1, 6, 12$ are shown in Figures 6(a), (b), and (c), respectively. The average soliton amplitudes predicted from (3.4) are also plotted for comparison. Figure 6(a) shows that in the case of small and moderate values of $\mu(0)$, the soliton still decays according to the analytical equation (3.4). Figures 6(b) and (c) indicate that when $\mu(0)$ is close to or above the lowest critical resonance value $\mu_1 = 2\pi$, the pulse deviates further from the DM soliton than in the case of $m = 0.1$, and the pulse amplitude oscillates with a period further away from the unit period of the dispersion map. When m increases, the distance scale for soliton evolution shrinks by a factor of m^2, as formula (3.4) predicts. For instance, when $m = 0.1$ and $\mu(0) = 12$, the lowest critical resonance is reached in the numerical solution at $z \approx 76$ (see Figure 4(a)), while when $m = 0.2$, the critical resonance in the numerical solution is reached at $z \approx 18$, i.e., four times faster.

5. Summary and discussion. In this paper, we have studied the nonlinear parametric resonance of DM solitons for average-anomalous dispersion ($D_0 > 0$) in the limit $m \to 0$ by both analytical and numerical methods. We have found that due to a resonance between the DM soliton and the dispersion map, the soliton

keeps shedding continuous-wave radiation and consequently decays. The radiation amplitude is on the order of m, while the decay rate of DM solitons is on the order of m^2. We have calculated the analytical approximations for the decay rate of DM solitons in the limits of small, intermediate, and large initial soliton amplitudes. We have shown that when the soliton passes through a critical resonance, it decays much faster. All these analytical results are found to be in excellent agreement with direct numerical simulations.

Resonances in the dispersion-periodic NLS equation (1.1) resemble a nonlinear generalization of parametric resonances in a linear Schrödinger equation studied recently in [17, 18]. The perturbation term in [17, 18] satisfies the assumption of being periodic in time and decaying fast in space. The nonlinear problem (1.1) does not satisfy this localization assumption. In addition, the periodic variations of $D_\epsilon(z)$ are not generally small perturbations of the mean term D_0 in real communication systems. Thus, rigorous analysis of the parametric resonance of DM solitons in dispersion-periodic NLS equation (1.1) with nonsmall dispersion variations needs further investigation.

Appendix A: Solutions of the first-order problem (2.14)–(2.15). We use Kaup's method [19] to solve the inhomogeneous problem (2.14)–(2.15) with the spectral decomposition for a linearized NLS operator. Since the potential of the problem can be rescaled as $\Phi(t;\mu) = \sqrt{2\mu}\,\Phi(T)$, where $\Phi(T) = \mathrm{sech} T$ and $T = \sqrt{2\mu}t$, we transform the variables as follows:

$$U_n^{(1)}(z,t;\mu) = 2d_n\sqrt{2\mu}\,V_n(Z,T), \quad Z = \mu z, \quad T = \sqrt{2\mu}t. \tag{A.1}$$

The system (2.14)–(2.15) in new variables transforms to the following:

$$i\frac{\partial V_n}{\partial Z} - (1+\lambda_n)\,V_n + \frac{\partial^2 V_n}{\partial T^2} + 2\,\mathrm{sech}^2 T\,\left(2V_n + \bar{V}_{-n}\right) = -\frac{1}{2}\Phi''(T), \tag{A.2}$$

$$-i\frac{\partial \bar{V}_{-n}}{\partial Z} - (1-\lambda_n)\,\bar{V}_{-n} + \frac{\partial^2 \bar{V}_{-n}}{\partial T^2} + 2\,\mathrm{sech}^2 T\,\left(2\bar{V}_{-n} + V_n\right) = -\frac{1}{2}\Phi''(T), \tag{A.3}$$

where

$$\lambda_n = \frac{2\pi n}{\mu}.$$

The system is written in matrix notations as

$$\mathcal{L}\begin{bmatrix} V_n \\ \bar{V}_{-n} \end{bmatrix} = \left(i\frac{\partial}{\partial Z} - \lambda_n\right)\begin{bmatrix} V_n \\ \bar{V}_{-n} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix}\Phi''(T), \tag{A.4}$$

where the linearized NLS operator is

$$\mathcal{L} = \begin{bmatrix} -\frac{\partial^2}{\partial T^2} + 1 - 4\,\mathrm{sech}^2 T & -2\,\mathrm{sech}^2 T \\ 2\,\mathrm{sech}^2 T & \frac{\partial^2}{\partial T^2} - 1 + 4\,\mathrm{sech}^2 T \end{bmatrix}. \tag{A.5}$$

The linearized NLS operator $\mathcal{L}$ possesses a complete set of eigenfunctions [19] that consists of eigenfunctions associated with two branches of the continuous spectrum and eigenfunctions associated with the zero eigenvalue of the discrete spectrum. The continuous spectrum eigenfunctions are

$$\boldsymbol{\psi}_1(T;k) = e^{ikT}\left[\left(1 - \frac{2ike^{-T}}{(k+i)^2\cosh T}\right)\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \frac{1}{(k+i)^2\cosh^2 T}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right] \tag{A.6}$$

and

$$\boldsymbol{\psi}_2(T;k) = e^{-ikT}\left[\left(1+\frac{2ike^{-T}}{(k-i)^2\cosh T}\right)\begin{pmatrix}1\\0\end{pmatrix}+\frac{1}{(k-i)^2\cosh^2 T}\begin{pmatrix}1\\1\end{pmatrix}\right], \tag{A.7}$$

such that $\mathcal{L}\boldsymbol{\psi}_1(T;k) = -(1+k^2)\boldsymbol{\psi}_1(T;k)$ and $\mathcal{L}\boldsymbol{\psi}_2(T;k) = (1+k^2)\boldsymbol{\psi}_2(T;k)$. The zero eigenvalue has algebraic multiplicity four and geometric multiplicity two. The eigenfunctions of the zero eigenvalue are

$$\boldsymbol{\phi}_1(T) = \begin{pmatrix}1\\-1\end{pmatrix}\operatorname{sech}T, \quad \boldsymbol{\phi}_2(T) = \begin{pmatrix}1\\1\end{pmatrix}\operatorname{sech}T\tanh T, \tag{A.8}$$

such that $\mathcal{L}\boldsymbol{\phi}_{1,2}(T) = 0$. The generalized eigenfunctions of the zero eigenvalue are

$$\boldsymbol{\phi}_1^d(T) = \begin{pmatrix}1\\1\end{pmatrix}(T\tanh T - 1)\operatorname{sech}T, \quad \boldsymbol{\phi}_2^d(T) = \begin{pmatrix}1\\-1\end{pmatrix}T\operatorname{sech}T, \tag{A.9}$$

such that $\mathcal{L}\boldsymbol{\phi}_{1,2}^d(T) = 2\boldsymbol{\phi}_{1,2}(T)$. Eigenfunctions of the linearized NLS operator $\mathcal{L}$ satisfy the orthogonality conditions

$$\langle\boldsymbol{\psi}_1(k')|\sigma_3|\boldsymbol{\psi}_1(k)\rangle = -2\pi\delta(k'-k), \quad \langle\boldsymbol{\psi}_2(k')|\sigma_3|\boldsymbol{\psi}_2(k)\rangle = 2\pi\delta(k'-k), \tag{A.10}$$

$$\langle\boldsymbol{\phi}_1|\sigma_3|\boldsymbol{\phi}_1^d\rangle = -2, \quad \langle\boldsymbol{\phi}_2|\sigma_3|\boldsymbol{\phi}_2^d\rangle = 2, \tag{A.11}$$

with respect to the inner product

$$\langle\mathbf{f}|\sigma_3|\mathbf{g}\rangle = \int_{-\infty}^{\infty}\left[\bar{f}_1(T)g_1(T) - \bar{f}_2(T)g_2(T)\right]dT. \tag{A.12}$$

All other inner products computed with eigenfunctions (A.6)–(A.9) are identically zero. The orthogonality conditions (A.10)–(A.11) are modified compared with the original definition in [19]. Orthogonality conditions similar to (A.10)–(A.11) were used by Kaup and Lakoba [20].

The right-hand side term of (A.4) can be decomposed through a complete set of eigenfunctions (A.6)–(A.9) as follows:

$$\begin{aligned}\mathbf{F} = \frac{1}{2}\begin{bmatrix}1\\-1\end{bmatrix}\Phi''(T) = &\int_{-\infty}^{\infty}\left[\alpha(k)\boldsymbol{\psi}_1(T;k) + \beta(k)\boldsymbol{\psi}_2(T;k)\right]dk\\ &+a\boldsymbol{\phi}_1(T) + b\boldsymbol{\phi}_2(T) + c\boldsymbol{\phi}_1^d(T) + d\boldsymbol{\phi}_2^d(T),\end{aligned} \tag{A.13}$$

where the expansion coefficients can be explicitly computed as

$$\alpha(k) = -\frac{1}{2\pi}\langle\boldsymbol{\psi}_1(k)|\sigma_3|\mathbf{F}\rangle = \frac{(k+i)^2}{8}\operatorname{sech}\frac{\pi k}{2}, \tag{A.14}$$

$$\beta(k) = \frac{1}{2\pi}\langle\boldsymbol{\psi}_2(k)|\sigma_3|\mathbf{F}\rangle = -\frac{(k-i)^2}{8}\operatorname{sech}\frac{\pi k}{2}, \tag{A.15}$$

$$a = -\frac{1}{2}\langle\boldsymbol{\phi}_1^d|\sigma_3|\mathbf{F}\rangle = -\frac{1}{2}, \quad b = \frac{1}{2}\langle\boldsymbol{\phi}_2^d|\sigma_3|\mathbf{F}\rangle = 0, \tag{A.16}$$

$$c = -\frac{1}{2}\langle \boldsymbol{\phi}_1 | \sigma_3 | \mathbf{F} \rangle = 0, \quad d = \frac{1}{2}\langle \boldsymbol{\phi}_2 | \sigma_3 | \mathbf{F} \rangle = 0. \tag{A.17}$$

Here we have used the exact value,

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\cos kT}{\cosh T} dT = \operatorname{sech}\left(\frac{\pi k}{2}\right).$$

The solution of (A.2)–(A.3) can be found by using the spectral decomposition:

$$\begin{aligned}\begin{bmatrix} V_n \\ \bar{V}_{-n} \end{bmatrix}(Z,T) &= \int_{-\infty}^{\infty} \left[\alpha_n(k,Z)\boldsymbol{\psi}_1(T;k) + \beta_n(k,Z)\boldsymbol{\psi}_2(T;k)\right] dk \\ &+ a_n(Z)\boldsymbol{\phi}_1(T) + b_n(Z)\boldsymbol{\phi}_2(T) + c_n(Z)\boldsymbol{\phi}_1^d(T) + d_n(Z)\boldsymbol{\phi}_2^d(T).\end{aligned} \tag{A.18}$$

Coefficients of the expansion satisfy a simple Z-evolution problem with zero initial conditions:

$$i\frac{\partial \alpha_n}{\partial Z} = (\lambda_n - 1 - k^2)\alpha_n - \alpha(k), \quad i\frac{\partial \beta_n}{\partial Z} = (\lambda_n + 1 + k^2)\beta_n - \beta(k), \tag{A.19}$$

$$i\frac{\partial a_n}{\partial Z} = \lambda_n a_n + 2c_n - a, \quad i\frac{\partial b_n}{\partial Z} = \lambda_n b_n + 2d_n - b, \tag{A.20}$$

and

$$i\frac{\partial c_n}{\partial Z} = \lambda_n c_n - c, \quad i\frac{\partial d_n}{\partial Z} = \lambda_n d_n - d. \tag{A.21}$$

The unique solution of the Z-evolution problem (A.19)–(A.21) is

$$\alpha_n(k,Z) = \frac{\alpha(k)}{\lambda_n - 1 - k^2}\left[1 - e^{-i(\lambda_n - 1 - k^2)Z}\right], \tag{A.22}$$

$$\beta_n(k,Z) = \frac{\beta(k)}{\lambda_n + 1 + k^2}\left[1 - e^{-i(\lambda_n + 1 + k^2)Z}\right], \tag{A.23}$$

$$a_n(Z) = -\frac{1}{2\lambda_n}\left[1 - e^{-i\lambda_n Z}\right], \quad b_n(Z) = 0, \tag{A.24}$$

and

$$c_n(Z) = 0, \quad d_n(Z) = 0. \tag{A.25}$$

Equations (A.24)–(A.25) are obtained with the use of (A.16)–(A.17).

Appendix B: Asymptotic limits for the first-order solution. We analyze the first-order solution $(V_n, \bar{V}_{-n})(Z,T)$ defined in the spectral representation form (A.18) of Appendix A with explicit spectral coefficients in (A.14)–(A.15) and (A.22)–(A.25). The asymptotic limit $Z \to \infty$ depends on a range of values of T.

(i) $|T| < \infty$ and $Z \to \infty$. The first-order solution is a sum of two terms, $V_n(Z,T) = W_n(T) + Q_n(Z,T)$, where $W_n(T)$ is generated by the inhomogeneous part of the system (A.2)–(A.3) and $Q_n(Z,T)$ is generated by the homogeneous part of the

system (A.2)–(A.3) in the initial-value problem. Using the spectral decomposition (A.18), we express $W_n(T)$ and $Q_n(Z,T)$ explicitly as

$$\left[\begin{array}{c} W_n \\ \bar{W}_{-n} \end{array} \right](T) = \int_{-\infty}^{\infty} \left[\frac{\alpha(k)}{\lambda_n - 1 - k^2} \boldsymbol{\psi}_1(T;k) + \frac{\beta(k)}{\lambda_n + 1 + k^2} \boldsymbol{\psi}_2(T;k) \right] dk - \frac{1}{2\lambda_n} \boldsymbol{\phi}_1(T) \tag{B.1}$$

and

$$\begin{aligned} \left[\begin{array}{c} Q_n \\ \bar{Q}_{-n} \end{array} \right](Z,T) = - \int_{-\infty}^{\infty} \left[\frac{\alpha(k) e^{-i(\lambda_n - 1 - k^2)Z}}{\lambda_n - 1 - k^2} \boldsymbol{\psi}_1(T;k) \right. \\ \left. + \frac{\beta(k) e^{-i(\lambda_n + 1 + k^2)Z}}{\lambda_n + 1 + k^2} \boldsymbol{\psi}_2(T;k) \right] dk + \frac{e^{-i\lambda_n Z}}{2\lambda_n} \boldsymbol{\phi}_1(T). \end{aligned} \tag{B.2}$$

We use formulas of generalized functions,

$$\lim_{Z \to \infty} \frac{e^{\pm iKZ}}{K} = \pm \pi i \delta(K) \tag{B.3}$$

and

$$\delta(k^2 + k_n^2) = 0, \quad \delta(k^2 - k_n^2) = \frac{1}{2k_n} \left[\delta(k - k_n) + \delta(k + k_n) \right], \tag{B.4}$$

and notice that the limit $Z \to \infty$ in (B.2) is nonzero only if the resonance equation $1 + k^2 \pm \lambda_n = 0$ has a solution for real k. We consider $n > 0$ such that $\lambda_n > 0$ and denote $k_n = \sqrt{\lambda_n - 1} \geq 0$ for $\lambda_n \geq 1$. The resonance condition $\lambda_n \geq 1$ is satisfied for $n \geq N_\mu$, where $N_\mu = [\frac{\mu}{2\pi}]$ is the integer ceiling of $\frac{\mu}{2\pi} > 0$. With the use of (B.3)–(B.4), we compute the limit $Z \to \infty$ for $Q_n(Z,T)$ at $n \geq N_\mu$ and finite T:

$$\lim_{Z \to \infty} \left[\begin{array}{c} Q_n \\ \bar{Q}_{-n} \end{array} \right](Z,T) = \frac{\pi i}{2k_n} \left[\alpha(k_n) \boldsymbol{\psi}_1(T;k_n) + \alpha(-k_n) \boldsymbol{\psi}_1(T;-k_n) \right]. \tag{B.5}$$

The first-order solution $V_n(Z,T) = W_n(T) + Q_n(Z,T)$ is bounded in T and Z in the limit $Z \to \infty$.

(ii) $|T| \to \infty$ and $Z \to \infty$. It follows from (B.3)–(B.4) that

$$\lim_{T \to \pm\infty} \frac{e^{ikT}}{(k - k_n)(k + k_n)} = \pm \frac{\pi i}{2k_n} \left[\delta(k - k_n) e^{ik_n T} - \delta(k + k_n) e^{-ik_n T} \right]. \tag{B.6}$$

Using this formula for $n \geq N_\mu$, we find from (B.1) and (B.5) that

$$\lim_{T \to \pm\infty} \left[\begin{array}{c} W_n \\ \bar{W}_{-n} \end{array} \right](T) = \mp \frac{\pi i}{16 k_n} \operatorname{sech} \frac{\pi k_n}{2} \left[e^{ik_n T} (k_n \pm i)^2 - e^{-ik_n T} (k_n \mp i)^2 \right] \left(\begin{array}{c} 0 \\ 1 \end{array} \right) \tag{B.7}$$

and

$$\lim_{T \to \pm\infty, Z \to \infty} \left[\begin{array}{c} Q_n \\ \bar{Q}_{-n} \end{array} \right](Z,T) = \frac{\pi i}{16 k_n} \operatorname{sech} \frac{\pi k_n}{2} \left[e^{ik_n T} (k_n \pm i)^2 + e^{-ik_n T} (k_n \mp i)^2 \right] \left(\begin{array}{c} 0 \\ 1 \end{array} \right). \tag{B.8}$$

As a result, the boundary values of the first-order solution $V_n(Z,T)$ satisfy the Sommerfeld radiation boundary conditions:

$$\lim_{|T|\to\infty, Z\to\infty} V_{-n}(Z,T) = -\frac{\pi i(k_n+i)^2}{8k_n}\operatorname{sech}\frac{\pi k_n}{2}e^{ik_n|T|}, \quad n \geq N_\mu. \tag{B.9}$$

(iii) $|T| \to \infty$ and $Z < \infty$. Using formula (B.6) in (B.1)–(B.2), we find that both terms cancel out since

$$\lim_{k\to\pm k_n}\left(1 - e^{-i(\lambda_n - 1 - k^2)Z}\right) = 0.$$

As a result, we have zero boundary values for $V_n(Z,T)$ in the limit $|T| \to \infty$ for finite Z:

$$\lim_{|T|\to\infty} V_n(Z,T) = 0. \tag{B.10}$$

The first-order solution represents radiative waves diverging from the NLS soliton. In the limit $Z \to \infty$, the radiative waves approach the Z-independent boundary values given by (B.9). In the intermediate region, where $|T| \to \infty$, $Z \to \infty$, and $\lim_{Z\to\infty} T/Z = C$, where $0 < C < \infty$, the radiative waves move with the group velocity $2k_n$, according to the intermediate asymptotic expression

$$\lim_{|T|\to\infty, Z\to\infty} V_{-n}(Z,T) = -\frac{\pi i(k_n+i)^2}{8k_n}\operatorname{sech}\frac{\pi k_n}{2}\; e^{ik_n|T|}\; H\left(2k_n - \frac{|T|}{Z}\right), \quad n \geq N_\mu, \tag{B.11}$$

where $H(z) = 1$ for $z > 0$ and $H(z) = 0$ for $z < 0$. The intermediate asymptotic expression includes (B.9) and (B.10) as particular cases.

Acknowledgments. The authors thank W. Kath, E. Kirr, B. Malomed, M. Weinstein, and V. Zharnitsky for useful discussions.

REFERENCES

[1] V. Cautaerts, A. Maruta, and Y. Kodama, *On the dispersion-managed soliton*, Chaos, 10 (2000), p. 515.

[2] S. Turitsyn, M. P. Fedoruk, E. G. Shapiro, V. K. Mezentsev, and E. G. Turitsyna, *Novel approaches to numerical modeling of periodic dispersion-managed fiber communication systems*, IEEE J. Quantum Electr., 6 (2000), pp. 263–275.

[3] J. H. Nijhof, W. Forysiak, and N. J. Doran, *The averaging method for finding exactly periodic dispersion-managed solitons*, IEEE J. Quantum Electr., 6 (2000), p. 330.

[4] D. E. Pelinovsky and V. Zharnitsky, *Averaging of dispersion-managed solitons: Existence and stability*, SIAM J. Appl. Math., 63 (2003), pp. 745–776.

[5] I. R. Gabitov and S. K. Turitsyn, *Averaged pulse dynamics in a cascaded transmission system with passive dispersion compensation*, Opt. Lett., 21 (1996), pp. 327–329.

[6] M. J. Ablowitz and G. Biondini, *Multiscale pulse dynamics in communication systems with strong dispersion management*, Opt. Lett., 23 (1998), pp. 1668–1670.

[7] V. Zharnitsky, E. Grenier, S. Turitsyn, C. K. R. T. Jones, and J. S. Hesthaven, *Ground states of dispersion-managed nonlinear Schrödinger equation*, Phys. Rev. E (3), 62 (2000), p. 7358.

[8] M. Kunze, *The singular perturbation limit of a variational problem from nonlinear fiber optics*, Phys. D, 180 (2003), pp. 108–114.

[9] P. M. Lushnikov, *Dispersion-managed soliton in a strong dispersion map limit*, Opt. Lett., 26 (2001), pp. 1535–1537.

[10] T. Yang and W. L. Kath, *Radiation loss of dispersion-managed solitons in optical fibers*, Phys. D, 149 (2001), p. 80.

[11] J. H. B. Nijhof, N. J. Doran, W. Forysiak, and F. M. Knox, *Stable soliton-like propagation in dispersion-managed systems with net anomalous, zero and normal dispersion*, Electron. Lett., 33 (1997), pp. 1726–1727.

[12] A. V. Mikhailov and V. Yu. Novokshenov, *The Riemann–Hilbert Problem for Analytic Description of the DM Solitons*, preprint, University of Leeds, Leeds, UK, 2003.

[13] F. K. Abdullaev, J. G. Caputo, and N. Flytzanis, *Envelope soliton propagation in media with temporally modulated dispersion*, Phys. Rev. E (3), 50 (1994), pp. 1552–1558.

[14] B. Malomed, D. F. Parker, and N. F. Smyth, *Resonant shape oscillations and decay of a soliton in a periodically inhomogeneous nonlinear optical fiber*, Phys. Rev. E (3), 48 (1993), p. 1418.

[15] R. Grimshaw, J. He, and B. A. Malomed, *Decay of a fundamental soliton in a periodically modulated nonlinear waveguide*, Phys. Scripta 53 (1996), pp. 385–393.

[16] F. K. Abdullaev and J. G. Caputo, *Validation of the variational approach for chirped pulses in fibers with periodic dispersion*, Phys. Rev. E (3), 58 (1998), p. 6637.

[17] P. D. Miller, A. Soffer, and M. I. Weinstein, *Metastability of breather modes of time-dependent potentials*, Nonlinearity, 13 (2000), pp. 507–568.

[18] E. Kirr and M. I. Weinstein, *Parametrically excited Hamiltonian partial differential equations*, SIAM J. Math. Anal., 33 (2001), pp. 16–52.

[19] D. J. Kaup, *Perturbation theory for solitons in optical fibers*, Phys. Rev. A (3), 42 (1990), pp. 5689–5694.

[20] D. J. Kaup and T. I. Lakoba, *Variational method: How it can generate false instabilities*, J. Math. Phys., 37 (1996), pp. 3442–3462.

[21] T. I. Lakoba, J. Yang, D. J. Kaup, and B. A. Malomed, *Conditions for stationary pulse propagation in the strong dispersion management regime*, Opt. Comm., 149 (1998), pp. 366–375.

[22] J. Yang, *Dynamics of embedded solitons in the extended KdV equations*, Stud. Appl. Math., 106 (2001), p. 337.

[23] D. Pelinovsky, *Radiative effects to the adiabatic dynamics of envelope-wave solitons*, Phys. D, 119 (1998), pp. 301–313.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1383–1407

COLLAPSE OF THE KELDYSH CHAINS AND STABILITY OF CONTINUOUS NONCONSERVATIVE SYSTEMS*

OLEG N. KIRILLOV† AND ALEXANDER P. SEYRANIAN†

Abstract. In the present paper, eigenvalue problems for non-self-adjoint linear differential operators smoothly dependent on a vector of real parameters are considered. Bifurcation of eigenvalues along smooth curves in the parameter space is studied. The case of a multiple eigenvalue with the Keldysh chain of arbitrary length is investigated. Explicit expressions describing bifurcation of eigenvalues are found. The obtained formulas use eigenfunctions and associated functions of the adjoint eigenvalue problems as well as the derivatives of the differential operator taken at the initial point of the parameter space. These results are important for the stability problems and sensitivity analysis of nonconservative systems. As a mechanical application, the extended Beck problem of stability of an elastic column subjected to a partially tangential follower force is considered and discussed in detail.

Key words. nonconservative system, non-self-adjoint differential operator, Keldysh chain, multiple eigenvalue, bifurcation, stability boundary

AMS subject classifications. 34B08, 34D10, 34L16

DOI. 10.1137/S0036139902414720

1. Introduction.

In nonconservative dynamic stability problems arising in mechanics and physics, energy is not conserved; it can be pumped into the system or taken out depending on problem parameters. Increase of energy leads to growth of vibrational amplitudes, i.e., to instability of vibrations. If energy is lost the amplitudes decay in time, which implies stability for the system. Unlike buckling problems of conservative systems for nonconservative ones, the dynamic method of stability study must be applied. Two types of instability in nonconservative systems are distinguished: static (divergence) and dynamic (flutter). Important examples of nonconservative systems are aircraft wings and panels vibrating in a flow, elastic missiles subjected to a jet thrust, which is a nonpotential follower force, and tubes conveying fluid; see Bolotin [1], Ziegler [2], Leipholz [3], and Paidoussis [4]. Sensitivity analysis of critical stability parameters for nonconservative systems was developed by Pedersen and Seyranian in [5]. A comprehensive review of nonconservative stability problems was given by Langthjem and Sugiyama in [6].

Non-self-adjoint operators naturally appear in nonconservative problems. In discrete problems such an operator is just a nonsymmetrical matrix. The general theory of non-self-adjoint operators going back to the works by Birkhoff was then developed by many mathematicians; see the review by Davies [7]. Keldysh [8] was the first to generalize the notion of the Jordan chain of vectors to a wide class of non-self-adjoint operators. For that reason it was called the Keldysh chain; see Gohberg and Krein [9] and Gohberg, Lancaster, and Rodman [10]. In the work by Vishik and Lyusternik [11], the perturbation theory for nonsymmetrical matrices and non-self-adjoint differential

*Received by the editors September 16, 2002; accepted for publication (in revised form) December 12, 2003; published electronically May 20, 2004. This work was supported in part by the Russian Foundation for Basic Research (RFBR-NSFC 02-01-39004, RFBR 03-01-00161) and by the United States Civilian Research and Development Foundation for the Independent States of the Former Soviet Union (CRDF-BRHE Y1-MP-06-19).

http://www.siam.org/journals/siap/64-4/41472.html

†Institute of Mechanics, Moscow State Lomonosov University, Michurinsky pr. 1, 119192 Moscow, Russia (kirillov@imec.msu.ru, seyran@imec.msu.ru).

operators, $L = L_0 + \epsilon L_1$ with ϵ as a small parameter, was developed. This practical and constructive theory allows one to find the perturbation coefficients of eigenvalues and eigenvectors in an explicit form. The paper by Vishik and Lyusternik [11] was not widely known in the Western (and even Russian) literature on the subject for a long time. However, its importance was highly appreciated in the paper by Moro, Burke, and Overton [12]. We extend this perturbation theory to multiparameter bifurcation analysis of eigenvalues of non-self-adjoint differential operators.

It is known that in the generic case the spectrum of a multiparameter family of nonsymmetrical matrices contains multiple eigenvalues with the Jordan chains; see Arnold [13]. In many important cases multiple eigenvalues define geometrical properties of the stability boundary of a corresponding system. At the same time, multiple eigenvalues create considerable mathematical difficulties due to their nondifferentiability with respect to parameters. An effective tool for analysis of stability boundary is the study of bifurcations of multiple eigenvalues due to change of parameters. For the discrete case, this method based on the perturbation theory [11] was developed in the works by Seyranian [14], Mailybaev and Seyranian [15], and Seyranian and Kirillov [16]. To perform the stability analysis in the continuous case, we need to consider bifurcations of eigenvalues in multiparameter families of non-self-adjoint differential operators. The study of generic properties of the spectrum of the multiparameter family of non-self-adjoint differential operators remains a difficult problem. It seems that in the infinite-dimensional case there is still no analogue to the Arnold theory of versal deformations of matrices, allowing us to classify the generic singularities of the bifurcation and stability diagrams: even in the self-adjoint case the progress is quite slow; see Teytel [17].

In our paper we combine the ideas of [8], [11], and [14]. This allows us to find explicit formulas describing bifurcation of multiple eigenvalues with the Keldysh chain of any length. These formulas suit for a wide class of non-self-adjoint eigenvalue problems arising in applications and take into account parameters both in the differential expression and in the boundary conditions. Besides, our approach allows one to study bifurcations of multiple eigenvalues both in regular and degenerate cases. An analogous approach was applied by Seyranian and Kliem to the investigation of stability problems for continuous conservative systems with gyroscopic forces [18].

The paper is organized in the following way. In section 2 basic relations for eigenvalue problems with general linear differential operators and boundary conditions are introduced.

In section 3 it is supposed that the differential expression and boundary conditions smoothly depend on a vector of real parameters. A formula describing splitting of a multiple eigenvalue with the Keldysh chain of arbitrary length depending on a change of the parameters is derived. Both regular and degenerate cases are studied. Finally, a formula for bifurcation of a semisimple multiple eigenvalue is obtained. These formulas generalize the results for splitting of eigenvalues obtained earlier for the finite-dimensional case in [14], [15], and [16]. The obtained formulas take into account both variations of the differential expression and boundary conditions due to change of the parameters and can be applied to a wide class of nonconservative problems.

In section 4 we apply the results of previous sections to stability problems of general nonconservative (circulatory) systems. General solutions in an explicit form via eigenvalues and eigenvectors are obtained. Then stability boundaries in the parameter space are investigated. It is shown that the smooth parts of the boundaries correspond to either simple zero (stability-divergence boundary) or double positive

eigenvalues with the Keldysh chain of length 2 (stability-flutter boundary). Normal vectors to the boundaries between stability, flutter, and divergence domains are found. It is remarkable that only information at a point of the stability boundary is needed for the calculation of the vectors.

Section 5 treats a mechanical example—an elastic column loaded by the partially tangential follower force. This problem is referred to as the extended Beck column. The stability boundaries of this two-parameter continuous system are carefully investigated and analyzed. Explicit expressions for eigenfunctions and associated functions are derived. With the use of these expressions, the linear and quadratic approximations of the stability boundary at its regular and singular points are constructed. The behavior of eigenvalues in the vicinity of the stability boundaries is studied by the perturbation approach, showing a good agreement with the numerical results.

2. Basic relations. Using the notation of Naimark [19] we consider an eigenvalue problem for a linear differential operator L defined by

$$l(u) = \lambda u, \quad U^s(u) = 0, \quad s = 1, \dots, m, \tag{2.1}$$

$$l(u) \equiv \sum_{i=0}^{m} a_i \frac{d^{m-i}u}{dx^{m-i}}, \quad U^s(u) \equiv \sum_{i=0}^{m-1} \left(\alpha_i^s \frac{d^i u}{dx^i}\bigg|_{x=0} + \beta_i^s \frac{d^i u}{dx^i}\bigg|_{x=1} \right).$$

The operators $U^s(u)$ are linear forms with respect to the variables $u(0)$, $u'(0), \dots,$ $u^{(m-1)}(0)$; $u(1)$, $u'(1), \dots,$ $u^{(m-1)}(1)$. These variables are values of the function $u \in C^{(m)}[0,1]$ and its derivatives up to $(m-1)$th order evaluated at the points $x = 0$ and $x = 1$. It is assumed that the forms U^s, $s = 1, 2, \dots, m$, are linearly independent.

The differential expression

$$l^*(v) \equiv \sum_{i=0}^{m} (-1)^{m-i} \, \overline{a_i} \, \frac{d^{m-i}v}{dx^{m-i}},$$

where the overbar denotes complex conjugation, is called *adjoint* to the differential expression $l(u)$ [19]. With the use of integration by parts it can be shown that

$$\int_0^1 l(u)\bar{v}dx = P(\alpha, \beta) + \int_0^1 u\overline{l^*(v)}dx, \tag{2.2}$$

where $P(\alpha, \beta)$ is a bilinear form of the variables

$$\alpha = (u(0), u'(0), \dots, u^{(m-1)}(0), u(1), u'(1), \dots, u^{(m-1)}(1)), \tag{2.3}$$

$$\beta = (v(0), v'(0), \dots, v^{(m-1)}(0), v(1), v'(1), \dots, v^{(m-1)}(1)). \tag{2.4}$$

Let us choose the forms $U^{m+1}, U^{m+2}, \dots, U^{2m}$ so that $U^1, U^2, \dots, U^{2m}$ are linearly independent. Then variables (2.3) can be expressed as linear combinations of the forms $U^1, U^2, \dots, U^{2m}$. Substituting these linear combinations into (2.2), we get the Lagrange identity [19]

$$(l(u), v) - (u, l^*(v)) = U^1 V^{2m} + \dots + U^{2m} V^1, \tag{2.5}$$

where $(u, v) = \int_0^1 u(x)\overline{v}(x)dx$ is the inner product of the functions $u, v \in C^m[0,1]$.

The coefficients at $U^1, U^2, \dots, U^{2m}$ are linear forms with respect to variables (2.4) and are denoted by $V^{2m}, \dots, V^2, V^1$, respectively. The forms $V^1, V^2, \dots, V^{2m}$ are linearly independent [19]. The boundary conditions

$$V^s(v) = 0, \quad s = 1, \dots, m,$$

are called *adjoint* to boundary conditions (2.1). The differential operator L^*, corresponding to the differential expression $l^*(v)$ and to the adjoint boundary conditions, is referred to as adjoint to the operator L, and we say that the eigenvalue problem

$$l^*(v) = \overline{\lambda} v, \quad V^s(v) = 0, \quad s = 1, \dots, m, \tag{2.6}$$

is adjoint to eigenvalue problem (2.1).

Due to the boundary conditions in (2.1) and (2.6), identity (2.5) for the adjoint operators L and L^* takes a simple form: $(l(u), v) = (u, l^*(v))$. If we consider differential expression $l(u)$ and assume that the function u satisfies the nonhomogeneous boundary conditions

$$U^s(u) = G^s, \quad s = 1, \dots, m, \tag{2.7}$$

then the Lagrange identity (2.5) yields

$$(l(u), v) - (u, l^*(v)) = G^1 V^{2m} + \cdots + G^m V^{m+1}. \tag{2.8}$$

This is valid since v satisfies the boundary conditions in (2.6).

3. Collapse of the Keldysh chain. Suppose that in eigenvalue problem (2.1) the coefficients of the differential expression $l(u)$ and the coefficients of the forms $U^s(u)$ are real functions, *smoothly* dependent on a vector of real parameters $\mathbf{p} = (p_1, p_2, \dots, p_n)$ on an open set $\Omega \subset R^n$. Let λ_0 be an eigenvalue of the operator L at the point $\mathbf{p} = \mathbf{p}_0$. We are interested in bifurcation of eigenvalues along the curves $\mathbf{p}(\epsilon) = \mathbf{p}_0 + \epsilon \mathbf{e} + \epsilon^2 \mathbf{d} + o(\epsilon^2)$, emitted from the initial point $\mathbf{p}_0$ in the parameter space. The vector $\mathbf{e} = (e_1, e_2, \dots, e_n)$ defines a direction of the curve, and $\epsilon \geq 0$ is a small parameter.

Due to variation of parameters the differential expression $l(u)$ and the forms $U^s(u)$ are expanded as

$$l(u) = l_0(u) + \epsilon l_1(u) + \epsilon^2 l_2(u) + \cdots, \quad U^s(u) = U_0^s(u) + \epsilon U_1^s(u) + \epsilon^2 U_2^s(u) + \cdots, \tag{3.1}$$

where $l_0 = l(u)|_{\mathbf{p}=\mathbf{p}_0}$, $U_0^s = U^s(u)|_{\mathbf{p}=\mathbf{p}_0}$, the differential expressions $l_1(u)$, $l_2(u)$ look like

$$l_1(u) = \sum_{i=1}^n e_i \frac{\partial l}{\partial p_i}(u), \quad l_2(u) = \sum_{i=1}^n d_i \frac{\partial l}{\partial p_i}(u) + \frac{1}{2} \sum_{i,j=1}^n e_i e_j \frac{\partial^2 l}{\partial p_i \partial p_j}(u), \tag{3.2}$$

and for the forms $U_1^s(u)$, $U_2^s(u)$ we have

$$U_1^s(u) = \sum_{i=1}^n e_i \frac{\partial U^s}{\partial p_i}(u), \quad U_2^s(u) = \sum_{i=1}^n d_i \frac{\partial U^s}{\partial p_i}(u) + \frac{1}{2} \sum_{i,j=1}^n e_i e_j \frac{\partial^2 U^s}{\partial p_i \partial p_j}(u). \tag{3.3}$$

All the derivatives in formulas (3.2) and (3.3) are evaluated at the point $\mathbf{p} = \mathbf{p}_0$. Thus, we deal with the *regular* perturbations which do not increase the order of the nonperturbed operator $L_0 = L(\mathbf{p}_0)$ [11].

Consider an eigenvalue λ_0 with *the Keldysh chain of length* $k > 0$. This means that at $\mathbf{p} = \mathbf{p}_0$ there exist an eigenfunction $u_0(x)$ and *associated* functions $u_1(x)$, $u_2(x), \dots, u_{k-1}(x)$, corresponding to the λ_0 and satisfying the equations and the boundary conditions

$$\begin{array}{cc} l_0(u_0) = \lambda_0 u_0, & U_0^s(u_0) = 0; \\ l_0(u_i) = \lambda_0 u_i + u_{i-1}, & U_0^s(u_i) = 0; \\ i = 1, \dots, k-1; \quad s = 1, \dots, m. & \end{array} \tag{3.4}$$

For the adjoint operator L^* we have

$$\begin{array}{cc} l_0^*(v_0) = \overline{\lambda_0} v_0, & V_0^s(v_0) = 0; \\ l_0^*(v_i) = \overline{\lambda_0} v_i + v_{i-1}, & V_0^s(v_i) = 0; \\ i = 1, \dots, k-1; \quad s = 1, \dots, m. & \end{array} \tag{3.5}$$

The notion of the Keldysh chain is an analogue of the Jordan chain of vectors when we consider eigenvalue problems for differential operators [8], [9], [19]. Eigenfunctions and associated functions of adjoint operators L and L^* are related by the following conditions:

$$(u_j, v_0) = 0, \quad j = 0, \dots, k-2, \quad (u_{k-1}, v_0) \equiv (u_0, v_{k-1}) \neq 0; \tag{3.6}$$

$$(u_{j-1}, v_i) \equiv (u_j, v_{i-1}), \quad i, j = 1, \dots, k-1. \tag{3.7}$$

This naturally follows from (3.4) and (3.5) with the relation $(l(u), v) = (u, l^*(v))$ stated for the adjoint operators.

A variation of the vector of parameters $\mathbf{p} = \mathbf{p}_0 + \epsilon \mathbf{e} + o(\epsilon)$ causes perturbation of eigenvalues and eigenfunctions. In the case of a multiple eigenvalue with the Keldysh chain of length k, the expansions for eigenvalues and eigenfunctions contain terms with fractional powers of the small parameter $\epsilon^{j/k}$, $j = 0, 1, 2, \dots$ [11]:

$$\lambda = \lambda_0 + \epsilon^{1/k} \lambda_1 + \epsilon^{2/k} \lambda_2 + \cdots, \quad u = u_0 + \epsilon^{1/k} w_1 + \epsilon^{2/k} w_2 + \cdots. \tag{3.8}$$

Substituting expansions (3.1) and (3.8) into eigenvalue problem (2.1), we get expressions that determine the first order perturbations of the eigenvalue λ_0 and eigenfunction u_0,

$$l_0(w_j) - \lambda_0 w_j = \lambda_j u_0 + \sum_{i=1}^{j-1>0} \lambda_{j-i} w_i, \quad U_0^s(w_j) = 0, \quad j = 1 \dots k-1, \tag{3.9}$$

$$l_0(w_k) - \lambda_0 w_k = \lambda_k u_0 - l_1(u_0) + \sum_{i=1}^{k-1>0} \lambda_{k-i} w_i, \quad U_0^s(w_k) = -U_1^s(u_0). \tag{3.10}$$

The functions w_j can be found from (3.4) and (3.9) in the form

$$w_j = \lambda_1^j u_j + \sum_{p=0}^{j-1} \gamma_{jp} u_p, \quad j = 1, \dots, k-1, \tag{3.11}$$

where γ_{jp} are arbitrary constants.

Consider the inner product of the function v_0 with the left- and right-hand sides of (3.10). Using then expression (3.11) for w_j, equations (3.6) and (3.7), and the Lagrange identity (2.8), which in this case has the form

$$(l_0(w_k) - \lambda_0 w_k, v_0) - (w_k, l_0^*(v_0) - \overline{\lambda_0} v_0) = -\sum_{s=1}^{m} U_1^s(u_0) V_0^{2m-s+1}(v_0),$$

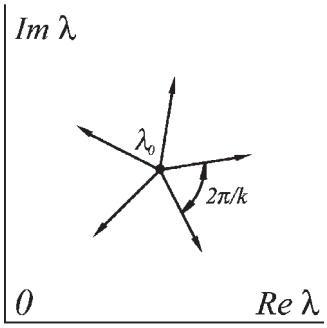


FIG. 3.1. *Splitting of the multiple eigenvalue λ_0 with the Keldysh chain of length $k = 5$.*

we get the coefficient λ_1 in the expansion of the eigenvalue λ,

$$\lambda_1^k = \frac{(l_1(u_0), v_0) - \sum_{s=1}^{m} U_1^s(u_0) V_0^{2m-s+1}(v_0)}{(u_{k-1}, v_0)}. \tag{3.12}$$

Introducing scalar product $\langle \mathbf{a}, \mathbf{b} \rangle$ of the vectors $\mathbf{a}, \mathbf{b} \in R^n$ and taking into account expressions (3.2) and (3.3) we can rewrite (3.12) in the form [14]

$$\lambda_1 = \sqrt[k]{\langle \mathbf{f}_k, \mathbf{e} \rangle + i \langle \mathbf{g}_k, \mathbf{e} \rangle}, \tag{3.13}$$

where the real vectors $\mathbf{f}_k$ and $\mathbf{g}_k$ correspond to the k-fold eigenvalue λ_0 at the point $\mathbf{p} = \mathbf{p}_0$ and their components are

$$f_k^j + i g_k^j = \frac{(\frac{\partial l}{\partial p_j}(u_0), v_0) - \sum_{s=1}^{m} \frac{\partial U^s}{\partial p_j}(u_0) V_0^{2m-s+1}(v_0)}{(u_{k-1}, v_0)}, \quad j = 1, \ldots, n. \tag{3.14}$$

The right-hand side of (3.13) takes k complex values. If the radicand in (3.13) is not zero, the expression $\lambda = \lambda_0 + \epsilon^{1/k} \lambda_1 + o(\epsilon^{1/k})$ describes the splitting of the k-fold eigenvalue with a change of parameters along a curve emitted in the direction $\mathbf{e}$, as shown in Figure 3.1.

We emphasize that for real eigenvalue λ_0 the vector $\mathbf{g}_k$=0 since the eigenfunctions and associated functions can be chosen real, and the coefficient λ_1 does not depend on normalization conditions.

After splitting, the length of the Keldysh chain decreases from k to 1, and we say that *collapse of the Keldysh chain* occurs.

In particular, (3.8) and (3.13) describe the behavior of a simple eigenvalue for $k = 1$, and for $k = 2$-splitting of a double eigenvalue λ_0 with the Keldysh chain of length 2, which are the most important cases in applications.

Consider now for a double eigenvalue the *degenerate* case

$$\langle \mathbf{f}_2, \mathbf{e}_* \rangle + i \langle \mathbf{g}_2, \mathbf{e}_* \rangle = 0. \tag{3.15}$$

It follows from condition (3.15) that the coefficient λ_1 in expansions (3.8) becomes zero. Substitution of expansions (3.8) into eigenvalue problem (2.1) gives equations determining second order terms λ_2 and w_2,

$$l_0(w_2) - \lambda_0 w_2 = \lambda_2 u_0 - l_1(u_0), \quad U_0^s(w_2) = -U_1^s(u_0), \tag{3.16}$$

$$l_0(w_4)-\lambda_0 w_4 = \lambda_4 u_0 - l_2(u_0)+\lambda_2 w_2 - l_1(w_2), \quad U_0^s(w_4) = -U_1^s(w_2)-U_2^s(u_0). \tag{3.17}$$

Multiplying both parts of (3.17) by v_0 and using the Lagrange identity (2.8), we get

$$\lambda_2(w_2,v_0)-(l_1(w_2),v_0)-(l_2(u_0),v_0)+\sum_{s=1}^{m}(U_1^s(w_2)+U_2^s(u_0))V_0^{2m-s+1}(v_0)=0. \tag{3.18}$$

Multiplication of (3.16) by v_1 with the use of (2.8) and (3.5) gives the term (w_2, v_0),

$$(w_2,v_0) = \lambda_2(u_0,v_1) - (l_1(u_0),v_1) + \sum_{s=1}^{m} U_1^s(u_0)V_0^{2m-s+1}(v_1). \tag{3.19}$$

The solution of (3.16) has the form $w_2 = \lambda_2 u_1 + \hat{w}_2 + \gamma u_0$, where γ is an arbitrary constant and $\hat{w}_2$ is a particular solution of the boundary value problem

$$l_0(\hat{w}_2) - \lambda_0 \hat{w}_2 = -l_1(u_0), \quad U_0^s(\hat{w}_2) = -U_1^s(u_0). \tag{3.20}$$

The solution of boundary value problem (3.20) exists due to degeneration condition (3.15), playing here the role of the solvability condition. Substituting (3.19) and the expression for w_2 into (3.18) we get the quadratic equation in λ_2,

$$\lambda_2^2 + \lambda_2 a_1 + a_2 = 0. \tag{3.21}$$

The coefficients a_1 and a_2 are determined by the expressions

$$a_1 = \frac{\sum_{s=1}^{m}[U_1^s(u_0)V_0^{2m-s+1}(v_1) + U_1^s(u_1)V_0^{2m-s+1}(v_0)]}{(u_0,v_1)} - \frac{(l_1(u_0),v_1)+(l_1(u_1),v_0)}{(u_0,v_1)}, \tag{3.22}$$

$$a_2 = \frac{-(l_1(\hat{w}_2),v_0) - (l_2(u_0),v_0) + \sum_{s=1}^{m}(U_1^s(\hat{w}_2)+U_2^s(u_0))V_0^{2m-s+1}(v_0)}{(u_0,v_1)}. \tag{3.23}$$

These coefficients can be written in the form

$$a_1 = \langle \mathbf{h}, \mathbf{e}_* \rangle + i\langle \mathbf{k}, \mathbf{e}_* \rangle, \quad a_2 = \langle \mathbf{H}\mathbf{e}_*, \mathbf{e}_* \rangle + i\langle \mathbf{K}\mathbf{e}_*, \mathbf{e}_* \rangle - \langle \mathbf{f}_2, \mathbf{d} \rangle - i\langle \mathbf{g}_2, \mathbf{d} \rangle, \tag{3.24}$$

where components of the real vectors $\mathbf{h}$, $\mathbf{k}$ are defined by the relationship

$$h^j + ik^j = \frac{\sum_{s=1}^{m}[\frac{\partial U^s}{\partial p_j}(u_0)V_0^{2m-s+1}(v_1) + \frac{\partial U^s}{\partial p_j}(u_1)V_0^{2m-s+1}(v_0)]}{(u_0,v_1)} - \frac{(\frac{\partial l}{\partial p_j}(u_0),v_1)+(\frac{\partial l}{\partial p_j}(u_1),v_0)}{(u_0,v_1)}, \tag{3.25}$$

and the real symmetric matrices $\mathbf{H}$ and $\mathbf{K}$ can be found according to (3.23) and (3.24). Thus, bifurcation of the double eigenvalue λ_0 in degenerate case (3.15) is described by the formula $\lambda = \lambda_0 + \epsilon\lambda_2 + o(\epsilon)$, where λ_2 are the two roots of (3.21).

Finally, we consider at $\mathbf{p} = \mathbf{p}_0$ a so-called [18] semisimple eigenvalue λ_0 of multiplicity k with k linearly independent eigenfunctions $u_0^1, u_0^2, \ldots, u_0^k$ satisfying eigenvalue problem (2.1). The complex-conjugate $\overline{\lambda}_0$ is the semisimple eigenvalue of adjoint eigenvalue problem (2.6) with the eigenfunctions $v_0^1, v_0^2, \ldots, v_0^k$.

Expansion of the parameters $\mathbf{p} = \mathbf{p}_0 + \epsilon\mathbf{e} + o(\epsilon)$ causes perturbations of the eigenvalue and eigenfunctions, which can be expressed as Taylor series with respect to the small parameter ϵ [11],

$$\lambda = \lambda_0 + \epsilon\lambda_1 + o(\epsilon), \quad u = w_0 + \epsilon w_1 + o(\epsilon). \tag{3.26}$$

Substituting these expansions as well as expansions (3.1) into eigenvalue problem (2.1) we get equations determining the functions w_0, w_1:

$$l_0(w_0) - \lambda_0 w_0 = 0, \quad U_0^s(w_0) = 0; \tag{3.27}$$

$$l_0(w_1) - \lambda_0 w_1 = -l_1(w_0) + \lambda_1 w_0, \quad U_0^s = -U_1^s(w_0). \tag{3.28}$$

A general solution of eigenvalue problem (3.27) has the form

$$w_0 = \sum_{i=1}^{k} \gamma_i u_0^i$$

with unknown coefficients γ_i. Taking the inner product of (3.28) and the functions v_0^1, $v_0^2, \ldots, v_0^k$ and using the Lagrange identity

$$(l_0(w_0) - \lambda_0 w_0, v_0^j) - (w_0, l_0^*(v_0^j) - \overline{\lambda_0} v_0^j) = -\sum_{s=1}^{m} U_1^s(w_0) V_0^{2m-s+1}(v_0^j),$$

we come to the system of equations on the coefficients $\gamma_1, \gamma_2, \ldots, \gamma_k$,

$$\sum_{i=1}^{k} \left((l_1(u_0^i), v_0^j) - \sum_{s=1}^{m} U_1^s(u_0^i) V_0^{2m-s+1}(v_0^j) - \lambda_1(u_0^i, v_0^j) \right) \gamma_i = 0, \quad j = 1, \ldots, k.$$

This system has a nontrivial solution if and only if

$$\det \left[(l_1(u_0^i), v_0^j) - \sum_{s=1}^{m} U_1^s(u_0^i) V_0^{2m-s+1}(v_0^j) - \lambda_1(u_0^i, v_0^j) \right] = 0. \tag{3.29}$$

The coefficients of (3.29) can also be expressed in terms of the vector of variation $\mathbf{e}$. For the sake of convenience we suppose that the eigenfunctions satisfy the orthonormality conditions

$$(u_0^\sigma, v_0^j) = \delta_{\sigma j}, \quad \sigma, j = 1, \ldots, k, \tag{3.30}$$

where $\delta_{\sigma j}$ is the Kronecker symbol. Introducing the real vectors $\mathbf{f}^{\sigma j}$ and $\mathbf{g}^{\sigma j}$ of dimension n with the components defined by the equation

$$f_r^{\sigma j} + i g_r^{\sigma j} = \left(\frac{\partial l}{\partial p_r}(u_0^\sigma), v_0^j \right) - \sum_{s=1}^{m} \frac{\partial U^s}{\partial p_r}(u_0^\sigma) V_0^{2m-s+1}(v_0^j), \quad r = 1, \ldots, n, \tag{3.31}$$

we can write (3.29) in the following form:

$$\det[\langle \mathbf{f}^{\sigma j} + i\mathbf{g}^{\sigma j}, \mathbf{e} \rangle - \lambda_1 \delta_{\sigma j}] = 0, \quad \sigma, j = 1, \ldots, k. \tag{3.32}$$

Equation (3.32) is a polynomial of kth order for the coefficients λ_1 in expansions (3.26), which describe splitting of a multiple eigenvalue λ_0.

4. Application to nonconservative stability problems. Let us consider nonconservative systems, described by a partial differential equation with the boundary conditions

$$\ddot{y} + l(y) = 0, \quad U^s(y) = 0, \quad s = 1, \ldots, m, \tag{4.1}$$

where $y = y(x,t)$, dots mean differentiation with respect to time t, while $l(y)$ and $U^s(y)$ are, respectively, the linear differential expression in terms of $x \in [0,1]$ and the boundary forms defined in section 2. Such systems are usually called *circulatory systems* [1], [2], [3]. Note that damping and gyroscopic forces are not involved in such systems. However, circulatory forces play an important role in aeroelasticity, plasma physics, gyrodynamics, and other fields.

Looking up the solution of (4.1) in the form

$$y(x,t) = u(x)e^{\pm it\sqrt{\lambda}}$$

we come to the eigenvalue problem (2.1). Recall that the coefficients of the differential expression $l(u)$ and the coefficients of the forms $U^s(u)$ are real functions, smoothly dependent on a vector of real parameters $\mathbf{p} = (p_1, p_2, \dots, p_n)$. It follows from the basic theorems of the theory of ordinary differential equations [20] that solutions $z_1, \dots, z_m$ of (2.1) with the initial conditions (δ_{ij} is the Kronecker symbol)

$$z_i^{(j-1)}(0) = \delta_{ij}, \quad i,j = 1, \dots, m,$$

forming the fundamental system of solutions of (2.1), smoothly depend on λ and $\mathbf{p}$. The characteristic determinant $\Delta \equiv \det \left\| U^i(z_j) \right\|$ is thus a smooth function of the spectral parameter λ and the vector $\mathbf{p}$: $\Delta = \Delta(\lambda, \mathbf{p})$.

We assume that at some fixed value $\mathbf{p}_0$ of the vector $\mathbf{p}$ the spectrum of the operator L formed by the differential expression $l(u)$ and boundary conditions $U^s(u) = 0$ is discrete. The eigenvalues λ can be simple or multiple roots of the characteristic equation $\Delta(\lambda, \mathbf{p}_0) = 0$.

If all eigenvalues λ_j are simple, then a general solution of (4.1) has the form

$$y(x,t) = \sum_{j=1}^{\infty} u_0^j(x)(\alpha_j e^{it\sqrt{\lambda_j}} + \beta_j e^{-it\sqrt{\lambda_j}}) \tag{4.2}$$

with arbitrary constants α_j, β_j. This form is also valid for semisimple eigenvalues λ_0 of algebraic multiplicity k ($\lambda_j = \lambda_{j+1} = \dots = \lambda_{j+k-1} = \lambda_0$), which means that the number of linearly independent eigenfunctions μ corresponding to λ_0 is equal to k. However, when $\mu < k$ in the general solution of (4.1), the *secular* terms proportional to $t^\sigma e^{\pm it\sqrt{\lambda_0}}$ with $\sigma < k$ appear.

Let $\lambda = \lambda_0$ be the k-fold root of the characteristic equation $\Delta(\lambda, \mathbf{p}_0) = 0$, and let the functions $u_0(x)$, $u_1(x), \dots, u_{k-1}(x)$ satisfying (3.4) be the Keldysh chain of length k corresponding to λ_0. In this case $\mu = 1$, and the partial solution of the boundary value problem (4.1) has the form

$$y(x,t) = (\alpha e^{it\sqrt{\lambda_0}} + \beta e^{-it\sqrt{\lambda_0}}) \sum_{r=0}^{k-1} y_r(x) \frac{t^{k-r-1}}{(k-r-1)!}, \tag{4.3}$$

$$y_r(x) = \sum_{j=0}^{r} (-1)^{r-j} (2i\sqrt{\lambda_0})^{r-2j} C_{r-j}^{r-2j} u_{r-j}(x), \quad C_{r-j}^{r-2j} = \begin{cases} 0, & 2j > r, \\ \frac{(r-j)!}{j!(r-2j)!}, & 2j \le r, \end{cases}$$

which can be verified by direct substitution to (4.1).

In the more complicated case of the multiple eigenvalue λ_0 of multiplicity $\sum_{s=1}^{\mu} k_s$ with μ Keldysh chains of lengths $k_1, k_2, \dots, k_\mu$, respectively, the solution corresponding to λ_0 is a sum of functions (4.3) over all μ Keldysh chains,

$$y(x,t) = \sum_{s=1}^{\mu} (\alpha_s e^{it\sqrt{\lambda_0}} + \beta_s e^{-it\sqrt{\lambda_0}}) \sum_{r=0}^{k_s-1} y_r^s(x) \frac{t^{k_s-r-1}}{(k_s-r-1)!}, \tag{4.4}$$

where the functions $y_r^s(x)$ are related to the functions $u_r^s(x)$ by the second of equations (4.3).

From (4.2)–(4.4) it is obvious that the linear system described by (4.1) is stable if and only if all the eigenvalues λ are nonnegative and semisimple. If all λ are real and some of them negative, then the circulatory system is statically unstable (divergence). Existence of at least one complex eigenvalue or a multiple real positive eigenvalue with the Keldysh chain of length $k > 1$ means flutter instability. Multiple zero eigenvalue with the Keldysh chain of length $k > 1$ causes divergence.

Let us now study the stability boundaries of circulatory systems in the parameter space. If at the point $\mathbf{p} = \mathbf{p}_0$ the characteristic equation has the k-fold real root $\lambda = \lambda_0$, i.e., $\Delta(\lambda_0, \mathbf{p}_0) = \partial\Delta/\partial\lambda = \cdots = \partial^{k-1}\Delta/\partial\lambda^{k-1} = 0$, $\partial^k\Delta/\partial\lambda^k \neq 0$, then according to Malgrange's preparation theorem [21] there exists a neighborhood $U_0 \subset R \times R^n$ of the point $(\lambda_0, \mathbf{p}_0)$, where $\Delta(\lambda, \mathbf{p})$ has the form

$$\Delta(\lambda, \mathbf{p}) = \left[(\lambda - \lambda_0)^k + \sum_{i=0}^{k-1}(\lambda - \lambda_0)^i a_i(\mathbf{p})\right] b(\lambda, \mathbf{p}). \tag{4.5}$$

The functions $a_0(\mathbf{p}), \ldots, a_{k-1}(\mathbf{p})$ and $b(\lambda, \mathbf{p})$ are real and smooth, $a_i(\mathbf{p}_0) = 0$, and $b(\lambda_0, \mathbf{p}_0) \neq 0$.

Let, for example, λ_0 be a simple real root of the equation $\Delta(\lambda, \mathbf{p}_0) = 0$. Then due to (4.5) we can write $\lambda = \lambda_0 - a_0(\mathbf{p})$, and λ remains real and simple in some neighborhood of the point $\mathbf{p}_0$. We can conclude from this fact that if at $\mathbf{p} = \mathbf{p}_0$ all the eigenvalues of eigenvalue problem (2.1) are positive and simple, then $\mathbf{p}_0$ is the inner point of the stability domain of circulatory system (4.1).

Similarly, the points of the parameter space, corresponding to either simple zero eigenvalue or real double eigenvalue with the Keldysh chain of length 2, form smooth surfaces of dimension $n-1$. Indeed, if $\lambda_0 = 0$ at $\mathbf{p} = \mathbf{p}_0$, then in the vicinity of $\mathbf{p}_0$ we have $\lambda = -a_0(\mathbf{p})$. The equation $a_0(\mathbf{p}) = 0$ defines a hypersurface in the parameter space.

If λ_0 is a double eigenvalue, then according to (4.1) its behavior near the point $\mathbf{p}_0$ is described by the quadratic equation

$$(\lambda - \lambda_0)^2 + a_1(\mathbf{p})(\lambda - \lambda_0) + a_0(\mathbf{p}) = 0. \tag{4.6}$$

It follows from (4.6) that the eigenvalue $\lambda(\mathbf{p})$ remains double in the neighborhood of the point $\mathbf{p}_0$ if $\mathbf{p}$ belongs to the hypersurface $a_1^2(\mathbf{p}) - 4a_0(\mathbf{p}) = 0$.

It is clear that the stability of the system in a neighborhood of the point $\mathbf{p}_0$ belonging to these hypersurfaces depends on the behavior of the zero or the double eigenvalues due to change of parameters if all other eigenvalues at the point $\mathbf{p}_0$ are positive and semisimple. According to (3.8) and (3.13), where we should put $k = 1$ or $k = 2$, the behavior of the simple zero eigenvalue is described by the formula

$$\lambda = \epsilon\langle \mathbf{f}_1, \mathbf{e}\rangle + o(\epsilon), \tag{4.7}$$

and the splitting of the real double λ_0 is governed by the expression

$$\lambda = \lambda_0 \pm \sqrt{\epsilon\langle \mathbf{f}_2, \mathbf{e}\rangle} + o(\epsilon^{1/2}). \tag{4.8}$$

The inequality $\langle \mathbf{f}_1, \mathbf{e}\rangle > 0$ defines a set of directions $\mathbf{e}$ such that the curves $\mathbf{p} = \mathbf{p}(\epsilon)$ emitted along these vectors lie in the stability domain, i.e., *a tangent cone* to the stability domain. The eigenvalue λ becomes negative for $\langle \mathbf{f}_1, \mathbf{e}\rangle < 0$. Consequently,

this inequality gives a tangent cone to the static instability (divergence) domain. The eigenvalue remains zero up to the terms of order ϵ^2 on the curves, emitted in the directions $\mathbf{e}$, such that $\langle \mathbf{f}_1, \mathbf{e} \rangle = 0$. Thus, the equation $\langle \mathbf{f}_1, \mathbf{p} - \mathbf{p}_0 \rangle = 0$ defines a tangent plane to the surface, where the operator L has a simple zero eigenvalue. If other eigenvalues remain simple and positive on this surface, then it forms a boundary between stability and divergence domains. The vector $\mathbf{f}_1$ is the normal vector to the boundary and is directed to the stability domain.

Analyzing splitting of the double eigenvalue with the formula (4.8) we see that the points of the parameter space, corresponding to the real double eigenvalue with the Keldysh chain of length 2, belong to the smooth parts of the boundary between the flutter domain and the stability domain if $\lambda_0 > 0$ or the divergence domain if $\lambda_0 < 0$. In this case the vector $\mathbf{f}_2$ is the normal vector to the flutter boundary looking at the stability or divergence domains, respectively.

Finally, we assume that at the point $\mathbf{p}_0$ there exists a double positive semisimple eigenvalue λ_0 with the two linearly independent eigenfunctions u_0^1 and u_0^2. We choose them, satisfying orthonormality condition (3.30). The splitting of this eigenvalue with a change of parameters is governed by (3.32), which for $k = 2$ takes the form

$$\lambda_1^2 - \lambda_1 \langle \mathbf{f}^{11} + \mathbf{f}^{22}, \mathbf{e} \rangle + \langle \mathbf{f}^{11}, \mathbf{e} \rangle \langle \mathbf{f}^{22}, \mathbf{e} \rangle - \langle \mathbf{f}^{12}, \mathbf{e} \rangle \langle \mathbf{f}^{21}, \mathbf{e} \rangle = 0, \tag{4.9}$$

where the real vectors $\mathbf{f}^{11}$, $\mathbf{f}^{12}$, $\mathbf{f}^{21}$, and $\mathbf{f}^{22}$ are defined in (3.31). The stability of the circulatory system near the point $\mathbf{p}_0$ depends on the sign of the discriminant D of quadratic equation (4.9), which can be written as

$$D = \langle \mathbf{f}^{11} - \mathbf{f}^{22}, \mathbf{e} \rangle^2 + \langle \mathbf{f}^{12} + \mathbf{f}^{21}, \mathbf{e} \rangle^2 - \langle \mathbf{f}^{12} - \mathbf{f}^{21}, \mathbf{e} \rangle^2. \tag{4.10}$$

The stability condition implies $D > 0$, while the flutter condition yields $D < 0$. The boundary between the flutter and stability domains in the parameter space is defined by the equality $D = 0$. It is easy to see from (4.10) that the equality $D = 0$ describes a circular cone $z^2 = x^2 + y^2$ in the space of three coordinates $x = \langle \mathbf{f}^{11} - \mathbf{f}^{22}, \mathbf{e} \rangle$, $y = \langle \mathbf{f}^{12} + \mathbf{f}^{21}, \mathbf{e} \rangle$, $z = \langle \mathbf{f}^{12} - \mathbf{f}^{21}, \mathbf{e} \rangle$, the flutter domain $z^2 > x^2 + y^2$ being inside the cone and the stability domain $z^2 < x^2 + y^2$ outside. The apex of the cone $(x = 0, y = 0, z = 0)$ corresponds to the double semisimple positive eigenvalue, while the skirts of the cone correspond to the double positive eigenvalues with the Keldysh chain.

Therefore, a double eigenvalue λ_0 defines a smooth surface of codimension 1 or the singularity of codimension 3 in the space of parameters $(p_1, \ldots, p_n)$ depending on the number of linearly independent eigenfunctions at λ_0. This result for matrices was established in [13]. Note that eigenvalue problems with the self-adjoint operators which are widely known in physics may have only semisimple multiple eigenvalues.

Eigenvalues of multiplicity higher than 2 are also responsible for the appearance of singularities on the stability boundaries, as is seen directly from (3.13) and Figure 3.1. Indeed, the eigenvalue with multiplicity $k > 1$ splits in the nondegenerate case into k distinct complex eigenvalues implying flutter instability of the system as it was discussed in the beginning of this section. This was also shown for the discrete circulatory systems in [16].

In aeroelasticity a condition of onset of flutter, which can be easily verified, is of major importance. For circulatory systems such a condition was derived by Plaut [22] in a form $(u_0, v_0) = 0$, where u_0 and v_0 are eigenfunctions of the adjoint problems. His derivation was based on the idea that the first derivative of the nonconservative load with respect to the spectral parameter λ becomes zero at the onset of flutter. This

question attracts current interest; see [23] and comments in [24]. However, Plaut's derivation is restricted by the assumptions that the flutter instability takes place due to interaction of only two eigenvalues in a one-parameter system and the nonconservative load parameter is a smooth function of λ. In our approach the "flutter condition" $(u_0, v_0) = 0$ is just a simple consequence of existence at a multiple eigenvalue λ_0 the Keldysh chain of length $k \geq 2$. Thus, this condition is satisfied at the flutter boundary of a multiparameter circulatory system.

5. Stability boundaries of the extended Beck problem. As an example of a continuous nonconservative mechanical system we consider stability of a uniform elastic cantilevered column of length L_c, loaded by a nonconservative force P; see Figure 5.1. It is assumed that the force P, which can be represented as the sum of a tangential follower force and a potential load, is acting at the free end of the column. The parameter $\eta \in [0, 1]$ measures the nonconservativity of the force P. The case $\eta = 1$ means that the column is loaded by purely tangential follower force (Beck's problem [25]). If $\eta = 0$, then the force P is potential (conservative). This problem was first considered by Dzhanelidze [26] and Kordas and Zyczkowski [27]. Note that the force P models the jet thrust acting on the free end of the column. Recent experiments on stability of such a column were carried out by Sugiyama; see [6]. We will investigate properties of the stability boundary in this problem.

Consider the transverse vibrations of the column in the plane OXY as in Figure 5.1. In the nondimensional variables

$$x = X/L_c, \quad y = Y/L_c, \quad \tau = t/\sqrt{\rho A L_c^4/EI}, \quad q = PL_c^2/EI,$$

the differential equation describing small in-plane vibrations of the column and the appropriate boundary conditions have the form

$$y''''(x,\tau) + qy''(x,\tau) + \ddot{y}(x,\tau) = 0,$$

$$y(0,\tau) = y'(0,\tau) = y''(1,\tau) = y'''(1,\tau) + (1-\eta)qy'(1,\tau) = 0.$$

Dots mean differentiation with respect to time τ, and primes denote differentiation with respect to coordinate x.

Separating time by $y(x,\tau) = u(x)\exp(i\sqrt{\lambda}\tau)$, we get the eigenvalue problem [27]

$$l(u) \equiv u'''' + qu'' = \lambda u, \tag{5.1}$$

$$\begin{aligned} U^1(u) &\equiv u(0) = 0, & U^3(u) &\equiv u''(1) = 0, \\ U^2(u) &\equiv u'(0) = 0, & U^4(u) &\equiv u'''(1) + (1-\eta)qu'(1) = 0. \end{aligned} \tag{5.2}$$

The corresponding adjoint eigenvalue problem looks like

$$l^*(v) \equiv v'''' + qv'' = \lambda v, \tag{5.3}$$

$$\begin{aligned} V^1(v) &\equiv -v(0) = 0, & V^3(v) &\equiv v''(1) + \eta q v(1) = 0, \\ V^2(v) &\equiv v'(0) = 0, & V^4(v) &\equiv -v'''(1) - qv'(1) = 0, \end{aligned} \tag{5.4}$$

and for the forms $V^5, \ldots, V^8$ we have

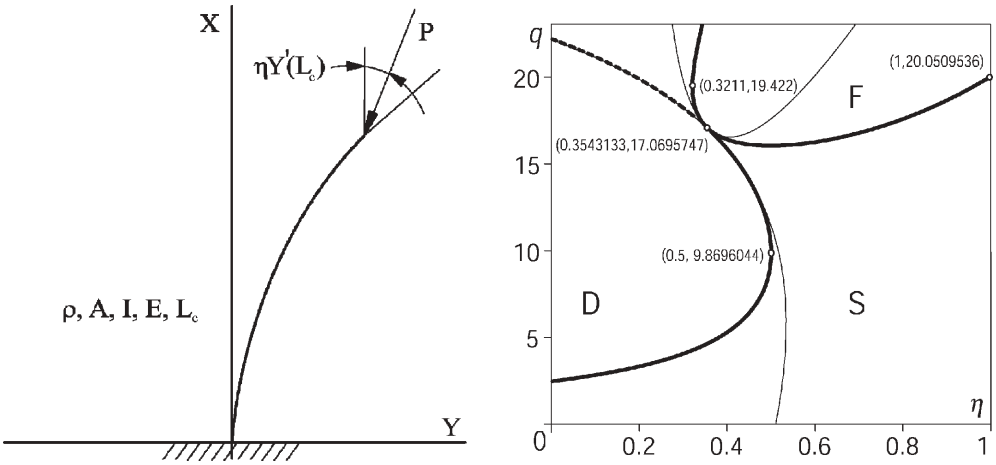


FIG. 5.1. *The extended Beck's problem and its stability diagram.*

$$V^5 \equiv v(1), \quad V^6 \equiv -v'(1), \quad V^7 \equiv -v''(0) - qv(0), \quad V^8 \equiv v'''(0) + qv'(0). \tag{5.5}$$

Substituting the general solution of differential equation (5.1)

$$u(x) = C_1 \cosh(ax) + C_2 \sinh(ax) + C_3 \cos(bx) + C_4 \sin(bx),$$

$$a = \sqrt{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \lambda}}, \quad b = \sqrt{\frac{q}{2} + \sqrt{\frac{q^2}{4} + \lambda}}, \quad \lambda \neq -\frac{q^2}{4},$$

into boundary conditions (5.2) we obtain the condition of the existence of a nontrivial solution $u(x)$ to eigenvalue problem (5.1), (5.2) in the form [27]

$$\Delta(\lambda, \eta, q) = 0, \tag{5.6}$$

$$\Delta \equiv (2\lambda + (1-\eta)q^2)(1 + \cosh(a)\cos(b)) + q(2\eta - 1)(q + ab\sinh(a)\sin(b)).$$

Equation (5.6) gives eigenvalues λ, depending on the parameters η and q.

The vertical equilibrium of the column is stable if all the eigenvalues λ are positive and semisimple; i.e., each eigenvalue has the same number of eigenfunctions as its algebraic multiplicity. After substitution $\lambda = 0$ in (5.6), it gives possible values of the parameters η and q at which the system loses stability statically [26], [27],

$$\eta(q) = \frac{\cos(\sqrt{q})}{\cos(\sqrt{q}) - 1}. \tag{5.7}$$

Equation (5.7) defines the curve of simple zero eigenvalues, the part of which forms the boundary between stability and divergence domains on the plane of parameters (η, q). The smooth parts of the flutter boundary consist of such points (η, q) that $\lambda(\eta, q)$ is a double real eigenvalue with the Keldysh chain. Calculation of the roots of characteristic equation (5.6) for different values of the load parameter q (at a fixed value of the parameter η) gives approximately the point where two simple eigenvalues form a double. Finding such points for different values of the parameter η we get the curve of double real eigenvalues.

The curves found subdivide the plane of the parameters (η, q) into stability (S), flutter (F), and divergence (D) domains; see Figure 5.1. The boundaries between these domains are shown in Figure 5.1 by the firm thick lines, while the dashed thick line shows the part of the curve of zero eigenvalues, which belongs to the divergence domain. Note that the boundary of stability domain has a singular point where the smoothness of the boundary is broken. The divergence boundary has two points with the vertical tangents. One can see that when the influence of the nonconservative part of the load q is small ($\eta < 0.5$), the column loses stability by divergence. Mainly nonconservative loads ($\eta > 0.5$) cause dynamical instability.

Our goal here is to demonstrate the advantages of the theory developed in the previous sections on the example of this stability problem. It will be applied for finding linear and quadratic approximations of the stability and instability domains both at singular and regular points of their boundaries. The explicit expression describing the overlapping of the frequency curves near the flutter boundary will be obtained and compared with the numerical results, given in [28]. Finally, we will obtain the exact coordinates of the singular point of the stability boundary, show that this point corresponds to the double zero eigenvalue with the Keldysh chain of length 2, and investigate the splitting of this eigenvalue in the vicinity of the singularity.

5.1. Bifurcation of eigenvalues in the vicinity of the flutter boundary. Consider a point $\mathbf{p}_0 = (\eta_0, q_0)$ of the flutter boundary, where the spectrum of the operator L contains double eigenvalue λ_0 with the Keldysh chain of length 2. Bifurcation of this eigenvalue is described by (4.8). Substituting the differential expression $l(u)$ from (5.1), the forms $U^1, \dots, U^4$ and $V^5, \dots, V^8$ from (5.2) and (5.5) into formula (3.14), we get the normal vector to the boundary,

$$\mathbf{f}_2 = \left(\frac{q_0 u_0'(1) v_0(1)}{\int_0^1 u_0 v_1 dx}, \frac{\int_0^1 u_0'' v_0 dx - (1-\eta_0) u_0'(1) v_0(1)}{\int_0^1 u_0 v_1 dx} \right). \tag{5.8}$$

For evaluation of the vector $\mathbf{f}_2$ it is essential to know the eigenfunctions u_0, v_0 as well as the associated functions u_1, v_1 at the double eigenvalue λ_0. The solution of eigenvalue problems (5.1), (5.2) and (5.3), (5.4) yields [5]

$$u_0(x) = \cosh(ax) - \cos(bx) + F(a\sin(bx) - b\sinh(ax)), \tag{5.9}$$

$$v_0(x) = \cosh(ax) - \cos(bx) + G(a\sin(bx) - b\sinh(ax)), \tag{5.10}$$

where the coefficients F and G depend on the parameters η and q:

$$F = \frac{a^2\cosh(a) + b^2\cos(b)}{ab(a\sinh(a) + b\sin(b))}, \quad G = \frac{(a^2+\eta q)\cosh(a) + (b^2 - \eta q)\cos(b)}{b(a^2+\eta q)\sinh(a) + a(b^2-\eta q)\sin(b)}. \tag{5.11}$$

Associated function u_1 is a solution of the boundary value problem (3.4), where we should put $k = 2$ and take the differential expression and the boundary forms from (5.1) and (5.2). A particular solution of the ordinary linear differential equation with constant coefficients

$$u_1'''' + q u_1'' - \lambda_0 u_1 = u_0,$$

whose right-hand side is the linear combination of trigonometric and hyperbolic functions (5.9), has the form

$$\hat{u}_1 = x(C_1\sin(bx) + C_2\cos(bx) + C_3\sinh(ax) + C_4\cosh(ax)).$$

Substitution of $\hat{u}_1$ into the second of equations (3.4) allows one to determine the coefficients $C_1, \dots, C_4$. After these coefficients are found one tries the solution of boundary value problem (3.4) in the form

$$u_1 = \hat{u}_1 + D_1 \sin(bx) + D_2 \cos(bx) + D_3 \sinh(ax) + D_4 \cosh(ax).$$

The unknown constants $D_1, \dots, D_4$ can be found from the boundary conditions (3.4). After all necessary manipulations we arrive at the associated function u_1,

$$u_1(x) = \frac{a\sin(bx) + b\sinh(ax) + F(a^2\cos(bx) - b^2\cosh(ax))}{2ab(a^2+b^2)}x + \frac{A_1\sinh(ax) - B_1\sin(bx)}{2ab(a^2+b^2)(a\sinh(a) + b\sin(b))^2}, \tag{5.12}$$

where the coefficient F is taken from (5.11), while for the coefficients A_1, B_1 we have the expressions

$$A_1 = \frac{\sin(b)(b^2\cos(b) - a^2\cosh(a)) + 2ab\cos(b)\sinh(a)}{a^2} q + b(a^2+b^2)(1+\cosh(a)\cos(b)),$$

$$B_1 = \frac{\sinh(a)(b^2\cos(b) - a^2\cosh(a)) - 2ab\cosh(a)\sin(b)}{b^2} q + a(a^2+b^2)(1+\cosh(a)\cos(b)).$$

Similarly, solving boundary value problem (3.5) with the differential expression and boundary forms from (5.1) and (5.2) we get the associated function v_1,

$$v_1(x) = \frac{a\sin(bx) + b\sinh(ax) + G(a^2\cos(bx) - b^2\cosh(ax))}{2ab(a^2+b^2)}x + \frac{A_2\sinh(ax) - B_2\sin(bx)}{2ab(a^2+b^2)(b(a^2+\eta q)\sinh(a) + a(b^2-\eta q)\sin(b))^2}, \tag{5.13}$$

where the coefficient G is defined in (5.11) and the coefficients A_2, B_2 are

$$A_2 = q\sin(b)[[-a^2b^2 + \eta((a^2+b^2)^2 - \eta q^2)]\cosh(a) + \cos(b)(b^2-\eta q)^2] + 2qab^3(1-2\eta)\sinh(a)\cos(b) + b(a^2+b^2)[a^2b^2 + \eta^2q^2 + (a^2b^2 + \eta(1-\eta)q^2)\cos(b)\cosh(a)];$$

$$B_2 = q\sinh(a)[[a^2b^2 - \eta((a^2+b^2)^2 - \eta q^2)]\cos(b) - \cosh(a)(a^2+\eta q)^2] - 2qba^3(1-2\eta)\sin(b)\cosh(a) + a(a^2+b^2)[a^2b^2 + \eta^2q^2 + (a^2b^2 + \eta(1-\eta)q^2)\cos(b)\cosh(a)].$$

A possibility of appearance of associated functions at the critical values of the nonconservative load in Beck's problem was noticed earlier in [29]. Nevertheless, the explicit expressions for the associated functions seem to be obtained first in the present paper. Note that although the eigenfunctions u_0, v_0 are defined up to arbitrary multipliers and associated functions u_1, v_1 are defined up to the addends C_1u_0 and C_2v_0, respectively, the vector $\mathbf{f}_2$ does not depend on these uncertainties.

Consider now the point $\mathbf{p}_0 = (1, 20.0509536)$, corresponding to the double eigenvalue $\lambda_0 = 121.347049$. This point is known as critical for the column subjected to a purely tangential follower force [25]. Substituting the values of λ_0 and $\mathbf{p}_0$ into (5.9)–(5.13) we obtain the functions u_0, v_0, u_1, v_1; see Figure 5.2.

Notice that for the case of tangential force ($\eta = 1$), the eigenfunction v_0 of the adjoint eigenvalue problem has a physical meaning. It is the vibrational mode for the loss of stability of a column loaded by a force with a fixed line of action; see [1] for the theory and [6] for the experiments.

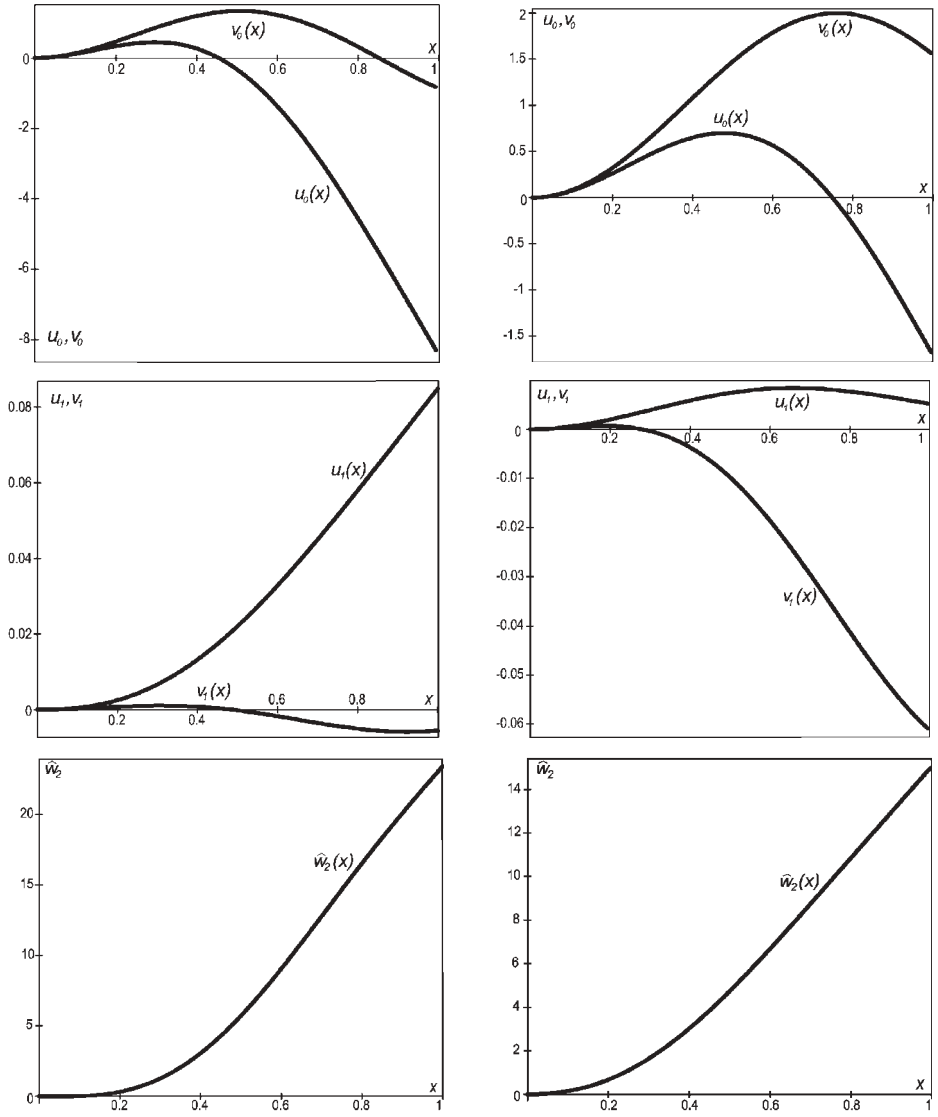


FIG. 5.2. *The eigenfunctions, associated functions, and functions* $\hat{w}_2(x)$ *at the points* $(1, 20.0509536)$ *and* $(0.35431330, 17.0695748)$*, the left and right columns, respectively.*

With the use of these functions, (5.8) gives the normal vector to the flutter boundary at the point $\mathbf{p}_0 = (1, 20.0509536)$,

$$\mathbf{f}_2 = (35458.3181, -2296.10536).$$

Let us look at the splitting of the double eigenvalue λ_0 due to change of parameters in different directions $\mathbf{e}$ on the parameter plane. Consider, for example, the vertical direction $\mathbf{e} = (0, 1)$. Taking into account that $\Delta\mathbf{p}$=$(0, q - q_0)$ we get from (4.8)

$$\lambda = 121.347049 \pm 47.9176936\sqrt{q_0 - q}. \tag{5.14}$$

For the horizontal variation $\Delta\mathbf{p} = (\eta - \eta_0, 0)$ corresponding to the vector $\mathbf{e} = (1, 0)$

TABLE 5.1
Splitting of the double eigenvalue near the point $\mathbf{p}_0 = (1, 20.0509536)$.

$(\Delta\eta, \Delta q)$	λ : Eqs. (5.14), (5.15), (5.19).	λ : Eq. (5.6)
$(0, \ 2{\cdot}10^{-5})$	$121.347049 \pm i0.21429444$	$121.342379 \pm i0.21422599$
$(0, -2{\cdot}10^{-5})$	121.132755 121.561343	121.128319 121.556963
$(\ 2{\cdot}10^{-5}, 0)$	122.189169 120.504929	122.188528 120.504432
$(-2{\cdot}10^{-5}, 0)$	$121.347049 \pm i0.84212016$	$121.338540 \pm i0.84201129$
$2{\cdot}10^{-3}\mathbf{e}_*$	120.762103 121.542851	120.755389 121.540824

we have

$$\lambda = 121.347049 \pm 188.303792\sqrt{\eta - \eta_0}. \tag{5.15}$$

The results of probing of a small neighborhood of the point $\mathbf{p}_0$ in different directions are summarized in Table 5.1. Thus, for example, for $q = q_0 + 0.00002$, i.e., when the new point is situated above the initial point $\mathbf{p}_0$, splitting yields $\lambda = 121.347049 \pm i0.21429444$, and thus the point $\mathbf{p}_0 + \Delta\mathbf{p}$ belongs to the flutter domain; see Figure 5.1. Characteristic equation (5.6) gives for the same values of parameters two complex conjugate eigenvalues, which differ from those found with the use of (5.14) only in a sixth digit; see Table 5.1.

Degeneration condition (3.15) defines the vector $\mathbf{e}_* = (-1, -15.4428097)$ tangent to the flutter boundary at the point $\mathbf{p}_0 = (1, 20.0509536)$. The double eigenvalue λ_0 splits in the tangent direction in accordance with (3.21),

$$\lambda = \lambda_0 - \frac{a_1}{2}\epsilon \pm \frac{\epsilon}{2}\sqrt{a_1^2 - 4a_2} + o(\epsilon). \tag{5.16}$$

Substitution of the differential expression $l(u)$ from (5.1) and the forms $U^1, \dots, U^4$, $V^5, \dots, V^8$ from (5.2) and (5.5) into (3.22) and (3.23) gives the coefficients a_1 and a_2 in the form

$$a_1 = \frac{e_2^* \int_0^1 (u_0'v_1' + v_0'u_1')dx - (e_2^*\eta_0 + e_1^*q_0)(v_1(1)u_0'(1) + v_0(1)u_1'(1))}{\int_0^1 u_0v_1dx}, \tag{5.17}$$

$$a_2 = \frac{e_2^* \int_0^1 v_0'\hat{w}_2'dx - (e_2^*\eta_0 + e_1^*q_0)v_0(1)\hat{w}_2'(1) - e_1^*e_2^*v_0(1)u_0'(1)}{\int_0^1 u_0v_1dx}.$$

The functions u_0, v_0, u_1, v_1 are presented by (5.9), (5.10), (5.12), and (5.13). The function $\hat{w}_2(x)$ (Figure 5.2) is a solution of boundary value problem (3.20), where the differential expressions l_0, l_1 and forms U_0^s, U_1^s are derived from differential expression (5.1) and boundary forms (5.2) according to (3.2) and (3.3):

$$\hat{w}_2(x) = \frac{b\sin(bx) - a\sinh(ax) + Fab(\cos(bx) + \cosh(ax))}{2(a^2 + b^2)}e_2^*x + \frac{A_3\sin(bx) - B_3\sinh(ax)}{2ab(a^2 + b^2)(b\sin(b) + a\sinh(a))^2}e_2^*. \tag{5.18}$$

The coefficient F in (5.18) is defined in (5.11), and for the coefficients A_3 and B_3 we have

$$\begin{aligned} A_3 =& -a(a^2+b^2)(q+ab\sin(b)\sinh(a)) + 2a^2b(b\sinh(a)\cos(b) - a\cosh(a)\sin(b)) \\ & - 2a^3\cosh(a)(a\sinh(a)+b\sin(b)), \\ B_3 =& -b(a^2+b^2)(q+ab\sin(b)\sinh(a)) + 2b^2a(b\sinh(a)\cos(b) - a\cosh(a)\sin(b)) \\ & + 2b^3\cos(b)(a\sinh(a)+b\sin(b)). \end{aligned}$$

With the use of the eigenfunctions, associated functions, and the function $\hat{w}_2$ we find from (5.17) the coefficients $a_1 = 194.571965$, $a_2 = -28633.4466$. Substitution of these coefficients into (5.16) gives approximate expressions for two simple eigenvalues which result from the splitting of the double λ_0 in the tangent direction to the stability boundary

$$\lambda_1 = 121.347049 - 292.473089\epsilon, \quad \lambda_2 = 121.347049 + 97.9011324\epsilon. \tag{5.19}$$

For example, take $\epsilon = 0.002$; then the double eigenvalue λ_0 splits into two positive eigenvalues (Table 5.1). This means that the tangent vector $\mathbf{e}_* = (-1, -15.4428097)$ lies in the stability domain, whence it follows that the flutter domain is convex at the point $\mathbf{p}_0$; see Figure 5.1. At the same values of the parameters, the characteristic equation has very close solutions (Table 5.1), showing thereby that formulas (5.19) give a good approximation to the directly computed eigenvalues.

Consider now the point $\mathbf{p}_0 = (0.32112653, 19.4220703)$ on the boundary between the flutter and divergence domains; see Figure 5.1. In this point there exists the negative double eigenvalue $\lambda_0 = -46.4046486$ with Keldysh chain of length 2. The normal vector to the flutter boundary evaluated at this point by formula (5.8) is

$$\mathbf{f}_2 = (-53123.691, 0).$$

The corresponding tangent vector to the boundary follows from degeneration condition (3.15),

$$\mathbf{e}_* = (0, 1).$$

One can see that the normal vector is parallel to the η-axis and is situated in the divergence domain, so the flutter boundary has a vertical tangent at the point $\mathbf{p}_0 = (0.32112653, 19.4220703)$; see Figure 5.1.

We are interested now in behavior of the frequency curves $\omega(q)$, where $\omega{=}\sqrt{\lambda}$ is a frequency of oscillations, in the vicinity of the point $\mathbf{p}_0$ of the boundary between the flutter and divergence domains. From the formula (3.21) it follows that along the curves $\mathbf{p} = \mathbf{p}_0 + \epsilon\mathbf{e}_* + \epsilon^2\mathbf{d} + o(\epsilon^2)$ the double eigenvalue splits according to

$$(\lambda - \lambda_0)^2 + \langle \mathbf{h}, \mathbf{e}_* \rangle (\lambda - \lambda_0)\epsilon + \langle \mathbf{H}\mathbf{e}_*, \mathbf{e}_* \rangle \epsilon^2 = \epsilon^2 \langle \mathbf{f}_2, \mathbf{d} \rangle + o(\epsilon^2); \tag{5.20}$$

see [30]. Taking into account that along the curve $\mathbf{p}(\epsilon)$ tangent to the flutter boundary at the point $\mathbf{p}_0$,

$$q - q_0 = \epsilon e_*^q + o(\epsilon), \quad \eta - \eta_0 = \epsilon e_*^\eta + \epsilon^2 d^\eta + o(\epsilon^2),$$

we convert expression (5.20) into

$$\left(\lambda - \lambda_0 + \frac{h^q}{2}(q - q_0)\right)^2 - \left(\frac{(h^q)^2}{4} - H^{qq}\right)(q - q_0)^2 = f_2^1(\eta - \eta_0). \tag{5.21}$$

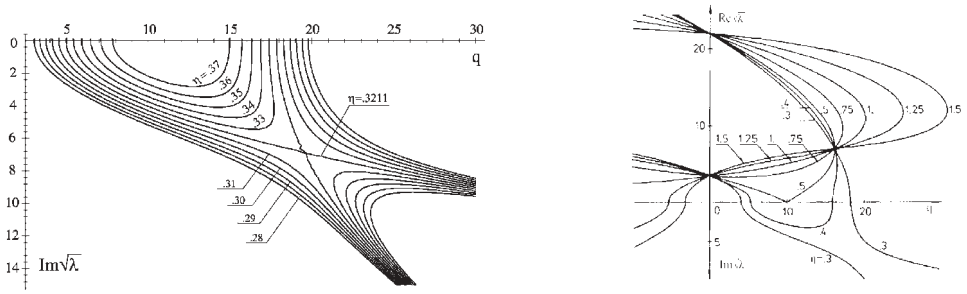


FIG. 5.3. *The frequency curves (from left to right) by* (5.21) *and by the work of* [28].

Formulas (3.22)–(3.25) give the components of the vector $\mathbf{h}$ and matrix $\mathbf{H}$,

$$h^\eta = 686267882692882, \quad h^q = 32.1039479,$$

$$H^{\eta\eta} = 0, \quad H^{\eta q} = 1917.18297, \quad H^{qq} = 93.4817323.$$

The double eigenvalue λ_0 does not split in the first approximation if the discriminant of (5.21) is zero. This condition gives us the quadratic approximation of the flutter boundary near the point $\mathbf{p}_0 = (0.32112653, 19.4220703)$,

$$\eta = 0.32112653 + 0.0030906(q - 19.4220703)^2. \tag{5.22}$$

Equation (5.22) shows that the flutter domain is convex at the point $\mathbf{p}_0$; see Figure 5.1. Formula (5.21) approximates in the vicinity of the point q_0, λ_0 the family of frequency curves $\omega(q) = \sqrt{\lambda(q)}$, parameterized by η. At $\eta = \eta_0$, (5.21) disintegrates into two parts:

$$q = 0.30878129(Im\sqrt{\lambda})^2 + 5.09318321, \quad q = 0.03464354(Im\sqrt{\lambda})^2 + 17.814449. \tag{5.23}$$

Parabolas (5.23) are symmetrical with respect to the axis q and are situated on the plane $(q, Im\sqrt{\lambda})$. At the points $(19.4220703, \pm 6.81209576)$ corresponding to two purely imaginary eigenfrequencies $\omega = \pm i 6.81209576$ these parabolas intersect.

In the left picture of Figure 5.3, the behavior of frequency curves described by (5.21) near one of the intersecting points is shown. One can see that at $\eta < \eta_0$ there exist two purely imaginary frequencies, meaning static instability. With the increase of η, frequency curves come closer together, overlap, and at $\eta > \eta_0$ move apart, forming a zone of complex eigenvalues (flutter). In the right picture of Figure 5.3, the dependence of the two lowest eigenfrequencies $\omega = \sqrt{\lambda}$ on the load q at the different values of the parameter $\eta \in [0.3, 1.5]$, obtained earlier in [28] by numerical solution of characteristic equation (5.6), is shown. Comparing the two pictures of Figure 5.3 we note a good qualitative and quantitative agreement in behavior of frequency curves calculated by two different methods in the range $\eta \in [0.3, 0.4]$.

5.2. Behavior of eigenvalues near the stability-divergence boundary. Consider a point $\mathbf{p}_0 = (\eta_0, q_0)$ on the boundary between the stability and divergence domains, where the spectrum of the eigenvalue problem (5.1), (5.2) contains a simple eigenvalue $\lambda_0 = 0$. Due to variation of parameters, a simple eigenvalue changes according to formula (4.7). Substituting the differential expression $l(u)$ from (5.1)

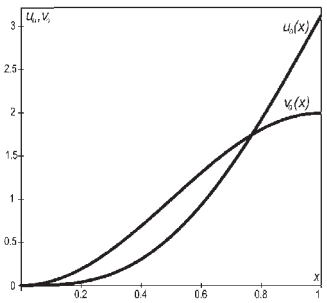


FIG. 5.4. *The eigenfunctions of the zero eigenvalue at the point* $\mathbf{p}_0 = (0.5, 9.86960440)$.

TABLE 5.2
Changing of zero eigenvalue near the point $\mathbf{p}_0 = (0.5, 9.86960440)$.

$(\Delta\eta, \Delta q)$	λ : Eq. (5.27)	λ : Eq. (5.6)
$(\ \ 10^{-4}, 0)$	0.00789568	0.00789498
$(-10^{-4}, 0)$	−0.00789568	−0.00789639

and the forms $U^1, \dots, U^4$ and $V^5, \dots, V^8$ from (5.2) and (5.5) into (3.14), we get the normal vector $\mathbf{f}_1$,

$$\mathbf{f}_1 = \left(\frac{q_0 u_0'(1) v_0(1)}{\int_0^1 u_0 v_0 dx}, \frac{\int_0^1 u_0'' v_0 dx - (1-\eta_0) u_0'(1) v_0(1)}{\int_0^1 u_0 v_0 dx} \right). \tag{5.24}$$

The eigenfunctions u_0 and v_0 at the simple zero eigenvalue have the form

$$u_0 = \sin(b) - xb\cos(b) - \sin(b)\cos(bx) + \cos(b)\sin(bx), \tag{5.25}$$

$$v_0 = 1 - \cos(bx), \quad b = \sqrt{q_0}. \tag{5.26}$$

These eigenfunctions are solutions of eigenvalue problems (5.1)–(5.4) at $\lambda_0 = 0$ and are presented in Figure 5.4.

Consider the point $\mathbf{p}_0 = (0.5, 9.86960440)$ on the divergence boundary described by (5.7). Substituting eigenfunctions (5.25) and (5.26) evaluated at this point into (5.24), we get the normal vector to the divergence boundary,

$$\mathbf{f}_1 = (78.9568352, 0).$$

Hence, the divergence boundary has the vertical tangent at the point $\mathbf{p}_0$; see Figure 5.1. Variation of the parameters $\Delta\mathbf{p} = (\eta - \eta_0, 0)$ changes the zero eigenvalue. According to (4.7) we have

$$\lambda = 78.956835(\eta - \eta_0). \tag{5.27}$$

One can see that for $\eta - \eta_0 < 0$ the eigenvalue $\lambda_0 = 0$ becomes negative. Therefore, the point $\mathbf{p}_0 + \Delta\mathbf{p}$ is inside the divergence domain; see Figure 5.1. If $\eta - \eta_0 > 0$, we come to the stability domain; see Table 5.2.

5.3. The singularity 0^2 of the stability boundary. Figure 5.1 clearly shows that the flutter domain has a common boundary with the domains of stability and divergence. Recall that at the points of the boundary between flutter and stability domains, the spectrum of the differential operator contains positive double eigenvalues, while at the points of the boundary between the flutter and divergence domains, double eigenvalues are negative.

Thus, the double eigenvalue becomes double zero at such point of the flutter boundary that separates stability and divergence domains. At the same time the point with double zero eigenvalue should belong to the curve of zero eigenvalues (5.7). Besides, due to (3.6) the orthogonality condition $\int_0^1 u_0 v_0 dx = 0$ must be true at the points of the flutter boundary. It is clear that this integral evaluated at the points of curve (5.7) becomes zero only at the point corresponding to the double zero eigenvalue.

Integrating the product of the eigenfunctions $u_0(x)$ and $v_0(x)$ from (5.25) and (5.26) over the range $[0, 1]$, we come to the transcendental equation for the ordinate of the desired point,

$$q_0 = (\sqrt{q_0} - 2\sin(\sqrt{q_0}))(\sqrt{q_0}(1 + 2\cos(\sqrt{q_0})) - 4\sin(\sqrt{q_0})). \tag{5.28}$$

The minimal element of the set of solutions of (5.28) at $q_0 > 0$ is $q_0 = 17.0695748$. Substituting this solution into (5.7), we find the corresponding value of the second parameter $\eta_0 = 0.35431330$.

Note that an equation similar to (5.28) was derived first in [27] from the analysis of characteristic equation (5.6) and without use of the eigenfunctions. However, formula (3.23) of [27] contains a misprint: the first term $k_2^2 l^2 \cos^2 k_2 l$ should be read as $k_2^2 l^2 \cos k_2 l$. Nevertheless, the coordinates of the singular point found in [27] are correct and coincide with those obtained from (5.28).

Thus, at the point $\mathbf{p}_0 = (0.35431330, 17.0695748)$ there exists the double eigenvalue $\lambda_0 = 0$ with the Keldysh chain of length 2. Following Arnold [13] we denote this point by the symbol 0^2, where the upper index means the length of the Keldysh chain corresponding to the double zero eigenvalue.

The bifurcation of a double eigenvalue is described by formula (4.8). To evaluate the normal vector $\mathbf{f}_2$ at this point, one needs to know the associated functions u_1, v_1 at the double zero eigenvalue along with the eigenfunctions u_0, v_0. Solving at $k = 2$ and $\lambda = 0$ boundary value problems (3.4) and (3.5) with the differential expressions and boundary forms from (5.1)–(5.4) we get

$$u_1 = -\frac{\cot(b)}{6b}x^3 + \frac{1}{2b^2}x^2 + \frac{\cot(b)(\cos(bx) - 1) + \sin(bx)}{2b^3}x + \frac{(bx - \sin(bx))(b + 2b\cos(b) - 2\sin(b))}{2b^4 \sin^2(b)}, \tag{5.29}$$

$$v_1 = \frac{x + x^2}{2b^2} + \frac{x - 1}{2b^3}\sin(bx) + \frac{b^2\cos(b) - \sin^2(b)}{b^4(b\cos(b) - \sin(b))}(\sin(bx) - bx), \tag{5.30}$$

where $b = \sqrt{q_0}$.

Substituting eigenfunctions (5.25) and (5.26) and associated function (5.30) into expression (5.8), we find the normal vector to the flutter boundary at the point $\mathbf{p}_0 = (0.35431330, 17.0695748)$,

$$\mathbf{f}_2 = (-24288.8139, -1024.49949). \tag{5.31}$$

TABLE 5.3
Splitting of the double zero near the singular point $\mathbf{p}_0 = (0.35431330, 17.0695748)$.

$(\Delta\eta, \Delta q)$	λ : Eqs. (5.32), (5.33)	λ : Eq. (5.6)
$(0, \ 10^{-4})$	$\pm i0.32007804$	$-0.00151188 \pm i0.32007586$
$(0, -10^{-4})$	0.32007804	0.32159210
	-0.32007804	-0.31856833
$(\ 10^{-4}, 0)$	$\pm i1.55848689$	$0.02668744 \pm i1.55823291$
$(-10^{-4}, 0)$	1.55848689	1.53205170
	-1.55848689	-1.58543004
$-10^{-5}\mathbf{e}_*$	-0.01207531	-0.01207543
	-0.00043108	-0.00043108
$10^{-5}\mathbf{e}_*$	0.01207531	0.01207520
	0.00043108	0.00043108

Knowing the normal vector allows us to study the neighborhood of the point of the flutter boundary in any direction $\mathbf{e}$ such that $\langle \mathbf{f}_2, \mathbf{e}\rangle \neq 0$. In particular, for two orthogonal directions $\mathbf{e} = (1, 0)$ and $\mathbf{e} = (0, 1)$, we get

$$\lambda = \pm 155.848689\sqrt{\eta_0 - \eta}, \quad \lambda = \pm 32.0078037\sqrt{q_0 - q}, \tag{5.32}$$

appropriately. It is easy to see that in the typical situation the double zero eigenvalue splits either into a complex conjugate pair or into two real eigenvalues, one of which is negative; see Table 5.3. Thus, the normal vector $\mathbf{f}_2$ at the point $\mathbf{p}_0$ is directed into the divergence domain. The inequality $\langle \mathbf{f}_2, \mathbf{e}\rangle > 0$ defines the tangent cone to this domain, and $\langle \mathbf{f}_2, \mathbf{e}\rangle < 0$ defines the tangent cone to the flutter domain; see Figure 5.1. Only curves, emitted in the tangent direction to the boundary, can reach the stability domain from the singular point.

Using the degeneration condition $\langle \mathbf{f}_2, \mathbf{e}_*\rangle = 0$, we find the tangent vector $\mathbf{e}_* = (1, -23.7079804)$. To examine whether this vector points to the stability domain, we should consider bifurcation of a double zero eigenvalue in the degenerate case. Substituting eigenfunctions (5.25) and (5.26), associated functions (5.29) and (5.30), and the function $\hat{w}_2$, which according to (5.18) takes the form

$$\hat{w}_2 = e_2^* x \frac{\cot(b)(\cos(bx) - 1) + \sin(bx)}{2b} + e_2^* \frac{bx - \sin(bx)}{2b\sin^2(b)}, \quad b = \sqrt{q_0},$$

into expressions (5.17) we find the coefficients of (5.16),

$$a_1 = 1250.63981, \quad a_2 = 52054.6889.$$

In accordance with (5.16) in the first approximation we have

$$\lambda_1 = 1207.53146\epsilon, \quad \lambda_2 = 43.1083501\epsilon. \tag{5.33}$$

It follows from (5.33) that the double zero eigenvalue splits into two positive simple eigenvalues (stability) only if the parameters change in the direction specified by the vector $\mathbf{e}_* = (1, -23.7079804)$; see Table 5.3. Changing the parameters in the opposite direction results in the splitting of the double $\lambda_0 = 0$ into two negative simple eigenvalues, which means static instability (divergence). Note that the approximate expressions for the eigenvalues are in a good agreement with the solutions of characteristic equation (5.6); see Table 5.3.

One can see that the tangent cone to the stability domain at the singular point is a ray on the plane of parameters. Stability domain in the vicinity of this point is a long narrow tongue (Figure 5.1). Our technique allows us to find the quadratic approximation of the flutter and divergence domains and therefore the stability domain near the singular point.

It is easy to see that (5.20), describing splitting of the double eigenvalue $\lambda_0 = 0$ along smooth curves tangent to the flutter boundary at the point $\mathbf{p} = \mathbf{p}_0$, can be rewritten as follows [30]:

$$\lambda^2 + \langle \mathbf{h}, \Delta\mathbf{p}\rangle\lambda + \langle \mathbf{H}\Delta\mathbf{p}, \Delta\mathbf{p}\rangle = \langle \mathbf{f}_2, \Delta\mathbf{p}\rangle + o(||\Delta\mathbf{p}||^2). \tag{5.34}$$

Components of the real vector $\mathbf{h}$ and real symmetrical matrix $\mathbf{H}$ are determined by formulas (3.22)–(3.25). Their evaluation at the singular point gives

$$\begin{gathered} h^\eta = -917.197355, \quad h^q = 14.0645660, \\ H^{\eta\eta} = 0, \quad H^{\eta q} = -690.854898, \quad H^{qq} = 34.3323737. \end{gathered} \tag{5.35}$$

Equation (5.34) provided that $\lambda = 0$ gives the quadratic approximation of the divergence boundary near the singular point,

$$f_2^\eta(\eta - \eta_0) + f_2^q(q - q_0) = 2H^{\eta q}(\eta - \eta_0)(q - q_0) + H^{qq}(q - q_0)^2. \tag{5.36}$$

The equality of the discriminant of (5.34) to zero guarantees the nonsplitting of the double zero eigenvalue and therefore defines the approximation of the flutter boundary

$$\begin{gathered} f_2^\eta(\eta - \eta_0) + f_2^q(q - q_0) \\ = (h^\eta(\eta - \eta_0) + h^q(q - q_0))^2/4 - (2H^{\eta q}(\eta - \eta_0)(q - q_0) + H^{qq}(q - q_0)^2). \end{gathered} \tag{5.37}$$

Substitution of the components of the normal vector $\mathbf{f}_2$ from (5.31) and the vector $\mathbf{h}$ and matrix $\mathbf{H}$ from (5.35) into (5.36) and (5.37) gives the quadratic approximations of the flutter and divergence domains in the vicinity of the point $\mathbf{p}_0 = (0.35431330, 17.0695748)$. These approximations are shown in Figure 5.1 by the thin solid lines. One can see that the approximation of the divergence domain is very good at far distances from the singular point while the approximation of the flutter domain is good only in the neighborhood of the point $\mathbf{p}_0$.

6. Conclusion. A new approach to obtain explicit formulas for the bifurcation of multiple eigenvalues of non-self-adjoint differential operators smoothly dependent on a vector of real parameters is presented. The formulas found use the derivatives of the differential expression and the boundary forms with respect to parameters as well as the functions of the Keldysh chain evaluated at the point of the parameter space corresponding to a multiple eigenvalue.

The results obtained let us study the splitting of the multiple eigenvalues in both regular and degenerate cases and serve as a basis for the sensitivity analysis of continuous nonconservative systems. This allows one to avoid the variational calculus in every specific problem to find sensitivities of eigenvalues or critical values of parameters.

Then the multiparameter stability problems of continuous circulatory systems are studied. It is found that the stability boundaries of these systems are smooth surfaces in the parameter space corresponding to simple zero (divergence) or double

real eigenvalues with Keldysh chain of length 2 (flutter). It is shown that the flutter condition for the circulatory systems is a simple consequence of the existence of the Keldysh chain of length $k \geq 2$.

The advantages of the proposed approach are illustrated by the mechanical example known as the extended Beck problem. With the use of the bifurcation analysis of eigenvalues, stability boundaries in this problem are investigated. Linear and quadratic approximations to the stability and instability domains at both regular and singular points of their boundaries are found and compared with the exact numerical values.

REFERENCES

[1] V. V. Bolotin, *Non-conservative Problems of the Theory of Elastic Stability*, Pergamon Press, Oxford, 1963.

[2] H. Ziegler, *Principles of Structural Stability*, Blaisdell, Waltham, MA, 1968.

[3] H. Leipholz, *Stability Theory*, John Wiley & Sons and B.G. Teubner, Stuttgart, 1987.

[4] M. P. Paidoussis, *Fluid-Structure Interactions: Slender Structures and Axial Flow,* Vol. 1, Academic Press, New York, London, 1998.

[5] P. Pedersen and A. P. Seyranian, *Sensitivity analysis for problems of dynamic stability*, Internat. J. Solids Structures, 19 (1983), pp. 315–335.

[6] M. A. Langthjem and Y. Sugiyama, *Dynamic stability of columns subjected to follower loads: A survey*, J. Sound Vibration, 238 (2001), pp. 809–851.

[7] E. B. Davies, *Non-self-adjoint differential operators. A review article*, Bull. London Math. Soc., 34 (2002), pp. 513–532.

[8] M. V. Keldysh, *On eigenvalues and eigenfunctions of some classes of nonselfadjoint equations*, Dokl. AN SSSR, 77 (1951), pp. 11–14 (in Russian).

[9] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, AMS, Providence, RI, 1969.

[10] I. C. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, San Diego, 1982.

[11] M. I. Vishik and L. A. Lyusternik, *Solution of some perturbation problems in the case of matrices and selfadjoint or non-selfadjoint equations*, Russian Math. Surveys, 15 (1960), pp. 1–73.

[12] J. Moro, J. V. Burke, and M. L. Overton, *On the Lidskii–Vishik–Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 793–817.

[13] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, Berlin, 1983.

[14] A. P. Seyranian, *Sensitivity analysis of multiple eigenvalues*, Mech. Structures Mach., 21 (1993), pp. 261–284.

[15] A. A. Mailybaev and A. P. Seyranian, *On singularities of a boundary of the stability domain*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 106–128.

[16] A. P. Seyranian and O. N. Kirillov, *Bifurcation diagrams and stability boundaries of circulatory systems*, Theoret. Appl. Mech., 26 (2001), pp. 135–168.

[17] M. Teytel, *How rare are multiple eigenvalues?*, Comm. Pure Appl. Math., 52 (1999), pp. 917–934.

[18] A. P. Seyranian and W. Kliem, *Bifurcations of eigenvalues of gyroscopic systems with parameters near stability boundaries*, Trans. ASME J. Appl. Mech., 68 (2001), pp. 199–205.

[19] M. A. Naimark, *Linear Differential Operators. Part* I, Frederick Ungar Publishing, New York, 1967.

[20] P. Hartman, *Ordinary Differential Equations*, Birkhauser, Boston, Basel, Stuttgart, 1982.

[21] C.-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Grundlehren Math. Wiss. 251, Springer-Verlag, New York, Berlin, Heidelberg, 1982.

[22] R. H. Plaut, *Determining the nature of instability in nonconservative problems*, AIAA J., 10 (1972), pp. 967–968.

[23] R. M. V. Pidaparti and D. Afolaby, *The role of eigenvectors in aeroelastic analysis*, J. Sound Vibration, 193 (1996), pp. 934–940.

[24] R. H. Plaut, *Comments on "The role of eigenvectors in aeroelastic analysis,"* J. Sound Vibration, 202 (1997), pp. 736–738.

[25] M. Beck, *Die Knicklast des einseitig eingespannten, tangential gedruckten Stabes*, ZAMM Z. Angew. Math. Mech., 3 (1952), pp. 225–228.
[26] G. Yu. Dzhanelidze, *On the stability of rods due to the action of follower forces*, Trud. Leningr. Polit. Instituta, 192 (1958), pp. 21–27 (in Russian).
[27] Z. Kordas and M. Zyczkowski, *On the loss of stability of a rod under a super-tangential force*, Arch. Mech. Stos., 15 (1963), pp. 7–31.
[28] P. Pedersen, *Influence of boundary conditions on the stability of a column under non-conservative load*, Internat. J. Solids Structures, 13 (1977), pp. 445–455.
[29] I. P. Andreichikov and V. I. Yudovich, *On the stability of viscoelastic rods*, Mekh. Tverd. Tela, 2 (1974), pp. 78–87 (in Russian).
[30] O. N. Kirillov and A. P. Seyranian, *Metamorphoses of characteristic curves in circulatory systems*, J. Appl. Math. Mech., 66 (2002), pp. 371–385.

SIAM J. Appl. Math.
Vol. 64, No. 4, pp. 1408–1433

DISCRETIZATION OF CONTINUOUS SPECTRA BASED ON PERFECTLY MATCHED LAYERS*

FRANK OLYSLAGER†

Abstract. As a tool of analysis in physics, wavefields are often expanded in a set of eigensolutions obtained from a Sturm–Liouville problem. For singular Sturm–Liouville problems subject to radiation boundary conditions, i.e., problems defined on an infinite domain, this set of eigensolutions has continuous parts. In this paper we will show that it is possible to approximate this continuous set of eigensolutions by a discrete set of eigensolutions of the same Sturm–Liouville operator but subject to Dirichlet boundary conditions in complex space. The idea of Dirichlet boundary conditions in complex space stems from the perfectly matched layer (PML) absorbing boundary condition. The PML was introduced in 1994 [J. P. Bérenger, *J. Comput. Phys.*, 114 (1994), pp. 185–200] as an absorbing termination of a finite difference time domain grid. These complex space Dirichlet boundary conditions have been used recently to close open electromagnetic waveguide structures. In the present paper we aim at developing a mathematical basis for the wavefields existing in such structures. On the one hand, this yields a better understanding of the properties of such waveguides and their applications in electromagnetic field problems. On the other hand, this opens the road for applications in other wavefields such as elastodynamics and quantum mechanics.

Key words. eigenfunction expansion, Green functions, waveguides, boundary value problems on infinite intervals, absorbing boundary conditions

AMS subject classifications. 34B24, 34B27, 34B40, 34L10, 78M25, 78A50

DOI. 10.1137/S0036139903430197

1. Introduction.

In 1994, Bérenger published [1] a new absorbing boundary condition for the finite difference time domain (FDTD) technique to solve electromagnetic field problems. That absorbing boundary condition, also called the perfectly matched layer (PML), has shown to be extremely useful in FDTD and finite element techniques. A PML is capable of absorbing without reflection the incident waves arriving from almost any direction and for all frequencies; i.e., it mimics infinite space. Originally Bérenger presented the PML in FDTD as a split field formalism. Later other formulations for a PML were given by interpreting a PML as an artificial uniaxial anisotropic medium [2] or as a layer with a complex thickness backed by a perfect electric conductor [3].

Originally, PMLs were used only in numerical, finite difference, or finite element–based techniques. Later it turned out that PMLs are also useful in semianalytical techniques. The first such application of PMLs aimed at the study of open waveguide discontinuities with the mode matching technique [4], [5]. An open waveguide is not bounded in the transverse plane orthogonal to the propagation direction. This means that the eigenmodes are solutions of a singular Sturm–Liouville problem over an infinite domain, subject to radiation boundary conditions. Hence, the modal spectrum of an open waveguide contains continuous parts corresponding to so-called radiation modes. These continuous parts impede the application of the mode matching technique. With the addition of a PML around the waveguide, the waveguide becomes a closed waveguide described by a Sturm–Liouville problem over a finite domain, having

*Received by the editors June 13, 2003; accepted for publication (in revised form) October 28, 2003; published electronically May 20, 2004.
http://www.siam.org/journals/siap/64-4/43019.html
†Electromagnetics Group, Department of Information Technology, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium (olyslag@intec.ugent.be).

only a discrete spectrum. Since the PML absorbs all incident waves without reflections, it is to be expected that this discrete spectrum is a good representation of the continuous spectrum of the open waveguide. In the subsequent years the usage of PMLs for mode matching techniques has shown to be very successful, especially in the study of optics; see, e.g., [6], [7].

Another semianalytical application of PMLs is intended for the determination of Green functions for electromagnetic fields in two and three dimensions in open layered media [8], [9]. To determine the Green function of a layered structure one usually first solves for the spectral Green function in the spatial Fourier domain with respect to the coordinates parallel to the layers. This spectral Green function is usually easily obtained in closed form. Its inverse Fourier transform cannot, in general, be evaluated in closed form and often relies on quite delicate numerical techniques. For the three-dimensional case these inverse Fourier transforms are known as Sommerfeld integrals. An expansion of the Green function in the eigenmodes of the waveguide consisting of the layered medium does not allow us to escape the numerical evaluation of an integral because the spectrum is again continuous due to the open nature of the structure. The spectral Green function has branch-cuts, corresponding to the continuous spectrum, apart from some discrete poles. With the addition of a PML, the layered medium again becomes closed, and the Green function can be expanded in a series of the discrete eigenmodes of the closed layered waveguide. The spectral Green function now has only discrete poles. By invoking Cauchy's residue theorem, this allows an expansion of the inverse Fourier transform in a series of residues in the discrete poles. This yields the modal expansion of the Green function.

In both these applications one could say that the addition of PMLs results in a discretization of continuous spectra.

All studies so far on semianalytical techniques mainly focussed on numerical accuracy and did not particularly explore the underlying mathematics. In the present paper we want to fill this gap and prove that field solutions in structures with PMLs converge to the field solutions in the original structures without a PML. At the same time, we will investigate the convergence properties of the discrete eigenmode series.

The usage of the PML technique to discretize continuous spectra is not restricted to electromagnetic field problems but can be applied to other and more general wave problems in physics and elsewhere. Our feeling is that this technique has advantageous applications in other branches of physics and especially in quantum physics, where one is often confronted with unbounded spaces, i.e., open structures, yielding continuous spectra of states. In order to avoid these continuous spectra, one often adds a box which again yields discrete spectra. The PML technique provides an alternative by adding a box with PML walls. We remark that this is easily implemented because the PML can be formulated as a Dirichlet boundary condition in complex rather than real space. By providing more mathematical rigor, it is hoped that the present paper will pave the way for the application of PMLs in semianalytical techniques outside the electromagnetic community. Noteworthy is a recent application of PMLs in economics for the prediction of the evolution of stock options [10].

In this paper we will treat the PML as a Dirichlet boundary condition placed at complex space coordinates. We will focus on the Green function problem; i.e., we will consider linear wave equations with Dirac distributions at the right-hand side. This poses no restriction, since arbitrary right-hand sides can be dealt with by considering convolutions with the Green functions. Our aim is to investigate when the Green function of a wave equation subject to radiation boundary conditions converges to the Green function of the same wave equation subject to Dirichlet boundary

conditions at points in complex space. First we will consider the Green function of a two-dimensional Helmholtz wave equation. This allows a simple treatment, and a closed form expression of the Green function is available, i.e., the Hankel function. By adding the Dirichlet boundary conditions, we obtain a new series representation for the Hankel function. It is shown that this series can be made as accurate as required by adjusting the position in complex space of the Dirichlet boundary conditions. Some of the results concerning this new series for the Hankel function were announced in [11]. We also investigate the conditions under which the series converges. In a second step this is generalized to the Green function of an inhomogeneous two-dimensional Helmholtz equation. In this case the Green function subject to radiation conditions is not expressible in closed form. However, the Green function subject to Dirichlet boundary conditions in complex space can be expanded in a series. For one particular case the convergence properties of the series are investigated. Last, some further generalizations are treated, such as Green functions of the homogeneous and inhomogeneous three-dimensional Helmholtz operator and of some vectorial Helmholtz wave equations. In all these cases the Green function is expanded in the discrete eigenfunctions of a wave operator, i.e., of a Sturm–Liouville problem, bounded by the Dirichlet boundary conditions. This wave operator is the spatial Fourier transform in one or more directions of the original wave equation. Along these directions the radiation condition is still imposed. In the last section we consider the situation in which the domain of the wave operator is subject to a Dirichlet boundary condition in all directions. The proofs of the various theorems are concentrated in Appendix A, and in Appendix B the asymptotic solutions of a transcendental equation are derived. A few numerical experiments illustrate some of our findings.

2. The basic problem. Consider the Green function that satisfies the two-dimensional Helmholtz equation

$$\nabla_{xy}^2 g(x,y) + g(x,y) = \delta(x)\delta(y) \tag{1}$$

over the entire xy-plane, i.e., $x, y \in \mathbf{R}$. To obtain a unique solution we demand that $g(x,y)$ represent outgoing waves when a $e^{j\omega t}$ time-dependence is assumed. This can be translated in the radiation condition [12]

$$\lim_{\rho\to+\infty}\left[\frac{\partial}{\partial\rho}g(x,y) + jg(x,y)\right] = 0, \tag{2}$$

with $\rho = \sqrt{x^2+y^2}$, which automatically selects the correct solution. It is well known that this solution is given by

$$g(x,y) = g(\rho) = \frac{j}{4}H_0^{(2)}(\rho), \tag{3}$$

with $H_0^{(2)}(\rho)$ the Hankel function of zeroth order and second kind.

Consider the spatial Fourier transform $G(\lambda, y)$ with respect to x of $g(\rho)$

$$G(\lambda,y) = \int_{-\infty}^{+\infty} g(\rho)e^{j\lambda x}dx. \tag{4}$$

The spectral Green function $G(\lambda, y)$ is a solution of

$$\frac{\mathrm{d}^2}{\mathrm{d}y^2}G(\lambda,y) + (1-\lambda^2)G(\lambda,y) = \delta(y) \tag{5}$$

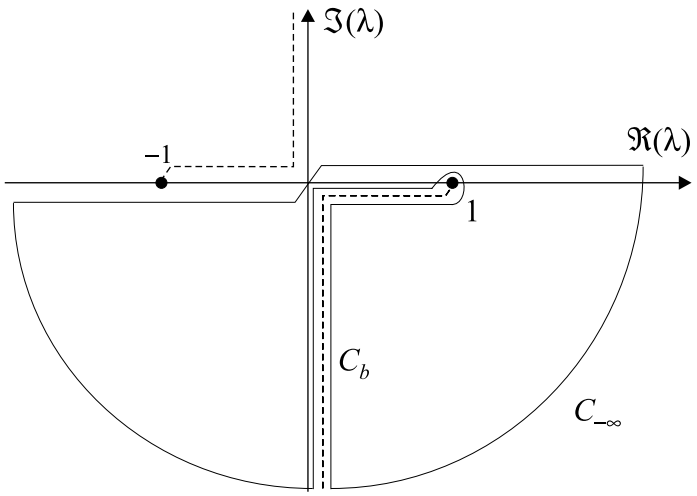


FIG. 1. *The complex λ-plane.*

and is given by

$$G(\lambda, y) = \frac{je^{-j\sqrt{1-\lambda^2}|y|}}{2\sqrt{1-\lambda^2}}. \tag{6}$$

Writing the inverse transform yields

$$\frac{j}{4}H_0^{(2)}(\rho) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{je^{-j\sqrt{1-\lambda^2}|y|}}{2\sqrt{1-\lambda^2}} e^{-j\lambda x} d\lambda. \tag{7}$$

If we choose the branch-cut of $\sqrt{\zeta}$ along $\arg(\zeta) = 0^+$, then the integrand has branch-cuts, emanating from $\lambda = \pm 1$, as indicated in the complex λ-plane in Figure 1. For $x > 0$ ($x < 0$) the integral along the real λ-axis can be replaced by an integral along the branch-cut in the lower (upper) half-plane, since the semicircle at infinity will not contribute. So far, this is a well known trivial matter.

Consider another Green function $\tilde{g}(x, y)$, a solution of (1) but subject to the boundary condition

$$\tilde{g}(x, y = \pm d) = 0, \tag{8}$$

$d \in \mathbf{R}$ that represents outgoing waves for $x \to \pm\infty$. The spectral Green function $\tilde{G}(\lambda, y)$ is readily found,

$$\tilde{G}(\lambda, y) = \frac{\sin[\sqrt{1-\lambda^2}(|y| - d)]}{2\sqrt{1-\lambda^2}\cos[\sqrt{1-\lambda^2}d]}, \tag{9}$$

and hence

$$\tilde{g}(x, y) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{\sin[\sqrt{1-\lambda^2}(|y| - d)]}{2\sqrt{1-\lambda^2}\cos[\sqrt{1-\lambda^2}d]} e^{-j\lambda x} d\lambda. \tag{10}$$

Let us extend d in (10) to the complex plane; i.e., assume that $d = \gamma e^{-j\alpha}$. Now we conjecture for all $x, y \in \mathbf{R}$ that

$$\forall \epsilon > 0, \qquad \exists \delta(\epsilon) > 0 : \gamma > \delta(\epsilon) > 0 \Rightarrow |g(x, y) - \tilde{g}(x, y)| < \epsilon \tag{11}$$

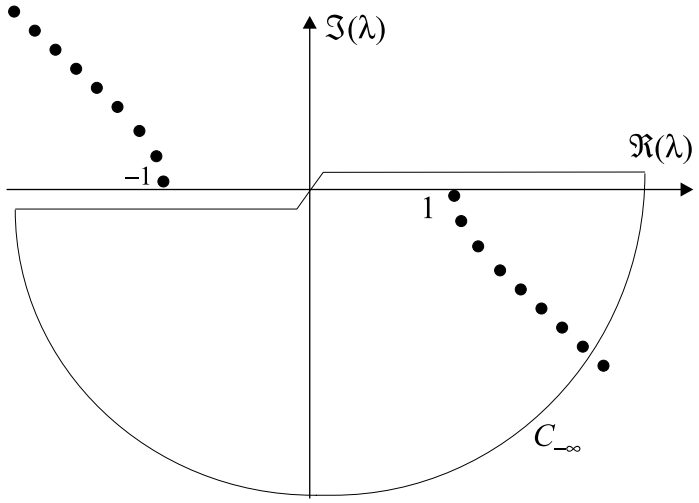


FIG. 2. *Complex λ-plane.*

when $0 < \alpha < \frac{\pi}{2}$. This can be seen from

$$g(x,y) - \tilde{g}(x,y) = -\frac{1}{\pi}\int_0^{+\infty}\left[\frac{\sinh[\sqrt{\lambda^2-1}(|y|-d)]}{\sqrt{\lambda^2-1}\cosh[\sqrt{\lambda^2-1}d]} + \frac{e^{-\sqrt{\lambda^2-1}|y|}}{\sqrt{\lambda^2-1}}\right]\cos(\lambda x)d\lambda$$

$$= \frac{1}{\pi}\int_0^{+\infty}\frac{\tanh(\sqrt{\lambda^2-1}\gamma e^{-j\alpha})-1}{\sqrt{\lambda^2-1}}\cosh(\sqrt{\lambda^2-1}|y|)\cos(\lambda x)d\lambda, \tag{12}$$

where the branch-cut of $\sqrt{\zeta}$ now is along $\arg(\zeta) = \pi^+$. The proof of (11) is an immediate consequence of Theorem A.4 in Appendix A. This means that we can use $\tilde{g}(x,y)$ as an approximation for $g(x,y)$, provided that γ is chosen large enough and that $0 < \alpha < \frac{\pi}{2}$.

Now consider the integral (10). The integrand has no branch-cuts but simple poles at

$$\lambda_n^{\pm} = \pm\sqrt{1 - \frac{(2n+1)^2\pi^2}{4d^2}}, \tag{13}$$

$n = 0, 1, 2, \ldots$. Figure 2 shows a typical distribution of these poles in the complex λ-plane. Let us assume that $x > 0$. Using Cauchy's theorem, the integral along the real axis can be replaced by a sum of the residues at the poles in the lower half-plane and an integral along the semicircle $C_{-\infty}$ at infinity in the lower half-plane. If the contribution of $C_{-\infty}$ vanishes, then (10) allows the following series representation:

$$\tilde{g}(x,y) = -\frac{j}{2d}\sum_{n=0}^{+\infty}\cos\frac{(2n+1)\pi y}{2d}\frac{e^{-j\sqrt{1-\frac{(2n+1)^2\pi^2}{4d^2}}x}}{\sqrt{1-\frac{(2n+1)^2\pi^2}{4d^2}}}. \tag{14}$$

Let us investigate when the contribution from $C_{-\infty}$ vanishes. At a semicircle with radius R in the lower half-plane, the spectral Green function $\tilde{G}(\lambda, y)$, with $\lambda = Re^{j\phi}$,

$-\pi \le \phi \le 0$, can be written as

$$\tilde{G}(\lambda, y) \approx -\frac{e^{-j\phi}}{2R} \frac{e^{Re^{j\phi}(|y|-\gamma e^{-j\alpha})} - e^{-Re^{j\phi}(|y|-\gamma e^{-j\alpha})}}{e^{Re^{j\phi}\gamma e^{-j\alpha}} + e^{-Re^{j\phi}\gamma e^{-j\alpha}}} e^{-j\lambda x}, \tag{15}$$

where we have assumed that R is large enough such that $\sqrt{1-\lambda^2} \approx jRe^{j\phi}$.

We distinguish two cases as follows.

1. Assume first that $\cos(\phi - \alpha) > 0$; this will be the case if $-\frac{\pi}{2} + \alpha \le \phi \le 0$. In this case we can write (15) as

$$\tilde{G}(\lambda, y) \approx -\frac{e^{-j\phi}}{2R} \frac{e^{R|y|e^{j\phi}} e^{-2R\gamma e^{j(\phi-\alpha)}} - e^{-Re^{j\phi}|y|}}{1 + e^{-2R\gamma e^{j(\phi-\alpha)}}} e^{-jRxe^{j\phi}}. \tag{16}$$

This function decays exponentially for increasing R if

$$|y| \cos\phi - 2\gamma \cos(\phi - \alpha) + x \sin\phi \le 0, \tag{17}$$

$$|y| \cos\phi + x \sin\phi \le 0. \tag{18}$$

Writing $x = \rho \cos\tau$ and $|y| = \rho \sin\tau$, with $0 \le \tau \le \frac{\pi}{2}$, this becomes

$$\rho \sin(\phi + \tau) - 2\gamma \cos(\phi - \alpha) \le 0, \tag{19}$$

$$\rho \sin(\phi - \tau) \le 0. \tag{20}$$

For the considered ranges of ϕ and τ the condition (20) is always satisfied. For the condition (19) it is easily checked that it is satisfied for $-\frac{\pi}{2} + \alpha \le \phi \le 0$ if it is satisfied for $\phi = 0$, i.e., if $\rho \sin\tau - 2\gamma \cos\alpha \le 0$. This can always be assured by taking γ large enough.

2. Now assume that $\cos(\phi - \alpha) < 0$, i.e., that $-\pi \le \phi \le -\frac{\pi}{2} + \alpha$. We can cast (15) as

$$\tilde{G}(\lambda, y) = -\frac{e^{-j\phi}}{2R} \frac{e^{R|y|e^{j\phi}} - e^{-Re^{j\phi}|y|} e^{2R\gamma e^{j(\phi-\alpha)}}}{e^{2R\gamma e^{j(\phi-\alpha)}} + 1} e^{-jRxe^{j\phi}}. \tag{21}$$

Exponential decay now demands that

$$\rho \sin(\phi + \tau) \le 0, \tag{22}$$

$$\rho \sin(\phi - \tau) + 2\gamma \cos(\phi - \alpha) \le 0. \tag{23}$$

The condition (23) can be assured by taking γ large enough. Indeed, one again checks that (23) is satisfied for all ϕ when it is satisfied for $\phi = -\pi$, i.e., when $\rho \sin\tau - 2\gamma \cos\alpha \le 0$. Condition (22) is a different matter. That condition is only satisfied for $-\pi \le \phi \le -\frac{\pi}{2} + \alpha$ when $\tau \le \frac{\pi}{2} - \alpha$.

By taking the limit for $R \to +\infty$, we have shown that the contribution of $C_{-\infty}$ will not vanish for x and y values for which $\tau > \frac{\pi}{2} - \alpha$. One could imagine closing the contour by a semicircle $C_{+\infty}$ at infinity in the upper half-plane to resolve this problem, but also the contribution of $C_{+\infty}$ will not vanish.

The previous steps can be repeated for $x < 0$ by closing the contour in the upper half-plane with $C_{+\infty}$ and considering the residues of the poles in the upper half-plane. The result is that a series representation of the integral (10) is possible in the nonhatched region of the xy-plane indicated in Figure 3. As a confirmation of this result one can check that the terms in the series (14) diverge exponentially in

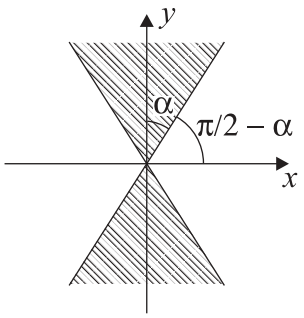


FIG. 3. *Region of convergence of the series* (14).

the hatched region of Figure 3. Outside the hatched region we have found a series approximation for the Green function (3), i.e., the Hankel function, that becomes more accurate if γ increases. In particular, we have that

$$H_0^{(2)}(x) \approx -\frac{2}{d}\sum_{n=0}^{+\infty} \frac{e^{-j\sqrt{1-\frac{(2n+1)^2\pi^2}{4d^2}}x}}{\sqrt{1-\frac{(2n+1)^2\pi^2}{4d^2}}} \tag{24}$$

for $x > 0$.

It is unnecessary to keep y real. One could as well analytically extend (14) to the complex plane with respect to y. Let us consider one particular example where we take $y = \tilde{y}e^{-j\alpha}$ with $\tilde{y} \in \mathbf{R}$. With this choice it is obvious that the series (14) is convergent for all $x > 0$ and for all $\tilde{y}$. It is then verified that $\tilde{g}(x, y = \pm d) = 0$ with d complex. Hence, we could say that $\tilde{g}(x, y)$ is the solution of (1) subject to Dirichlet boundary conditions in the complex y-plane. As was already the case on the real axis, the series (14) will not converge in the entire complex y-plane.

Applying Cauchy's theorem on the right-hand side of (7) allows us to write $g(x, y)$ for $x > 0$ as

$$\begin{aligned} g(x,y) &= \frac{1}{2\pi}\int_{C_b} \frac{je^{-j\sqrt{1-\lambda^2}|y|}}{2\sqrt{1-\lambda^2}} e^{-j\lambda x} d\lambda \\ &= -\frac{1}{2\pi}\int_0^1 \frac{e^{-j\lambda x}\cos(|\sqrt{1-\lambda^2}||y|)}{|\sqrt{1-\lambda^2}|} d\lambda + \frac{1}{2\pi}\int_0^{+\infty} \frac{e^{-\kappa x}\cos(|\sqrt{1+\kappa^2}||y|)}{|\sqrt{1+\kappa^2}|} d\kappa, \end{aligned} \tag{25}$$

where we have closed the contour with $C_{-\infty}$ and an integration along the branch-cut C_b as indicated on Figure 1. The right-hand side is nothing but an expansion of the Green function $g(x, y)$ in the continuous spectrum of eigenfunctions of the operator

$$\frac{\mathrm{d}^2}{\mathrm{d}y^2} + 1 \tag{26}$$

along the entire y-axis. The spectrum is continuous because the domain of the operator is infinite in the y-direction. The first term on the right-hand side represents propagating plane waves, and the second term on the right-hand side represents evanescent or inhomogeneous plane waves. The series (14) is also an expansion in eigenfunctions of the operator (26) but now subject to the Dirichlet boundary conditions at $y = \pm d$. These eigenfunctions constitute a discrete set. Physically we can thus say that we

have constructed a discrete spectrum representation of the continuous spectrum. In the analysis we considered the Green function problem, but it is clear that one could as well consider the field due to a general source (i.e., an arbitrary function $s(x, y)$ on the right-hand side in (1)) by taking the convolution with the Green function.

The expression (14) for the Green function $\tilde{g}(x, y)$ can also be obtained by a different route. Let us, for the moment, assume that the resolution

$$\delta(y) = \frac{1}{d} \sum_{n=0}^{+\infty} \cos \frac{(2n+1)\pi y}{2d} \tag{27}$$

of the $\delta(y)$ distribution remains valid for complex d. It then follows that the solution of (5) for $\tilde{G}(\lambda, y)$ can be written as

$$\tilde{G}(\lambda, y) = \frac{1}{d} \sum_{n=0}^{+\infty} \frac{\cos \frac{(2n+1)\pi y}{2d}}{1 - \lambda^2 - \frac{(2n+1)^2\pi^2}{4d^2}}, \tag{28}$$

and its inverse Fourier transform as

$$\tilde{g}(x, y) = \frac{1}{2\pi d} \sum_{n=0}^{+\infty} \cos \frac{(2n+1)\pi y}{2d} \int_{-\infty}^{+\infty} \frac{e^{-j\lambda x}}{1 - \lambda^2 - \frac{(2n+1)^2\pi^2}{4d^2}} d\lambda. \tag{29}$$

The integral is easily evaluated in closed form by again invoking Cauchy's theorem and closing, for $x > 0$, the contour along $C_{-\infty}$, the contribution of which always vanishes. The result is again (14). Since we know that (14) is invalid in the hatched region of Figure 3, we find that the resolution (27) of the Dirac distribution is invalid for real y. This is in agreement with [13]. However, for $y = \tilde{y}e^{-j\alpha}$ with $\tilde{y} \in \mathbf{R}$, (27) is a good resolution of the Dirac distribution in $-\gamma < \tilde{y} < \gamma$ since (14) is then valid for all x. This means that

$$\int_{-d}^{+d} h(y) \frac{1}{d} \sum_{n=0}^{+\infty} \cos \frac{(2n+1)\pi y}{2d} dy = h(0), \tag{30}$$

where $h(y)$ is holomorphic and where the path connecting $-d$ and d in the complex y-plane is arbitrary.

In Figure 4 the number of digits of accuracy of the series (24) is shown for three different distances x as a function of the parameter $q = \log_{10}(\gamma)$. It is seen that the accuracy increases with γ and that γ has to increase for a given accuracy when x increases in agreement with the aforementioned conclusion. In Figure 5 the number of terms needed in the series is shown to approximate the series sum with an accuracy of 10^{-7}. The results are plotted for the same distances x as a function of the parameter q.

3. A more general problem. Consider the Green function satisfying the inhomogeneous two-dimensional Helmholtz equation

$$\nabla_{xy}^2 g(x, y) + k^2(y) g(x, y) = \delta(x)\delta(y - y'), \tag{31}$$

where we assume that $y' \leq 0$, that $k^2(y) \in \mathbf{R}^+$, and that $k^2(y) = 1$ for $y > 0$. We also assume that $g(x, y)$ satisfies the radiation condition (2) in the half-plane $y > 0$. The spectral Green function $G(\lambda, y)$ satisfies

$$\frac{\mathrm{d}^2}{\mathrm{d}y^2} G(\lambda, y) + [k^2(y) - \lambda^2] G(\lambda, y) = \delta(y - y'). \tag{32}$$

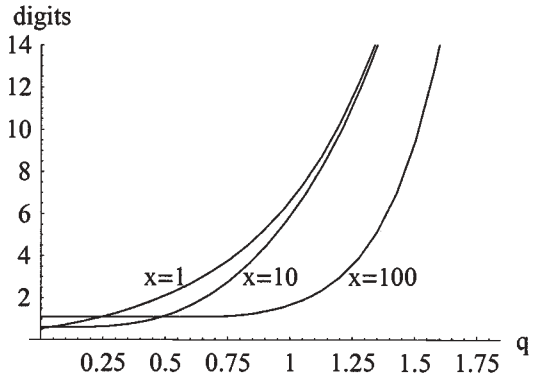


FIG. 4. *Number of digits of accuracy of* (24) *as a function of* $q = \log_{10}(\gamma)$ *for* $x = 1, 10, 100$.

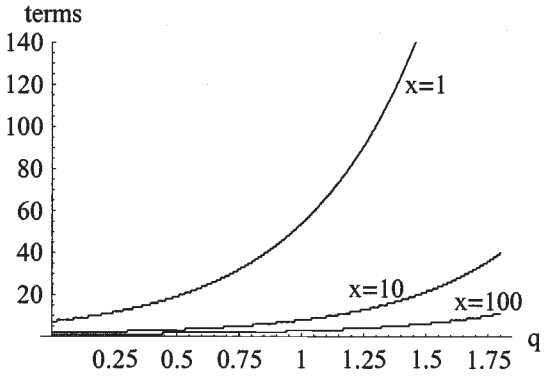


FIG. 5. *Number of terms needed in* (24) *as a function of* $q = \log_{10}(\gamma)$ *for an accuracy of* 10^{-7} *for* $x = 1, 10, 100$.

Note that $G(\lambda, y)$ is an even function of λ since $g(x, y)$ is an even function of x. In the region $y > 0$ we can write

$$G(\lambda, y) = G(\lambda, 0)e^{-j\sqrt{1-\lambda^2}y}. \tag{33}$$

Consider another Green function $\tilde{g}(x, y)$ that also satisfies (31) but subject to the boundary condition $\tilde{g}(x, y = d) = 0$. If $R(\lambda)$ is the reflection coefficient for an incident wave $e^{j\sqrt{1-\lambda^2}y}$ on the structure at $y = 0$, then we can expand $\tilde{G}(\lambda, y)$ in the region $y > 0$ as a series of waves

$$\begin{aligned}\tilde{G}(\lambda, y) &= G(\lambda, 0)\sum_{n=0}^{+\infty}[-R(\lambda)e^{-2jd\sqrt{1-\lambda^2}}]^n(e^{-j\sqrt{1-\lambda^2}y} - e^{j\sqrt{1-\lambda^2}(y-2d)}) \\ &= G(\lambda, 0)\frac{e^{-j\sqrt{1-\lambda^2}y} - e^{j\sqrt{1-\lambda^2}(y-2d)}}{1 + R(\lambda)e^{-2jd\sqrt{1-\lambda^2}}}. \end{aligned} \tag{34}$$

The term $n = 0$ corresponds to the wave $G(\lambda, y)$ and the down-propagating reflected

wave at $y = d$,

$$G_1(\lambda, y) = -G(\lambda, 0)e^{j\sqrt{1-\lambda^2}(y-2d)}, \tag{35}$$

with inverse Fourier transform

$$g_1(x, y) = -\frac{1}{\pi}\int_0^{+\infty} G(\lambda, 0)e^{-\sqrt{\lambda^2-1}(2d-y)}\cos(\lambda x)d\lambda. \tag{36}$$

From Theorem A.5 in Appendix A it follows, with $d = \gamma e^{-j\alpha}$ for all x, that

$$\forall \epsilon > 0, \qquad \exists \delta(\epsilon) > 0 : \gamma > \delta(\epsilon) > 0 \Rightarrow |g_1(x, 0)| < \epsilon. \tag{37}$$

This means that γ can be chosen such that $g_1(x, 0)$ becomes negligibly small. This in turn means that the waves generated by the incidence of $g_1(x, y)$ on the region $y < 0$ will be negligible compared to $g(x, y)$. In other words, for all x,

$$\forall \epsilon > 0, \qquad \exists \delta(\epsilon) > 0 : \gamma > \delta(\epsilon) > 0 \Rightarrow |g(x, y) - \tilde{g}(x, y)| < \epsilon \tag{38}$$

for $y < 0$.

Return to problem (31) and assume that $k^2(y)$ takes constant values $(k^+)^2$ for $y > a^+$ and $(k^-)^2$ for $y < a^-$, and assume that $g(x, y)$ represents outgoing waves in the regions $y > a^+$ and $y < a^-$; then this Green function can be approximated by another Green function $\tilde{g}(x, y)$ that satisfies (31) with $a^- \leq y' \leq a^+$ but subject to the boundary conditions $\tilde{g}(x, a^+ + d^+) = \tilde{g}(x, a^- - d^-) = 0$ with $d^\pm = \gamma^\pm e^{-j\alpha^\pm}$. If $0 < \alpha^\pm < \pi/2$, then the difference $|g(x, y) - \tilde{g}(x, y)|$ can be made as small as required by taking $\gamma^\pm$ sufficiently large.

The spectral Green function $G(\lambda, y)$ will have branch-cuts emanating from $\pm k^+$ and $\pm k^-$, together with some discrete poles [14], [15]. The discrete poles and the branch-cuts constitute the spectrum of the operator

$$\frac{\mathrm{d}^2}{\mathrm{d}y^2} + k^2(y), \tag{39}$$

where the continuous part, i.e., the branch-cuts, are again due to the infinite domain. Due to the Dirichlet boundary conditions in the y-direction, the spectral Green function $\tilde{G}(\lambda, y)$ will have only discrete poles and no branch-cuts; i.e., the operator (39) has a discrete spectrum. This means that $\tilde{g}(x, y)$ can be written as a sum of residues of a number of these discrete poles if the contributions from semicircles at infinity vanish. This series will approximate $g(x, y)$.

Let us consider one particular example where we look for solutions of (31) with $k^2(y) = 1$ for $y > y' > 0$ and $k^2(y) = k^2$ for $y < y'$. We bound the domain at $y = 0$ by imposing $g(x, y = 0) = 0$. The spectral Green function $G(\lambda, y)$ is readily given by

$$G(\lambda, y) = -\frac{\sin[\kappa_1(\lambda)y]}{j\kappa_2(\lambda)\sin[\kappa_1(\lambda)y'] + \kappa_1(\lambda)\cos[\kappa_1(\lambda)y']} \tag{40}$$

for $0 < y < y'$ and by

$$G(\lambda, y) = -\frac{e^{j\kappa_2(\lambda)y'}e^{-j\kappa_2(\lambda)y}\sin[\kappa_1(\lambda)y']}{j\kappa_2(\lambda)\sin[\kappa_1(\lambda)y'] + \kappa_1(\lambda)\cos[\kappa_1(\lambda)y']} \tag{41}$$

for $y > y'$ with $\kappa_1(\lambda) = \sqrt{k^2 - \lambda^2}$ and $\kappa_2(\lambda) = \sqrt{1 - \lambda^2}$. These functions clearly have branch-cuts emanating from $\lambda = \pm 1$, and their inverse Fourier transforms are not available in closed form. The spectral Green function $\tilde{G}(\lambda, y)$, on the other hand, is given by

$$\tilde{G}(\lambda, y) = -\frac{\sin[\kappa_2(\lambda)d]\sin[\kappa_1(\lambda)y]}{N(\lambda)} \tag{42}$$

for $0 < y < y'$ and by

$$\tilde{G}(\lambda, y) = -\frac{\sin[\kappa_1(\lambda)y']\sin[\kappa_2(\lambda)(y' + d - y)]}{N(\lambda)} \tag{43}$$

for $y > y'$ with $N(\lambda) = \kappa_2(\lambda)\sin[\kappa_1(\lambda)y']\cos[\kappa_2(\lambda)d] + \kappa_1(\lambda)\cos[\kappa_1(\lambda)y']\sin[\kappa_2(\lambda)d]$. This spectral Green function has no branch cuts but only poles at the zeroes of $N(\lambda)$.

Let us examine the case $y > y'$. The space domain Green function $\tilde{g}(x, y)$ can be calculated using Cauchy's theorem by closing the contour with $C_{-\infty}$ for $x > 0$. The result is

$$\begin{aligned} \tilde{g}(x, y) &= -\frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{\sin[\kappa_2(\lambda)(y' + d - y)]\sin[\kappa_1(\lambda)y']e^{-j\lambda x}}{N(\lambda)} d\lambda \\ &= j\sum_{i=0}^{+\infty} \frac{\sin[\kappa_2(\lambda_i)(y' + d - y)]\sin[\kappa_1(\lambda_i)y']e^{-j\lambda_i x}}{N'(\lambda_i)}, \end{aligned} \tag{44}$$

with $N(\lambda_i) = 0$ and $N'(\lambda) = \frac{d}{d\lambda}N(\lambda)$. In (44) it is assumed that the contribution of $C_{-\infty}$ vanishes. Let us prove that this assumption is correct by showing that the series in (44) converges for $x > 0$. To do so we look at the asymptotic form of the terms in the series (44). For this we need the asymptotic roots of the equation $N(\lambda_i) = 0$. This is discussed in Appendix B. It turns out that there are two branches of solutions.

The asymptotic solutions corresponding to the first branch can be approximated by the solutions of the equation

$$\kappa_2(\lambda)\cosh(\lambda d) + \kappa_1(\lambda)\sinh(\lambda d) = 0; \tag{45}$$

i.e., these solutions are independent of the value of y'. According to Appendix B, an asymptotic expression for the solutions of this equation is given by

$$\lambda_i \approx -\frac{1}{d}\log\left[\frac{2(i\pi + \alpha)}{\sqrt{k^2 - 1}\gamma}\right] - \frac{i\pi + \alpha}{d}j. \tag{46}$$

These values of λ_i allow us to approximate the terms in the series (44) for large values of i as

$$j\frac{\sin[\kappa_2(\lambda_i)(y' + d - y)]\sin[\kappa_1(\lambda_i)y']e^{-j\lambda_i x}}{N'(\lambda_i)} \approx a_i e^{-\frac{i\pi x}{d}} e^{-j\frac{i\pi\Delta y}{d}}, \tag{47}$$

with $\Delta y = y - y'$ and a_i such that $\lim_{i\to+\infty} a_i e^{-\epsilon i} = 0$ for all $\epsilon > 0$. For real values of y these terms are exponentially decaying if $x > \Delta y \tan\alpha$, which gives us the convergence region of the part of the series corresponding to the first branch. This region corresponds again to the nonhatched region of Figure 3 when we replace the y-parameter by Δy. In [9] the solutions corresponding to this first branch are called PML surface waves.

The solutions of the second branch asymptotically satisfy the equation

$$\kappa_2(\lambda)\sinh(\lambda y') - \kappa_1(\lambda)\cosh(\lambda y') = 0, \tag{48}$$

which shows that these solutions are independent of d. An asymptotic expression for the solutions of (48) is given by

$$\lambda_i \approx \frac{1}{y'}\log\left[\frac{(2i+1)\pi}{y'\sqrt{k^2-1}}\right] - \frac{(2i+1)\pi}{2y'}j. \tag{49}$$

For these values of λ_i the terms in the series (44) asymptotically behave as

$$j\frac{\sin[\kappa_2(\lambda_i)(y'+d-y)]\sin[\kappa_1(\lambda_i)y']e^{-j\lambda_i x}}{N'(\lambda_i)} \approx b_i e^{-\frac{(2i+1)\pi x}{2y'}} e^{-j\frac{(2i+1)\pi\Delta y}{2y'}}, \tag{50}$$

with $\lim_{i\to+\infty} b_i e^{-\epsilon i} = 0$ for all $\epsilon > 0$. For real values of y these terms are always decaying exponentially for $x > 0$, proving the convergence of the part of the series corresponding to the second branch. In [9] the solutions corresponding to the second branch are called pseudoleaky surface waves.

It can be shown that the corresponding series for $0 < y < y'$,

$$\tilde{g}(x,y) = j\sum_{i=0}^{+\infty}\frac{\sin[\kappa_2(\lambda_i)d]\sin[\kappa_1(\lambda_i)y]e^{-j\lambda_i x}}{N'(\lambda_i)}, \tag{51}$$

is convergent for all $x > 0$. Note also that for $\Delta y = \Delta\tilde{y}e^{-j\alpha}$ and $0 < \Delta\tilde{y} < \gamma$ both terms (47) and (50) decay exponentially for all $x > 0$.

Remarkable is that previous convergence analysis breaks down when $k = 1$. For $k = 1$ the conclusions regarding the convergence of (44) and (51) are invalid. This finds its origin in the fact that the asymptotic solutions do not separate anymore into two independent parts where one part depends only on d and the other only on y'. We will not pursue the case $k = 1$.

We can also construct a resolution of the $\delta(y-y')$ distribution in the eigenfunctions of the operator (39). Let us write for $0 < y < y'$

$$\delta(y-y') = \sum_{i=0}^{+\infty}\alpha_i\sin[\kappa_1(\lambda_i)y], \tag{52}$$

with yet unknown expansion coefficients α_i. From (32) it then follows that

$$\tilde{G}(\lambda,y) = \sum_{i=0}^{+\infty}\frac{\alpha_i}{\lambda_i^2-\lambda^2}\sin[\kappa_1(\lambda_i)y]. \tag{53}$$

Inverse Fourier transformation and identification with the series expansion (51) yields that

$$\alpha_i = -\frac{\lambda_i\sin[\kappa_2(\lambda_i)d]}{\pi N'(\lambda_i)}. \tag{54}$$

As a numerical example, assume that $k = 1.75$, that $y' = 2$, and that $d = \gamma e^{-j\pi/8}$, i.e., that $\alpha = \pi/8$. Figure 6 shows the solutions of $N(\lambda) = 0$ by large dots when $\gamma = 10$. There is a solution that is almost located on the real axis; its value is given

TABLE 1
Shift of the pole on the real axis to the complex plane as a function of γ.

γ	λ
10	$1.34274364519737586140 - 9.16986697792\,10^{-9}j$
20	$1.34274365934371669251 - 9.9180\,10^{-16}j$
∞	1.34274365934371713609

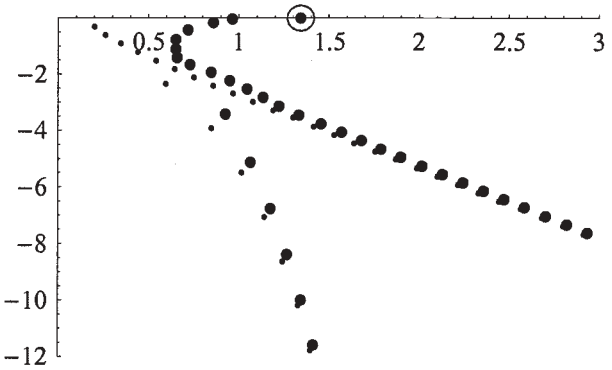


FIG. 6. *Solutions of $N(\lambda) = 0$ when $d = 10e^{-j\pi/8}$, $k = 1.75$, $y' = 2$.*

in Table 1. For $\gamma = 20$ this solution slightly shifts, as is also shown in Table 1. This solution corresponds to a zero of the denominator of (40) or (41) shown on the last row of Table 1 and indicated on Figure 6 by a circle. This is the only pole of the spectral Green function $G(\lambda, y)$. The asymptotic solutions (46) and (49) are also indicated on Figure 6 by small dots. For higher order solutions these asymptotic solutions better approximate the exact solutions. These asymptotic solutions allow for a very efficient determination of the exact solutions. One first calculates the asymptotic solutions (46) for $i = n$, $i = n + 1$, and $i = n + 2$, with n sufficiently large. Then one uses a Newton iteration technique starting from these asymptotic solutions to find the exact solutions. A quadratic extrapolation from these three exact solutions yields an approximate solution with index $n - 1$. The Newton technique is then used again to polish this solution to the exact solution. This process of quadratic extrapolation is repeated for decreasing index solutions. The same can be done starting from (49). The process breaks down at the very low index solutions. These low index solutions can be found by using a combined contour integration [9] and Newton technique. The solution that is located almost on the real axis can be found by first determining the real zero of the denominator of (40) and then polishing this solution to the corresponding solution of $N(\lambda) = 0$.

Assume that $y = y' = 2$ and $\gamma = 10$. Figure 7 shows the number of terms needed in the series (51) to approximate the series with a relative accuracy of 0.01, 0.0001, and 0.000001 as a function of the distance $s = \log_{10}(x)$. We have ordered the poles λ_i by decreasing imaginary part. Evidently, when x becomes larger, fewer terms are needed. When γ increases, the spacing between the poles decreases (cf. (46)), and hence more poles are needed to approximate the series to a given accuracy.

Finally, Figure 8 shows the absolute accuracy $a = |g(x, y') - \tilde{g}(x, y')|$ as a function of $s = \log_{10}(x)$ for $\gamma = 10$ and 20. In the series (51) we limited the sum to poles with

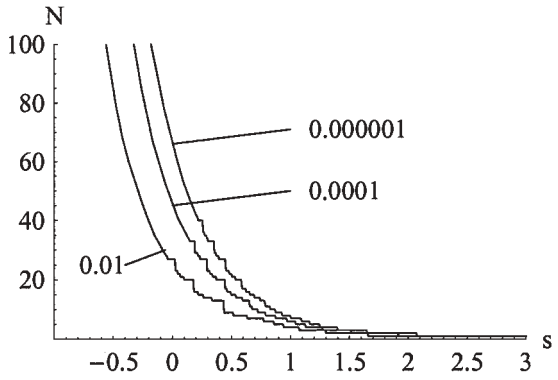


FIG. 7. *Number of terms in* (51) *as a function of* $s = \log_{10}(x)$ *for a relative accuracy of* $10^{-2}, 10^{-4}, 10^{-6}$.

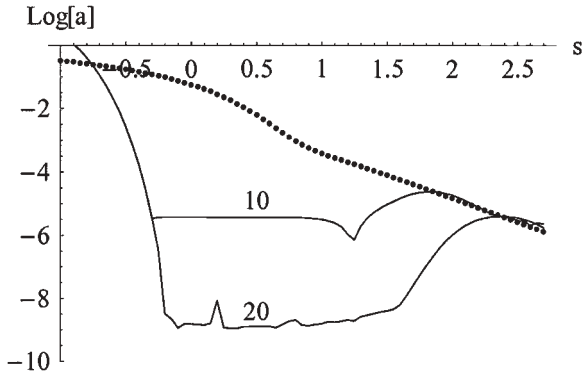


FIG. 8. *Absolute accuracy* $a = |g(x,y) - \tilde{g}(x,y)|$ *as a function of* $s = \log_{10}(x)$ *for* $\gamma = 10, 20$.

an imaginary part larger than -30. This explains the loss of accuracy for small x. As expected, when γ increases, the accuracy increases. The small ripple on the curve for $\gamma = 20$ is due to numerical noise in the evaluation of $g(x, y')$. The Green function $g(x, y')$ has been evaluated by a careful numerical integration of the inverse Fourier transformation of (40). For larger values of x we note a decrease in the accuracy. This is consistent with the results in section 2. However, for even larger values of x the accuracy increases again. In this regime it is only the pole of Table 1 that contributes to the series. In this regime, $g(x, y')$ is also approximated by the residue of (40) or (41) in this pole for $y = y'$. This residue is shown by a dotted line on the figure. For even larger values of x, not shown on Figure 8, the accuracy will decrease again because the small imaginary part of the pole close to the real axis will start to create significant exponential decay. Indeed, $\lim_{x\to+\infty} \tilde{g}(x, y') = 0$!

4. Further generalizations. The Green function of the three-dimensional Helmholtz-equation

$$\nabla^2_{xyz} g(x,y,z) + g(x,y,z) = \delta(x)\delta(y)\delta(z) \tag{55}$$

satisfying the Sommerfeld radiation condition is obviously

$$g(x,y,z) = g(r) = -\frac{e^{-jr}}{4\pi r}, \tag{56}$$

with $r = \sqrt{x^2+y^2+z^2}$. Expressing this as an inverse Fourier transform in the x- and y-direction of the spectral Green function yields

$$\begin{aligned} g(x,y,z) &= \frac{1}{2\pi}\int_0^{+\infty} \frac{je^{-j\sqrt{1-\lambda^2}|z|}}{2\sqrt{1-\lambda^2}} J_0(\lambda\sqrt{x^2+y^2})\lambda d\lambda \\ &= \frac{1}{4\pi}\int_{+\infty e^{-j\pi}}^{+\infty} \frac{je^{-j\sqrt{1-\lambda^2}|z|}}{2\sqrt{1-\lambda^2}} H_0^{(2)}(\lambda\sqrt{x^2+y^2})\lambda d\lambda, \end{aligned} \tag{57}$$

where $\lambda = \sqrt{\lambda_x^2+\lambda_y^2}$, with λ_x and λ_y the spectral variables in the x- and y-direction. This inverse Fourier transform is a Sommerfeld type of integral, which for this simple case can be evaluated in closed form. The spectral Green function has branch-cuts corresponding to the continuous set of eigenfunctions of $\nabla_{xy}+1$ consisting of propagating and evanescent cylindrical waves. For this simple case only the angular independent cylindrical waves are involved.

Consider another Green function $\tilde{g}(x,y,z)$ satisfying (55) but subject to the condition $\tilde{g}(x,y,z=\pm d)=0$. The corresponding spectral Green function is

$$\tilde{G}(\lambda,z) = \frac{\sin[\sqrt{1-\lambda^2}(|z|-d)]}{2\sqrt{1-\lambda^2}\cos[\sqrt{1-\lambda^2}d]}, \tag{58}$$

which obviously only has poles. Using Cauchy's theorem it is possible to expand $\tilde{g}(x,y,z)$ as the following series:

$$\tilde{g}(x,y,z) = -\frac{j}{4d}\sum_{n=0}^{+\infty}\cos\left[\frac{(2n+1)\pi z}{2d}\right] H_0^{(2)}\left[\sqrt{1-\frac{(2n+1)^2\pi^2}{4d^2}}\sqrt{x^2+y^2}\right]. \tag{59}$$

Since for large arguments the function $H_0^{(2)}(\zeta)$ behaves as $\sqrt{\frac{2j}{\pi\zeta}}e^{-j\zeta}$, it follows, by comparison with (14), that the series is convergent in the nonhatched region of Figure 3 if we replace y by z and x by ρ. From Theorem A.7 in Appendix A it follows again for a complex $d=\gamma e^{-j\alpha}$ and for all $x,y,z\in\mathbf{R}$ that

$$\forall\epsilon>0, \qquad \exists\delta(\epsilon)>0 : \gamma>\delta(\epsilon)>0 \Rightarrow |g(x,y,z)-\tilde{g}(x,y,z)|<\epsilon, \tag{60}$$

with $0<\alpha<\frac{\pi}{2}$. This can obviously be generalized to Green functions satisfying equations of the form

$$\nabla_{xyz}^2 g(x,y,z) + k^2(z)g(x,y,z) = \delta(x)\delta(y)\delta(z-z'). \tag{61}$$

In this case the Sommerfeld integrals cannot be evaluated in closed form. Numerical evaluation of Sommerfeld integrals is not always easy because the integrands are often highly oscillatory and have poles and branch-cuts. With the technique outlined here, a series representation is obtained for these Sommerfeld integrals. There are other discrete representations for Sommerfeld integrals, but those rely on rational approximations, i.e., Padé approximations [16]. To obtain these rational approximations Prony's techniques, or equivalents, are needed. Here, the series representation

follows in a natural way and requires only the determination of the complex zeroes of a transcendental function. In our series the x, y, z, and z'-dependence is available in explicit form, making this series representation especially attractive if the Green function is needed for many values of x, y, z, or z' and for further analytic manipulations. Other series representations [17] for Sommerfeld integrals can be obtained by transforming the integral to its steepest descent path and adding integral contributions from the branch-cuts and discrete contributions from the residues in the poles. The steepest descent integral and branch-cut integrals can then be approximated by Gaussian quadrature sums yielding a series representation. This technique is based on a pure numerical evaluation of the integrals and becomes complicated when poles, branch-cuts, and/or the steepest descent path come into each other's neighborhood. More on the evaluation of Sommerfeld integrals can be found in [18].

We can also derive another approximate series representation for the Green function $g(x, y, z)$ by considering a Green function $\hat{g}(x, y, z)$ that also satisfies (55) but now subject to the condition $\hat{g}(\rho = \sqrt{x^2 + y^2} = d, z) = 0$. By using a Fourier transform along the z-axis, we can express $\hat{g}(x, y, z)$ as

$$\hat{g}(x, y, z) = \frac{j}{4d} \sum_{n=1}^{+\infty} \frac{\sqrt{1-\lambda_n^2} J_0(\sqrt{1-\lambda_n^2}\rho) Y_0(\sqrt{1-\lambda_n^2} d)}{\lambda_n J_1(\sqrt{1-\lambda_n^2} d)} e^{-j\lambda_n z} \tag{62}$$

with $\lambda_n = \sqrt{1-\frac{\alpha_n^2}{d^2}}$, $n = 1, 2, \ldots$, where $\alpha_n \in \mathbf{R}$ are the zeroes of J_0. With $d = \gamma e^{-j\alpha}$, $0 < \alpha < \frac{\pi}{2}$, and γ sufficiently large, the function $\hat{g}(x, y, z)$ will again approximate $g(x, y, z)$. The condition $\hat{g}(\rho = d, z) = 0$ is equivalent to adding a curvilinear PML [19].

Still another approximation $\check{g}(x, y, z)$ can be obtained by imposing that $\check{g}(x = \pm d, y = \pm d, z) = 0$. The result is

$$\check{g}(x, y, z) = -\frac{j}{2d^2} \sum_{n,m=0}^{+\infty} \cos\frac{(2n+1)\pi x}{2d} \cos\frac{(2m+1)\pi y}{2d} \frac{e^{-j\kappa_{nm}|z|}}{\kappa_{nm}}, \tag{63}$$

with

$$\kappa_{nm} = \sqrt{1 - \frac{(2n+1)^2\pi^2}{4d^2} - \frac{(2m+1)^2\pi^2}{4d^2}}. \tag{64}$$

In analogy with section 2, this series will converge when $|z| \leq |x| \tan\alpha$ and $|z| \leq |y| \tan\alpha$.

Also Green functions satisfying equations of the form

$$\nabla_{xyz}^2 g(x, y, z) + k^2(x, y) g(x, y, z) = \delta(x - x')\delta(y - y')\delta(z) \tag{65}$$

can be treated in the same way as $\hat{g}(x, y, z)$ or $\check{g}(x, y, z)$.

Further generalizations include Green functions of other more complex wave operators such as

$$\nabla_{xyz} \cdot \overline{\overline{A}}(z) \cdot \nabla_{xyz} g(x, y, z) + \mathbf{b}(z) \cdot \nabla_{xyz} g(x, y, z) + k^2(z) g(x, y, z)$$

$$= \delta(x)\delta(y)\delta(z - z'), \tag{66}$$

or higher order operators. Equations of this form are obtained when considering electromagnetic field problems in anisotropic media (see [20] and references therein).

However, for that type of problems the present analysis needs to be adapted and further investigated; it will not just suffice to place a Dirichlet boundary condition in complex space. Indeed, adding PMLs to anisotropic media can cause problems, as has been reported in [21]. For an overview of some closed form Green functions of wave equations in infinite homogeneous space, i.e., of wave equations with constant coefficients, we refer the reader to [22].

In [9] series representations were derived for the Green tensor $\overline{\overline{g}}(x,y,z)$ of a vectorial wave equation of the form

$$\nabla_{xyz} \times [\alpha(z) \nabla_{xyz} \times \overline{\overline{g}}(x,y,z)] - k^2(z) \overline{\overline{g}}(x,y,z) = \delta(x)\delta(y)\delta(z-z')\overline{\overline{I}}, \tag{67}$$

with $\overline{\overline{I}}$ the unit tensor. The study in [9] focussed on a numerical investigation of the accuracy of the series.

5. Further remarks. In the previous sections we considered Green functions of wave equations with coefficients that were independent of at least one coordinate along which the outgoing wave condition was imposed. Along this direction we could apply a spatial Fourier transformation and write the result as a series of eigenfunctions.

Let us now consider cases where there is no such direction. As a simplest possible case consider

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} g(x) + g(x) = \delta(x), \tag{68}$$

where $g(x)$ represents outgoing waves. The solution obviously is

$$g(x) = -\frac{e^{-j|x|}}{2j}. \tag{69}$$

If we consider another Green function $\tilde{g}(x)$ satisfying (68) but subject to $\tilde{g}(x = \pm d) = 0$, then there is no invariant direction. Obviously $\tilde{g}(x)$ is given by

$$\tilde{g}(x) = \frac{1}{2} \frac{\sin(|x| - d)}{\cos d} = -\frac{1}{2j} \frac{e^{-j|x|} - e^{j|x|-2jd}}{1 + e^{-2jd}}. \tag{70}$$

If $d = \gamma e^{-j\alpha}$ with $0 < \alpha < \pi/2$, then one easily verifies for all $x \in \mathbf{R}$ that

$$|\tilde{g}(x) - g(x)| \leq e^{-2\gamma \sin \alpha}, \tag{71}$$

proving the convergence of $\tilde{g}(x)$ to $g(x)$ when γ increases. Using the result (28) for $\lambda = 0$, we can formally write

$$\tilde{g}(x) = \frac{1}{d} \sum_{n=0}^{+\infty} \frac{\cos \frac{(2n+1)\pi x}{2d}}{1 - \frac{(2n+1)^2 \pi^2}{4d^2}}. \tag{72}$$

For $x \neq 0$ this series diverges again, illustrating that the resolution (27) is invalid on the real axis.

Another case is the Green function (56) of the three-dimensional Helmholtz equation (55). We approximate this Green function by another Green function $\tilde{g}(x,y,z)$ that satisfies $\tilde{g}(x = \pm d, y = \pm d, z = \pm d) = 0$. Using the result (63), one can show that this Green function can be written as

$$\tilde{g}(x,y,z) = \frac{1}{2d^2} \sum_{n=0}^{+\infty} \sum_{m=0}^{+\infty} \cos \frac{(2n+1)\pi x}{2d} \cos \frac{(2m+1)\pi y}{2d} \frac{\sin[\kappa_{nm}(|z| - d)]}{\kappa_{nm} \cos[\kappa_{nm} d]}, \tag{73}$$

with κ_{nm} defined in (64). This series is still another approximation for the Green function (56) which converges if $|z| \leq |x| \tan\alpha$ and $|z| \leq |y| \tan\alpha$. Evidently more approximations for that Green function can be found by, e.g., imposing Dirichlet boundary conditions on a spherical or a finite cylindrical surface in complex space.

6. Conclusions. We have shown that the continuous spectrum of a singular Sturm–Liouville eigenvalue problem subject to radiation boundary conditions can be approximated by the discrete spectrum of the same Sturm–Liouville problem but now subject to Dirichlet boundary conditions at points in complex space. This allows new approximate series expansions for Green functions. The conditions under which these series converge were investigated.

Appendix A. Theorems and lemmas.

LEMMA A.1.

$$\forall \zeta \in \mathbf{C}, \qquad \xi = \Re(\zeta) > 1 \Rightarrow |\tanh\zeta - 1| < 4e^{-2\xi}. \tag{74}$$

Proof. From

$$|\tanh\zeta - 1| = \left|\frac{2e^{-2\zeta}}{1+e^{-2\zeta}}\right| = \frac{2e^{-2\xi}}{|1+e^{-2\zeta}|}, \tag{75}$$

it follows that

$$|\tanh\zeta - 1| < \frac{2e^{-2\xi}}{1-e^{-2\xi}} = \frac{2}{e^{2\xi}-1} < \frac{2}{\frac{e^{2\xi}}{2}} = 4e^{-2\xi} \tag{76}$$

if $\xi > 1$, since $e^{2\xi} - 1 > e^{2\xi}/2$ when $\xi > 1$. □

THEOREM A.2. *If* $0 < \alpha < \pi/2$ *and* $x \in \mathbf{R}$*, then*

$$\forall \epsilon > 0, \quad \exists \delta(\epsilon) > 0 : \gamma > \delta(\epsilon) \Rightarrow \left| \int_0^{+\infty} \frac{\tanh(\sqrt{\lambda^2-1}\gamma e^{-j\alpha}) - 1}{\sqrt{\lambda^2-1}} \cos(\lambda x) d\lambda \right| < \epsilon, \tag{77}$$

where the branch-cut of $\sqrt{\zeta}$ *is along* $\arg(\zeta) = \pi^+$.

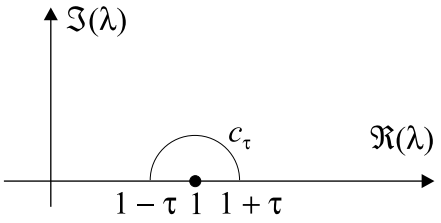


FIG. 9. *Deformed contour in the complex λ-plane.*

Proof. With the choice of the branch-cut the integrand in (77) is holomorphic in the first quadrant of the complex λ-plane. This means that we can deform the integration path as indicated in Figure 9. The part of the integration on the real λ axis from $\lambda = 1 - \tau$ to $\lambda = 1 + \tau$, with $0 < \tau < 1$, is replaced by a semicircle c_τ, as shown in Figure 9. Hence we have

$$\left|\int_0^{+\infty} \ldots d\lambda\right| \leq \left|\int_0^{1-\tau} \ldots d\lambda\right| + \left|\int_{c_\tau} \ldots d\lambda\right| + \left|\int_{1+\tau}^{+\infty} \ldots d\lambda\right|. \tag{78}$$

Using Lemma A.1, $|\cos\lambda x| \le 1$, and $|\lambda^2 - 1| \ge -\tau^2 + 2\tau$, we find for the first term on the right-hand side that

$$\left|\int_0^{1-\tau} \dots d\lambda\right| < \int_0^{1-\tau} \frac{4e^{-2\sqrt{1-\lambda^2}\gamma\sin\alpha}}{\sqrt{-\tau^2+2\tau}} d\lambda < \frac{4(1-\tau)}{\sqrt{-\tau^2+2\tau}} e^{-2\sqrt{-\tau^2+2\tau}\gamma\sin\alpha} \tag{79}$$

if

$$\gamma > \frac{1}{\sqrt{-\tau^2+2\tau}\sin\alpha}. \tag{80}$$

The second term on the right-hand side of (78) can be written as

$$\left|\int_{c_\tau} \dots d\lambda\right| = \left|\int_0^{\pi} \frac{\tanh(\sqrt{\tau^2 e^{-2j\phi} - 2\tau e^{-j\phi}}\gamma e^{-j\alpha}) - 1}{\sqrt{\tau^2 e^{-2j\phi} - 2\tau e^{-j\phi}}} \cos[(1-\tau e^{-j\phi})x]\tau e^{-j\phi} d\phi\right|. \tag{81}$$

For $0 \le \phi \le \pi$ we have that

$$|\sqrt{\tau^2 e^{-2j\phi} - 2\tau e^{-j\phi}}| > \sqrt{-\tau^2+2\tau}, \tag{82}$$

that

$$\Re(\sqrt{\tau^2 e^{-2j\phi} - 2\tau e^{-j\phi}} e^{-j\alpha}) > \sigma = \min[\sqrt{-2\tau+\tau^2}\sin\alpha, \sqrt{\tau^2+2\tau}\cos\alpha] > 0, \tag{83}$$

and also that

$$|\cos[(1-\tau e^{-j\phi})x]| < 2e^{\tau|x|}. \tag{84}$$

Using Lemma A.1, this all allows us to write for the second term on the right-hand side of (78) that

$$\left|\int_{c_\tau} \dots d\lambda\right| < \frac{8\pi\tau}{\sqrt{-\tau^2+2\tau}} e^{-2\sigma\gamma+\tau|x|} \tag{85}$$

if

$$\gamma > \frac{1}{\sigma}. \tag{86}$$

Using Lemma A.1, $|\cos\lambda x| \le 1$, and $|\lambda^2 - 1| \ge \tau^2 + 2\tau$, we find for the third term on the right-hand side of (78) that

$$\left|\int_{1+\tau}^{+\infty} \dots d\lambda\right| < \int_{1+\tau}^{+\infty} \frac{4e^{-2\sqrt{\lambda^2-1}\gamma\cos\alpha}}{\sqrt{\tau^2+2\tau}} d\lambda \tag{87}$$

if

$$\gamma > \frac{1}{\sqrt{\tau^2+2\tau}\cos\alpha}. \tag{88}$$

If we now restrict τ further such that $(2-\sqrt{3})/\sqrt{3} < \tau < 1$, then $\lambda/2 < \sqrt{\lambda^2-1}$ for $\lambda > 1+\tau$. Now we find

$$\int_{1+\tau}^{+\infty} \frac{4e^{-2\sqrt{\lambda^2-1}\gamma\cos\alpha}}{\sqrt{\tau^2+2\tau}} d\lambda < \frac{4}{\sqrt{\tau^2+2\tau}} \int_{1+\tau}^{+\infty} e^{-\lambda\gamma\cos\alpha} d\lambda = \frac{4e^{-(1+\tau)\gamma\cos\alpha}}{\sqrt{\tau^2+2\tau}\gamma\cos\alpha}. \tag{89}$$

Taking (79), (85), and (89) together allows us to write that

$$\left|\int_0^{+\infty} \dots d\lambda\right| < f(\gamma), \tag{90}$$

with

$$f(\gamma)=\frac{4(1-\tau)}{\sqrt{-\tau^2+2\tau}}e^{-2\sqrt{-\tau^2+2\tau}\gamma\sin\alpha}+\frac{8\pi\tau}{\sqrt{-\tau^2+2\tau}}e^{-2\sigma\gamma+\tau|x|}+\frac{4e^{-(1+\tau)\gamma\cos\alpha}}{\sqrt{\tau^2+2\tau}\gamma\cos\alpha} \tag{91}$$

if

$$\gamma > \max\left[\frac{1}{\sqrt{-\tau^2+2\tau}\sin\alpha}, \frac{1}{\sqrt{\tau^2+2\tau}\cos\alpha}\right]. \tag{92}$$

Note that the condition (92) also implies that $\gamma > 1/\sigma$. Since $f(\gamma)$ is a monotonic decreasing function of γ, we have shown that

$$\left|\int_0^{+\infty} \dots d\lambda\right| < \epsilon \tag{93}$$

if

$$\gamma > \max\left[f^{-1}(\epsilon), \frac{1}{\sin\alpha\sqrt{-\tau^2+2\tau}}, \frac{1}{\cos\alpha\sqrt{\tau^2+2\tau}}\right]. \tag{94}$$

□

From the proof it follows that, for a given ϵ, γ needs to be taken larger when $|x|$ increases. γ also needs to be taken large when α comes close to 0 or $\pi/2$.

LEMMA A.3.

$$\forall\zeta\in\mathbf{C}, \qquad \xi=\Re(\zeta)>0 \Rightarrow |\cosh\zeta| < e^{\xi}. \tag{95}$$

Proof. The proof is trivial. □

THEOREM A.4. *If* $0<\alpha<\pi/2$ *and* $x,y\in\mathbf{R}$, *then*

$$\forall\epsilon>0, \quad \exists\delta(\epsilon)>0:$$

$$\gamma>\delta(\epsilon)\Rightarrow\left|\int_0^{+\infty}\frac{\tanh(\sqrt{\lambda^2-1}\gamma e^{-j\alpha})-1}{\sqrt{\lambda^2-1}}\cosh(\sqrt{\lambda^2-1}|y|)\cos(\lambda x)d\lambda\right|<\epsilon, \tag{96}$$

where the branch-cut of $\sqrt{\zeta}$ *is along* $\arg(\zeta)=\pi^+$.

Proof. This can be proven in the same manner as Theorem A.2. Let us settle for mentioning the differences.

From Lemma A.3 it follows for the first term on the right-hand side of (78) that

$$|\cosh(\sqrt{\lambda^2-1}|y|)| < e^{\sqrt{-\tau^2+2\tau}|y|}, \tag{97}$$

for the second term on the right-hand side of (78) that

$$|\cosh(\sqrt{\lambda^2-1}|y|)| < e^{\sqrt{\tau^2+2\tau}|y|}, \tag{98}$$

and for the third term on the right-hand side of (78) that

$$|\cosh(\sqrt{\lambda^2-1}|y|)| < e^{\sqrt{\lambda^2-1}|y|}, \tag{99}$$

which means that in (89) we have to replace $\gamma\cos\alpha$ by $\gamma\cos\alpha - |y|/2$. We also have to impose that

$$\gamma > \frac{|y|}{2\cos\alpha}, \tag{100}$$

in order to assure that the integrals in (89) remain convergent.

Taking all this together, the function $f(\gamma)$ now becomes

$$f(\gamma) = \frac{4(1-\tau)}{\sqrt{-\tau^2+2\tau}} e^{-\sqrt{-\tau^2+2\tau}(2\gamma\sin\alpha - |y|)} + \frac{8\pi\tau}{\sqrt{-\tau^2+2\tau}} e^{-2\sigma\gamma+\tau|x|+\sqrt{\tau^2+2\tau}|y|}$$

$$+ \frac{4e^{-(1+\tau)(\gamma\cos\alpha-|y|/2)}}{\sqrt{\tau^2+2\tau}(\gamma\cos\alpha - |y|/2)}, \tag{101}$$

which still is a monotonic decreasing function of γ. Hence, the proof is complete by adding the condition (100) to (94). □

From the proof it now also follows that, for a given ϵ, γ needs to be taken larger when $|y|$ increases.

THEOREM A.5. *If $0 < \alpha < \pi/2$, if $F(\lambda)$ is holomorphic along the contour of Figure 9, and if $\int_0^{+\infty} F(\lambda)\cos(\lambda x)d\lambda$ exists for all $x > 0$, then*

$$\forall\epsilon > 0, \qquad \exists\delta(\epsilon) > 0 : \gamma > \delta(\epsilon) \Rightarrow \left| \int_0^{+\infty} F(\lambda) e^{-\sqrt{\lambda^2-1}2d} \cos(\lambda x) d\lambda \right| < \epsilon, \tag{102}$$

where the branch-cut of $\sqrt{\zeta}$ is along $\arg(\zeta) = \pi^+$.

Proof. Let us deform the integration path and split the integration as in Theorem A.2. The first contribution is

$$\left| \int_0^{1-\tau} \ldots d\lambda \right| = \left| \int_0^{1-\tau} F(\lambda) e^{j\sqrt{1-\lambda^2}2\gamma\cos\alpha} e^{-\sqrt{1-\lambda^2}2\gamma\sin\alpha} \cos(\lambda x) d\lambda \right|$$

$$< e^{-\sqrt{2\tau-\tau^2}2\gamma\sin\alpha} \int_0^{1-\tau} |F(\lambda)\cos(\lambda x)|\, d\lambda. \tag{103}$$

Using the definition of σ from Theorem A.2, the second contribution obeys

$$\left| \int_{c_\tau} \ldots d\lambda \right| < \int_{c_\tau} |F(\lambda)\cos(\lambda x)|\, e^{-\sigma\gamma} d\lambda$$

$$< e^{-\sigma\gamma} \int_{c_\tau} |F(\lambda)\cos(\lambda x)|\, d\lambda. \tag{104}$$

For the third term we have

$$\left| \int_{1+\tau}^{+\infty} \ldots d\lambda \right| < \int_{1+\tau}^{+\infty} |F(\lambda)| e^{-\sqrt{\lambda^2-1}2\gamma\cos\alpha} d\lambda. \tag{105}$$

Since $\int_0^{+\infty} F(\lambda)\cos(\lambda x)d\lambda$ exists for all $x > 0$, the limit $\lim_{\lambda\to+\infty} F(\lambda)$ exists. Because $F(\lambda)$ is holomorphic, there exists a number M such that $|F(\lambda)| < M$ for $\lambda \geq 1 + \tau$. Hence, we can proceed as in (89) and write

$$\left|\int_{1+\tau}^{+\infty} \dots d\lambda\right| < \int_{1+\tau}^{+\infty} |F(\lambda)| e^{-\sqrt{\lambda^2-1}2\gamma\cos\alpha} d\lambda = \frac{M e^{-(1+\tau)\gamma\cos\alpha}}{\gamma\cos\alpha} \tag{106}$$

when $(2-\sqrt{3})/\sqrt{3} < \tau < 1$.

The three parts of the integration are, as in previous theorems, again bounded by an exponential decreasing function of γ. This completes the proof. Note that the proof remains valid if $F(\lambda)$ has simple poles along the real axis. These poles can be avoided in a similar way as the branch-point. □

LEMMA A.6.

$$\forall\zeta \in \mathbf{C}, \qquad |J_0(\zeta)| \leq e^{|\zeta|}. \tag{107}$$

Proof. Because

$$e^{|\zeta|} = \left[e^{|\zeta|/2}\right]^2 = \left[\sum_{n=0}^{+\infty} \frac{|\zeta|^n}{n!2^n}\right]^2 \geq \sum_{n=0}^{+\infty} \frac{|\zeta|^{2n}}{(n!)^2 2^{2n}}, \tag{108}$$

we prove that

$$|J_0(\zeta)| = \left|\sum_{n=0}^{+\infty} \frac{(-1)^n\zeta^{2n}}{(n!)^2 2^{2n}}\right| \leq \sum_{n=0}^{+\infty} \frac{|\zeta|^{2n}}{(n!)^2 2^{2n}} \leq e^{|\zeta|}. \tag{109}$$ □

THEOREM A.7. *If* $0 < \alpha < \pi/2$, $z \in \mathbf{R}$, *and* $\rho > 0$, *then*

$$\forall\epsilon > 0, \qquad \exists\delta(\epsilon) > 0 :$$

$$\gamma > \delta(\epsilon) \Rightarrow \left|\int_0^{+\infty} \frac{\tanh(\sqrt{\lambda^2-1}\gamma e^{-j\alpha}) - 1}{\sqrt{\lambda^2-1}} \cosh(\sqrt{\lambda^2-1}|z|) J_0(\lambda\rho)\lambda d\lambda\right| < \epsilon, \tag{110}$$

where the branch-cut of $\sqrt{\zeta}$ *is along* $\arg(\zeta) = \pi^+$.

Proof. This can be proven in the same manner as Theorems A.2 and A.4 by replacing $\cos(\lambda x)$ by $\lambda J_0(\lambda\rho)$ and y by z. Let us simply mention the changes.

Using the fact that $|J_0(\lambda\rho)| \leq 1$, we now find that first term on the right-hand side of (78) as

$$\left|\int_0^{1-\tau} \dots d\lambda\right| < \frac{4(1-\tau)^2}{\sqrt{-\tau^2+2\tau}} e^{-\sqrt{-\tau^2+2\tau}(2\gamma\sin\alpha - |\rho|)}. \tag{111}$$

Along the second part of the integration c_τ it follows from Lemma A.6 that

$$|J_0[(1-\tau e^{j\phi})\rho]| < e^{|1-\tau e^{j\phi}||\rho|} \leq e^{(1+\tau)|\rho|}. \tag{112}$$

This allows us to bound the second term in (78) as

$$\left|\int_{c_\tau} \dots d\lambda\right| < \frac{4\pi\tau^2}{\sqrt{-\tau^2+2\tau}} e^{-2\sigma\gamma + (1+\tau)|\rho| + \sqrt{\tau^2+2\tau}|z|}. \tag{113}$$

For the third term in (78) we can again use the fact that $|J_0(\lambda\rho)| \leq 1$, resulting in

$$\left|\int_{1+\tau}^{+\infty} \dots d\lambda\right| < \frac{4}{\sqrt{\tau^2+2\tau}} \int_{1+\tau}^{+\infty} e^{-\lambda(\gamma\cos\alpha - |z|/2)} \lambda d\lambda$$

$$= \frac{4[(\gamma\cos\alpha - |z|/2)(1+\tau)+1]e^{-(1+\tau)(\gamma\cos\alpha-|z|/2)}}{\sqrt{\tau^2+2\tau}(\gamma\cos\alpha - |z|/2)^2}. \tag{114}$$

Summing the results (111), (113), and (114) results in a new function $f(\gamma)$ that is still monotonic decreasing. This completes the proof. □

Appendix B. Asymptotic expressions. In this appendix we derive asymptotic expressions for the solutions of the equation

$$\kappa_2(\lambda)\sin[\kappa_1(\lambda)d_1]\cos[\kappa_2(\lambda)d_2] + \kappa_1(\lambda)\cos[\kappa_1(\lambda)d_1]\sin[\kappa_2(\lambda)d_2] = 0 \tag{115}$$

for large values of $|\lambda|$. In (115), $\kappa_i = \sqrt{k_i^2 - \lambda^2}$, $i = 1, 2$, and we assume that $k_1 > k_2$. Here also d_1 and d_2 can be complex and satisfy $0 \geq \arg(d_1) > \arg(d_2) > -\pi/2$. In this appendix the argument function arg has its branch-cut along the negative real axis. Since (115) is a quadratic equation in λ, we restrict the analysis to $\Im(\lambda) < 0$. For large values of $|\lambda|$ one can approximate

$$\kappa_i \approx j\lambda. \tag{116}$$

This allows us to write (115) as

$$\kappa_2(\lambda)\sinh(\lambda d_1)\cosh(\lambda d_2) + \kappa_1(\lambda)\cosh(\lambda d_1)\sinh(\lambda d_2) = 0 \tag{117}$$

for large values of $|\lambda|$. It will become clear further on why we did not approximate κ_i in front of the hyperbolic functions. If $\lim_{|\lambda|\to+\infty} \arg(\lambda d_i) \neq -\pi/2$, then

$$|\sinh(\lambda d_i)| \approx |\cosh(\lambda d_i)| \approx \frac{e^{|\Re(\lambda d_i)|}}{2}. \tag{118}$$

If $\lim_{|\lambda|\to+\infty} \arg(\lambda d_i) = -\pi/2$, then

$$\lim_{|\lambda|\to+\infty} \frac{\Re(\lambda d_i)}{\Im(\lambda d_i)} = 0, \tag{119}$$

but not necessarily $\lim_{|\lambda|\to+\infty} \Re(\lambda d_i) = 0$. Note that $\lim_{|\lambda|\to+\infty} \arg(\lambda d_i) = \pi/2$ is impossible with the argument constraints on d_i and λ. We have to distinguish three different cases.

In the first case we assume that

$$\lim_{|\lambda|\to+\infty} \arg(\lambda d_1) = -\frac{\pi}{2}. \tag{120}$$

This means that

$$\lim_{|\lambda|\to+\infty} \Re(\lambda d_2) = -\infty. \tag{121}$$

Hence, for large $|\lambda|$ we can approximate

$$-\sinh(\lambda d_2) \approx \cosh(\lambda d_2) \approx \frac{e^{-\lambda d_2}}{2} \tag{122}$$

and reduce (117) to

$$\kappa_2(\lambda)\sinh(\lambda d_1) - \kappa_1(\lambda)\cosh(\lambda d_1) = 0 \tag{123}$$

or

$$e^{2\lambda d_1} = \frac{\kappa_2(\lambda) + \kappa_1(\lambda)}{\kappa_2(\lambda) - \kappa_1(\lambda)}. \tag{124}$$

In the denominator on the right-hand side we cannot use the approximation (116), but we have to take an extra term into account, i.e.,

$$\kappa_i(\lambda) \approx j\lambda\left(1 - \frac{k_i^2}{2\lambda^2}\right). \tag{125}$$

This allows us to recast (124) as

$$e^{p_1} = \pm\frac{p_1}{\sqrt{k_1^2 - k_2^2 d_1}}, \tag{126}$$

with $p_1 = \lambda d_1$. From (126) it follows that

$$e^{\Re(p_1)} = \frac{2|p_1|}{\sqrt{k_1^2 - k_2^2|d_1|}}. \tag{127}$$

In view of (119) we can approximate $|p_1| \approx |\Im(p_1)|$ such that

$$\Re(p_1) = \log\left[\frac{2|\Im(p_1)|}{\sqrt{k_1^2 - k_2^2|d_1|}}\right], \tag{128}$$

and also

$$e^{j\Im(p_1)} = \pm\frac{p_1|d_1|}{d_1|\Im(p_1)|} \approx \pm\frac{\Im(p_1)|d_1|}{d_1|\Im(p_1)|}, \tag{129}$$

which gives

$$\Im(p_{1,i}) = -\arg(d_1) - (2i+1)\frac{\pi}{2}. \tag{130}$$

Hence, asymptotically the equation has a first infinite set of solutions of the form

$$\lambda_i = \frac{1}{d_1}\log\frac{2[\arg d_1 + (2i+1)\frac{\pi}{2}]}{\sqrt{k_1^2 - k_2^2|d_1|}} - \frac{j[\arg d_1 + (2i+1)\frac{\pi}{2}]}{d_1}. \tag{131}$$

In the second case we assume that

$$\lim_{|\lambda|\to+\infty}\arg(\lambda d_2) = -\frac{\pi}{2}. \tag{132}$$

This means that

$$\lim_{|\lambda|\to+\infty}\Re(\lambda d_1) = +\infty, \tag{133}$$

resulting in the approximate equation

$$\kappa_2(\lambda)\cosh(\lambda d_2) + \kappa_1(\lambda)\cosh(\lambda d_2) = 0, \tag{134}$$

or, proceeding as above,

$$e^{p_2} = \pm j\frac{\sqrt{k_1^2 - k_2^2}d_2}{p_2}, \tag{135}$$

with $p_2 = \lambda d_2$. This equation can be solved in the same way as (127). The result is

$$\lambda_i = -\frac{1}{d_2}\log\left[\frac{2(i\pi - \arg d_2)}{\sqrt{k_1^2 - k_2^2}|d_2|}\right] - \frac{j(i\pi - \arg d_2)}{d_2}. \tag{136}$$

In the remaining case neither (120) nor (132) is valid. In this case (117) reduces to

$$[\kappa_1(\lambda) \pm \kappa_2(\lambda)]e^{2[|\Re(\lambda d_1)|+|\Re(\lambda d_2)|]} = 0, \tag{137}$$

which has no solutions.

In [23] a similar analysis was given. However, (123) and (134) were derived on physical grounds without proof of the nonexistence of other possible solutions. The solutions (131) and (136) were also derived indirectly as asymptotic approximations of a special function, and the initial conditions were not taken as general as considered here.

REFERENCES

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

[2] Z. S. Zacks, D. M. Kingsland, R. Lee, and J. F. Lee, *A perfectly matched anisotropic absorber for use as an absorbing boundary condition*, IEEE Trans. Antennas and Propagation, 43 (1995), pp. 1460–1463.

[3] W. C. Chew, J. M. Jin, and E. Michielssen, *Complex coordinate system as a generalized absorbing boundary condition*, in Proceedings of the 13th Annual Review of Progress in Applied Computational Electromagnetics, Monterey, CA, 1997, Vol. 2, pp. 909–914.

[4] H. Derudder, D. De Zutter, and F. Olyslager, *Analysis of waveguide discontinuities using perfectly matched layers*, Electron. Lett., 34 (1998), pp. 2138–2140.

[5] H. Derudder, F. Olyslager, D. De Zutter, and S. Van den Berghe, *Efficient mode-matching analysis of discontinuities in finite planar substrates using perfectly matched layers*, IEEE Trans. Antennas and Propagation, 49 (2001), pp. 185–195.

[6] P. Bienstman, H. Derudder, R. Baets, F. Olyslager, and D. De Zutter, *Analysis of cylindrical waveguide discontinuities using vectorial eigenmodes and perfectly matched layers*, IEEE Trans. Microwave Theory Techn., 49 (2001), pp. 349–354.

[7] P. Bienstman and R. Baets, *Optical modelling of photonic crystals and VCSELs using eigenmode expansion and perfectly matched layers*, Optical and Quantum Electronics, 33 (2001), pp. 327–341.

[8] H. Derudder, F. Olyslager, and D. De Zutter, *An efficient series expansion for the 2D Green's function of a microstrip substrate using perfectly matched layers*, IEEE Microwave Guided Wave Lett., 9 (1999), pp. 505–507.

[9] F. Olyslager and H. Derudder, *Series representation of Green dyadics for layered media using PMLs*, IEEE Trans. Antennas and Propagation, 51 (2003), pp. 2319–2326.

[10] S. Gaal, *private communication*, 2003.

[11] F. Olyslager, *Series approximation for Green functions*, in Proceedings of the 2002 Workshop on Mathematical Modelling of Wave Phenomena, B. Nilsson, ed., Växjö University Press, Sweden, 2004, to appear.

[12] A. Sommerfeld, *Partielle Differentialgleichungen der Physik*, Dieterichsche Verlagsbuchhandlung, Wiesbaden, Germany, 1947.

[13] L. F. Knockaert and D. De Zutter, *On the completeness of eigenmodes in a parallel plate waveguide with a perfectly matched layer termination*, IEEE Trans. Antennas and Propagation, 50 (2002), pp. 1650–1653.

[14] L. B. Felsen and N. Marcuvitz, *Radiation and Scattering of Waves*, IEEE Press, New York, 1994.

[15] F. Olyslager, *Electromagnetic Waveguides and Transmission Lines*, Oxford University Press, Oxford, UK, 1999.

[16] Y. L. Chow, J. J. Yang, D. G. Fang, and G. E. Howard, *A closed-form spatial Green's function for the thick microstrip substrate*, IEEE Trans. Microwave Theory Techn., 39 (1991), pp. 588–592.

[17] B. Hu and W. C. Chew, *Fast inhomogeneous plane wave algorithm for electromagnetic solutions in layered medium structures: Two-dimensional case*, Radio Science, 35 (2000), pp. 31–43.

[18] K. A. Michalski, *Extrapolation methods for Sommerfeld integral tails*, IEEE Trans. Antennas and Propagation, 46 (1998), pp. 1405–1418.

[19] F. L. Taxeira and W. C. Chew, *Systematic derivation of anisotropic PML absorbing media in cylindrical and spherical coordinates*, IEEE Microwave Guided Wave Lett., 7 (1997), pp. 371–373.

[20] F. Olyslager and I. V. Lindell, *Field decomposition and factorization of the Helmholtz determinant operator for bianisotropic media*, IEEE Trans. Antennas and Propagation, 49 (2001), pp. 660–665.

[21] P. Joly, *Perfectly matched layers techniques: Stability and instability results*, in Proceedings of the 2002 Workshop on Mathematical Modelling of Wave Phenomena, B. Nilsson, ed., Växjö University Press, Sweden, 2004, to appear.

[22] I. V. Lindell and F. Olyslager, *Polynomial Operators and Green Functions*, Progr. Electromagnetic Res. 30, J. Kong, ed., Elsevier, New York, 2001, pp. 59–84.

[23] H. Rogier and D. De Zutter, *Berenger and leaky modes in microstrip substrates terminated by a perfectly matched layer*, IEEE Trans. Microwave Theory Techn., 49 (2001), pp. 712–715.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1434–1456

A DIFFUSIONAL-THERMAL THEORY OF NEAR-STOICHIOMETRIC PREMIXED FLAMES*

ELIANA S. ANTONIOU†, JOHN K. BECHTOLD‡, AND MOSHE MATALON§

Abstract. In this paper we present a diffusional-thermal theory of premixed flames for near-stoichiometric conditions. Our theory exhibits an explicit dependence on the equivalence ratio as well as on two distinct Lewis numbers which correspond to the fuel and the oxidizer. Normally, the deficient component in the mixture is totally depleted in the reaction zone. However, for curved or strained flames, it is possible for the initially excess reactant to be consumed at the reaction zone if it is the less mobile of the two species, while the initially deficient species leaks through. The form of the derived jump conditions for temperature and enthalpy gradients across the reaction sheet depends on which of the two species is consumed. This can have important implications on predicted flame dynamics. For example, we show that, as a result of preferential diffusion, portions of a corrugated flame may burn rich while neighboring regions burn lean. This results in leakage of fuel and oxidizer through the premixed flame which are then consumed downstream by trailing diffusion flame tongues. Furthermore, the extinction characteristics of strained flames are found to depend on whether fuel or oxidizer is ultimately depleted.

Key words. flame theory, extinction

AMS subject classifications. 80A25, 80A32

DOI. 10.1137/S003613990342428X

1. Introduction.

Many technologies employ combustion as a source of energy, for example, engines, furnaces, and jets, and important issues regarding efficiency and pollutants depend significantly on the mixture composition. The mixture composition is known to have a significant effect on flame shape and dynamics. For example, flames in light hydrocarbon mixtures are observed to propagate with a smooth surface when the mixture is rich, while the flame surface can take on a cellular appearance when the mixture is lean [1]. Similarly, a great deal of soot formation typically occurs in rich mixtures, while substantially less is produced when the mixture is lean [2].

Despite the recognized importance of mixture strength in practical combustion applications, most theoretical studies are based on single reactant models in which the chemistry is represented by a one-step overall irreversible reaction and proceeds at a rate that depends on the concentration of the deficient component in the mixture. All other reactants appear in relatively large amounts, so only a minimal amount is consumed. These models are therefore valid only when the mixture is far from stoichiometry, i.e., very lean or very rich. They don't exhibit any dependence on the mixture strength, usually measured in terms of the equivalence ratio, which is the ratio of the mass of fuel to oxidizer in the fresh mixture. Many practical combustion systems operate in a regime closer to stoichiometry, and, as noted above, this is a

*Received by the editors March 10, 2003; accepted for publication (in revised form) August 11, 2003; published electronically May 20, 2004. This work was partially supported by the National Science Foundation under grants CTS0074320 and DMS0072588.

http://www.siam.org/journals/siap/64-4/42428.html

†Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102. Current address: Department of Mathematics, William Paterson University, Wayne, NJ 07470 (AntoniouE@wpunj.edu).

‡Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102 (jobech@m.njit.edu).

§Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208 (matalon@nwu.edu).

transition regime where the composition may have an effect on flame dynamics. The objective of this paper is to examine these effects within the context of a diffusional-thermal model accounting for the entire spectrum of mixture compositions, from lean to rich, including the near-stoichiometric regime.

As noted, most earlier theories considered the mixture to be far removed from stoichiometry. As such, all results depend on a single Lewis number, the ratio of thermal diffusivity of the mixture to mass diffusivity of the deficient reactant. It has been shown, for example, that an initially smooth flame will lose stability to a cellular surface when this Lewis number is less than a critical value, slightly less than unity [3]. It has also been observed that a flame in a nonuniform flow can be extinguished by stretch when the Lewis number is greater than one but not otherwise [4, 5]. Therefore, predicted behavior in a given mixture can be quite different depending on whether conditions are lean or rich, but single-reactant models cannot describe the transition from one burning regime to another as stoichiometry is crossed. For conditions close to stoichiometry, both fuel and oxidizer can be expected to play a role in flame characteristics, and a scheme that depends on both reactants must be used. In this case, two Lewis numbers, one associated with the fuel and the other with the oxidizer, as well as the equivalence ratio are all important parameters affecting flame response.

Theoretical investigations of stoichiometric, or near-stoichiometric, flames have mostly been limited to planar or perturbed planar flames. Expressions for the flame speed and temperature, illustrating their dependence on the equivalence ratio, were given by Sen and Ludford [6] and by Mitani [7]. In the context of a diffusional-thermal model, Joulin and Mitani [8] showed that conclusions from the stability analyses of off-stoichiometric flames remain valid under near-stoichiometric conditions, with an average Lewis number replacing the one based on the deficient reactant. In that study, the initially deficient reactant is always consumed by the flame, while a small amount of the abundant species leaks through. Similar conclusions were reached by Jackson [9], who included the effect of thermal expansion in the linear stability analysis, and by Sivashinsky [10], who presented a weakly nonlinear analysis of the problem at exact stoichiometric conditions. Matalon, Cui, and Bechtold [11] have recently shown that hydrodynamic models of flames of arbitrary shape in general flows are modified in a similar way. Although their model can be used to treat nonplanar flames, the flame structure remains quasi–one-dimensional. Since these theories all consider the flame structure to be planar or nearly planar, with the Lewis numbers close to unity, the initially deficient reactant is nearly always consumed at the reaction front, and since temperature perturbations behind the flame are generally weak, the burning characteristics are only slightly modified. However, when the two species diffuse at unequal rates, that is, when the associated Lewis numbers are distinct, and when the flame structure is nonplanar, flame behavior is expected to be quite different. Recently, Bechtold and Matalon [12] derived a slowly-varying-flame model for two-reactant flames in near-stoichiometric mixtures and found that, for strained or curved flames, it is possible for a less mobile reactant that is initially in excess in the mixture to be locally deficient, and hence consumed, at the reaction zone. This can have important consequences on the flame temperature, and hence the burning characteristics. The asymptotic structure of premixed flames has also been analyzed for moderately rich methane-air flames using a four-step kinetic scheme [13]; these authors also find that the response of strained flames depends on the Lewis numbers of both fuel and oxygen.

Still lacking is a diffusional-thermal model of near-stoichiometric flames; the main goal of this paper is to fill that void. The model is derived in a formal asymptotic

way from the full system of equations governing premixed flames in a two-reactant mixture. The analysis is carried out considering the limit of large activation energy, which enables an analytical resolution of the nonlinear reaction rate terms, and the assumption of near-unity Lewis numbers, which is required to obtain a consistent closed model. The model differs from the corresponding single-reactant model in two significant ways. First, it consists of two equations governing the transport of the two reactants, and second, the form of the derived jump conditions for the gradients across the reaction sheet depends on which of the two species is ultimately consumed by the reaction. This can lead to quantitative and qualitative differences in flame structure and in predicted dynamic behaviors. We show that it is possible along a corrugated flame to have regions burning fuel-lean while neighboring regions burn fuel-rich. The excess oxidizer leaking through the first region and fuel leaking through the other both burn along diffusion flame tongues that trail behind the premixed flame. For strained flames we provide a description of the extinction characteristics, showing that the flame response depends significantly on which of the two species is ultimately consumed at the reaction zone.

2. Formulation. Consider a premixed combustible mixture, consisting of an excess (E) and deficient (D) reactant, in which reaction proceeds according to

$$\nu_E \mathcal{M}_E + \nu_D \mathcal{M}_D \rightarrow \text{Products},$$

where $\mathcal{M}_i$ are the chemical symbols for species i, and ν_i are the stoichiometric coefficients. An important parameter in this analysis is the equivalence ratio

$$\phi = \frac{Y_F/\nu_F W_F}{Y_O/\nu_O W_O},$$

which is the ratio of the mass fraction of the fuel to oxidizer in the fresh mixture to their stoichiometric ratio. Here Y_i denote the mass fractions of species i and W_i their molecular weights. Values of ϕ larger than unity correspond to fuel-rich mixtures, and values of ϕ less than one correspond to lean mixtures. To avoid discussing lean and rich mixtures separately it is convenient to introduce the parameter

$$\Phi = \frac{Y_E/\nu_E W_E}{Y_D/\nu_D W_D},$$

based on the ratio of the mass fractions of excess to deficient reactants. As defined, Φ is always greater than one; it is equal to ϕ for rich mixtures and $1/\phi$ for lean mixtures.

The governing equations are made dimensionless by using the adiabatic flame speed, S_f^0, and thermal thickness, $l_D = \lambda/\hat{\rho} c_p S_f^0$, as the characteristic velocity and length, respectively, where λ is the thermal conductivity, c_p the specific heat, and $\hat{\rho}$ the density of the fresh mixture. The time scale is chosen to be l_D/S_f^0. All other variables are made dimensionless by their values in the fresh mixture, which are denoted by the subscript u. The governing equations for temperature and species mass fractions are

$$\frac{DT}{Dt} - \nabla^2 T = \frac{q}{Y_{D,u}}\Omega, \tag{2.1}$$

$$\frac{DY_D}{Dt} - Le_D^{-1}\nabla^2 Y_D = -\Omega, \tag{2.2}$$

$$\frac{DY_E}{Dt} - Le_E^{-1}\nabla^2 Y_E = -\nu\Omega. \tag{2.3}$$

These are coupled to the equations of hydrodynamics, which we avoid writing here since our objective is a diffusional-thermal model. The two sets of equations decouple by considering weak thermal expansion. Here $D/Dt = \partial/\partial t + \mathbf{V}\cdot\nabla$ is the convective derivative with $\mathbf{V}$ the prescribed velocity field. The parameters appearing in these equations include the Lewis numbers $Le_i = \lambda/\hat{\rho}c_p\mathcal{D}_i$, the heat release $q = QY_{D,u}/c_p\hat{T}_u\nu_D W_D$, and the mass-weighted stoichiometric coefficient ratio $\nu = \nu_E W_E/\nu_D W_D$. The reaction rate term on the right-hand side of (2.1)–(2.3) has the form

$$\Omega = DY_E Y_D \exp\left\{\frac{\beta T_a^2}{q}\left(\frac{1}{T_a} - \frac{1}{T}\right)\right\}, \tag{2.4}$$

where D is the Damköhler number, a ratio of the flow time to the chemical time, given by

$$D = \frac{\lambda}{c_p(S_f^0)^2}\frac{\nu_D B}{W_E}e^{-\frac{\beta T_a}{q}}. \tag{2.5}$$

Here $\beta = E(\hat{T}_a - \hat{T}_u)/R\hat{T}_a^2$ is the Zeldovich number, and $\hat{T}_a$ is the adiabatic flame temperature, which in dimensionless form is expressed as $T_a = 1 + q$. In the limit of large activation energy and considering near unity Lewis numbers and near-stoichiometric conditions,

$$Le_i^{-1} = 1 - \beta^{-1}l_i, \qquad \Phi = 1 + \beta^{-1}\varphi, \tag{2.6}$$

the flame speed is given by

$$S_f^0 = \left[\frac{2\lambda\nu_D B}{c_p W_E}\beta^{-3}(\varphi+2)\nu Y_{D,u}\right]^{1/2} e^{-\beta T_a/2q}, \tag{2.7}$$

and thus the Damköhler number can be written $D = \beta^3/[2(\varphi+2)\nu Y_{D,u}]$.

It is convenient in our formulation to introduce the enthalpy functions

$$H_D = T + q\frac{Y_D}{Y_{D,u}}, \qquad H_E = T + \frac{q}{\nu}\frac{Y_E}{Y_{D,u}}. \tag{2.8}$$

Temperature gradients behind the flame are assumed small, i.e., $O(\beta^{-1})$, and correspondingly only an $O(\beta^{-1})$ amount of reactants can leak through. Thus, the enthalpy variables can be expanded as

$$H_D = 1 + q + \beta^{-1}qh_D + \cdots, \qquad H_E = 1 + q\Phi + \beta^{-1}qh_E + \cdots, \tag{2.9}$$

and equations for the enthalpy perturbations, h_i, which follow from (2.1)–(2.3), are given by

$$\frac{Dh_D}{Dt} - \nabla^2 h_D = \frac{l_D}{q}\nabla^2 T, \tag{2.10}$$

$$\frac{Dh_E}{Dt} - \nabla^2 h_E = \frac{l_E}{q}\nabla^2 T. \tag{2.11}$$

2.1. Reaction zone analysis. In the asymptotic limit $\beta \to \infty$, the reaction rate is negligible everywhere except where $T \sim T_a$, i.e., along the reaction sheet given by $x = f(y,z,t)$ (to fix ideas, we will let $x > f(y,z,t)$ correspond to the burned region). We expand all variables, on either side of the sheet, in the form

$$T = T^{(0)} + \beta^{-1} T^{(1)} + \cdots.$$

To leading order, the temperature is continuous across the sheet, while its gradient suffers a jump which is determined from the internal structure of the reaction zone. Integration of (2.10) and (2.11) determines that, to leading order, the two enthalpy variables are also continuous across the flame sheet, and that

$$h_D^* = \frac{1}{q} T^{(1)} + \frac{Y_D^{(1)}}{Y_{D,u}}, \tag{2.12}$$

$$h_E^* = \frac{1}{q} T^{(1)} + \frac{Y_E^{(1)}}{\nu Y_{D,u}} - \varphi. \tag{2.13}$$

Here the right-hand side is to be evaluated at $n = 0^+$ or $n = 0^-$, where n denotes the distance normal to the sheet. In the following, we will evaluate these quantities at $n = 0^+$ unless otherwise noted. Stretching the normal coordinate, integrating the temperature equation, and then matching with the solution outside the sheet yields

$$\left.\frac{\partial T^{(0)}}{\partial n}\right|_{n=0^-} = \frac{q}{\sqrt{2+\varphi}} \sqrt{2 + \left(Y_D^{(1)} + \frac{1}{\nu} Y_E^{(1)}\right)/Y_{D,u}}\ e^{T_b^{(1)}/2q}, \tag{2.14}$$

where $T_b^{(1)} = T^{(1)}$ evaluated at $n = 0^+$. This condition can also be written as

$$\left.\frac{\partial T^{(0)}}{\partial n}\right|_{n=0^-} = \frac{q}{\sqrt{2+\varphi}} \sqrt{2 + Y_{LE}/Y_{D,u}}\ e^{T_b^{(1)}/2q}, \tag{2.15}$$

where Y_{LE} denotes the reactant which is locally in excess. For a planar flame, the initially deficient reactant is always consumed at the reaction sheet, i.e., $Y_D^{(1)} = 0$, and thus $Y_{LE} = Y_E^{(1)}/\nu$. In general, however, it is possible, due to disparate diffusivities, for the initially excess species to be locally deficient, and hence, totally consumed by the reaction. In this case, a small amount of the initially deficient reactant leaks through and $Y_{LE} = Y_D^{(1)}$. The condition (2.15) on the temperature gradient depends on the temperature perturbation behind the flame and the amount of reactant that leaks through the sheet. Both of these quantities can be expressed in terms of the enthalpy perturbations.

To determine which species is consumed and which leaks through, we subtract (2.12) from (2.13), which gives

$$\left(\frac{1}{\nu} Y_E^{(1)} - Y_D^{(1)}\right)/Y_{D,u} = \varphi + h_E^* - h_D^*. \tag{2.16}$$

Recognizing that the mass fractions cannot be negative, the sign of the right-hand side of (2.16) provides the necessary information. If the right-hand side is positive, then $Y_D^{(1)} = 0$, $T_b^{(1)}/q = h_D^*$, and (2.16) determines the amount of initially excess

reactant that leaks through. On the other hand, if the right-hand side of (2.16) is negative, then $Y_E^{(1)} = 0$, $T_b^{(1)}/q = h_E^* + \varphi$, and (2.16) determines the amount of initially deficient reactant that leaks through. When these are inserted into (2.15), we obtain

$$\left.\frac{\partial T^{(0)}}{\partial n}\right|_{n=0^-} = q\sqrt{\frac{2+|\varphi+h_E^*-h_D^*|}{2+\varphi}} \times \exp\left\{\frac{\varphi+h_E^*+h_D^*-|\varphi+h_E^*-h_D^*|}{4}\right\}. \tag{2.17}$$

To summarize, in the limit $\beta \to \infty$, the problem reduces to solving the reaction-free equation for the leading order temperature, $T^{(0)}$, together with the equations for the enthalpies, h_E and h_D, on either side of the reaction sheet. Solutions must satisfy the derived jump conditions as well as appropriate initial and boundary conditions. Our objective is a diffusional-thermal model, accounting for the entire spectrum of mixture composition, that is free of hydrodynamic disturbances. It is therefore commonplace to consider weak thermal expansion ($q \ll 1$), formally done by expanding the temperature as $T^{(0)} = 1 + q\tau + \cdots$, to obtain

$$\frac{D\tau}{Dt} - \nabla^2\tau = 0, \quad x < f(y,z,t), \tag{2.18}$$

$$\tau = 1, \quad x > f(y,z,t), \tag{2.19}$$

$$\frac{Dh_i}{Dt} - \nabla^2 h_i = l_i\nabla^2\tau, \quad x \neq f(y,x,t). \tag{2.20}$$

The jump conditions at the reaction sheet $x = f(y,x,t)$ are

$$[\tau] = 0, \quad [h_i] = 0, \tag{2.21}$$

$$\left[\frac{\partial h_i}{\partial n}\right] + l_i\left[\frac{\partial \tau}{\partial n}\right] = 0, \tag{2.22}$$

$$\left[\frac{\partial \tau}{\partial n}\right] = -\sqrt{\frac{2+|\varphi+h_E^*-h_D^*|}{2+\varphi}}\ \exp\left\{\frac{\varphi+h_E^*+h_D^*-|\varphi+h_E^*-h_D^*|}{4}\right\}, \tag{2.23}$$

where we have used the notation $[\tau] = \tau(\text{burned}) - \tau(\text{unburned})$.

The above system is seen to differ from the single-reactant theory [14] in several ways. First, two species are accounted for, and thus the model has an explicit dependence on both Lewis numbers as well as the equivalence ratio. Second, the jump in the temperature gradient is sensitive not only to the temperature perturbation behind the flame but also to the local species concentrations. Note that for off-stoichiometric mixtures, $\varphi \to \infty$, the inequality $\varphi + h_E^* - h_D^* > 0$ is always satisfied, and the right-hand side of (2.23) reduces to $-e^{h_D^*/2}$, in agreement with single-reactant theory. Our model appropriately describes the flame behavior for conditions far removed from stoichiometry as well, and thus spans the whole range from lean to rich conditions.

3. Weakly corrugated flames. In this section we examine the spontaneous self-corrugation of a planar near-stoichiometric flame using linear and weakly nonlinear stability results. Although in general the flame behaves similarly to that in a mixture far from stoichiometry, there are some differences in the details, as we shall see.

The stability of a planar flame in a near-stoichiometric mixture has been studied previously by Joulin and Mitani [8] and Sivashinsky [10], and their results are readily obtained from our model. When small disturbances are introduced and the linearized equations are solved for the perturbed variables, the following dispersion relation is found for the growth rate ω:

$$64\omega^3 + \omega^2(192k^2 + 32 + 8l - l^2) + 2\omega(12k^2 + 1)(8k^2 + l + 2) + k^2(8k^2 + l + 2)^2 = 0, \tag{3.1}$$

where k is the wavenumber and l is an effective (reduced) Lewis number defined by

$$l = \frac{l_E + l_D + l_D\varphi}{2 + \varphi}. \tag{3.2}$$

This is identical to the result of the single-reactant theory with l appearing in the place of l_D, and, in fact, $l \to l_D$ as $\varphi \to \infty$.

When conditions are stoichiometric or very close to stoichiometric, the perturbed flame may burn rich along some portions and lean along others, depending on the relative magnitudes of the enthalpy perturbations. From the linear theory, the difference in the enthalpy perturbations on the burned side of the flame is found to be

$$h_E^* - h_D^* = -\frac{A(l_E - l_D)}{2(1 + 4k^2)}[1 - \sqrt{1 + 4(\omega + k^2)}]e^{iky+\omega t}, \tag{3.3}$$

where A is the amplitude of the disturbance. Because of the sinusoidal nature of the perturbations, the difference in enthalpy variables, $h_E^* - h_D^*$, changes sign along the front provided there exists preferential diffusion ($l_E \neq l_D$). Thus both forms of the jump relation for the temperature gradient (2.23) are applied along different segments of the corrugated flame. The two conditions, when linearized, are equivalent, and therefore the stability characteristics are not modified despite the local differences in mixture composition along the front. However, the structure of the resulting corrugated front beyond the instability threshold may be affected by these differences.

A weakly nonlinear analysis can be performed for $l + 2 \ll 1$. Upon introducing the appropriate scalings, we find that the species that leaks through is determined by the sign of the quantity

$$\mathcal{E} = \varphi + (l_D - l_E)\frac{\partial^2 f}{\partial y^2}, \tag{3.4}$$

where y is the transverse coordinate. Although the concentrations along the corrugated front vary in concave/convex segments, the overall flame dynamics remain governed by the Kuramoto–Sivashinsky equation as envisaged by Sivashinsky [10] under exact stoichiometric conditions. Thus, the curvature of the front plays a role in determining which species leaks through. However, the jump conditions remain identical to all orders considered, and the Kuramoto–Sivashinsky equation is found to govern the flame dynamics regardless of whether a particular region burns rich or lean.

3.1. Trailing diffusion flames. The excess of fuel and oxidizer leaking along neighboring regions of the premixed flame front burns in trailing diffusion flame tongues attached at the stoichiometric points. Here we analyze the diffusion flame structure. The system we investigate is reminiscent of the classical Burke–Schumann flame [15], but with several important differences. First, the inlet boundary is a *corrugated* premixed flame as opposed to a burner rim. Second, the distribution of species on either side of the diffusion flame is not uniform. Third, the temperature is everywhere within $O(\beta^{-1})$ of its adiabatic value and species concentrations are small $(O(\beta^{-1}))$. Finally, the Damköhler number is not a free controlled parameter; its value is fixed by the presence of the premixed flame.

To analyze the region behind the premixed flame, we introduce the expansions

$$T = T_a + \beta^{-1} q\theta, \qquad Y_i = \beta^{-1} Z_i,$$

which are inserted into the governing equations (2.1)–(2.3) to yield the system

$$\frac{\partial \theta}{\partial x} - \frac{\partial^2 \theta}{\partial x^2} - \frac{\partial^2 \theta}{\partial y^2} = \frac{\beta^2}{2(\varphi+2)\nu Y_{D,u}^2} Z_D Z_E \, e^{\theta}, \tag{3.5}$$

$$\frac{\partial Z_D}{\partial x} - \frac{\partial^2 Z_D}{\partial x^2} - \frac{\partial^2 Z_D}{\partial y^2} = -\frac{\beta^2}{2(\varphi+2)\nu Y_{D,u}} Z_D Z_E \, e^{\theta}, \tag{3.6}$$

$$\frac{\partial Z_E}{\partial x} - \frac{\partial^2 Z_E}{\partial x^2} - \frac{\partial^2 Z_E}{\partial y^2} = -\frac{\beta^2 \nu}{2(\varphi+2)\nu Y_{D,u}} Z_D Z_E \, e^{\theta}. \tag{3.7}$$

We assume a steady corrugated premixed flame front at $x = A\cos(ky)$, where A, k are the amplitude and wavenumber, respectively. From the linear analysis the temperature and concentration perturbations along the premixed flame front are given by

$$\begin{aligned} &\theta = l_D B \cos(ky), \quad Z_D = 0, \\ &Z_E = \nu Y_{D,u}[(l_E - l_D) B \cos(ky) + \varphi] \qquad \text{for } \mathcal{E} > 0, \end{aligned} \tag{3.8}$$

and

$$\begin{aligned} &\theta = l_E B \cos(ky) + \varphi, \quad Z_E = 0, \\ &Z_D = -Y_{D,u}[(l_E - l_D) B \cos(ky) + \varphi] \qquad \text{for } \mathcal{E} < 0, \end{aligned} \tag{3.9}$$

where $B = A[\sqrt{1+4k^2}-1]/2(1+4k^2)$. These are the appropriate boundary conditions to be applied for the determination of the solution in the burned region.

We rescale $y \to y\pi/k$ and introduce the variable $W = \frac{1}{\nu} Z_E - Z_D$, which satisfies the system

$$\frac{dW}{dx} - \frac{d^2 W}{dx^2} - \frac{k^2}{\pi^2} \frac{d^2 W}{dy^2} = 0, \tag{3.10}$$

$$W = Y_{D,u}[(l_E - l_D) B \cos(\pi y) + \varphi] \quad \text{at } x = A\cos(\pi y). \tag{3.11}$$

Recall that the amplitude of the perturbation, A, has been assumed small, and thus we seek solutions in the form of a power series in A. We rescale $B \to AB$, $\varphi \to A\varphi$, $W \to AW$ and at leading order we find

$$W = Y_{D,u}\varphi + Y_{D,u}(l_E - l_D)Be^{-rx}\cos\pi y, \tag{3.12}$$

where $r = \frac{1}{2}[\sqrt{1+4k^2}-1]$. We now restrict our attention to $0 < y < 1$ and determine the diffusion flame sheet location by setting $W = 0$, which yields the expression

$$y = \frac{1}{\pi}\arccos\left(-\frac{a_0}{a_1}e^{rx}\right),$$

where

$$a_0 = Y_{D,u}\varphi, \quad a_1 = Y_{D,u}(l_E - l_D)B.$$

Under exact stoichiometric conditions, $\varphi = 0$ $(a_0 = 0)$, the diffusion flame location is $y = 1/2$ and extends to infinity, parallel to the x-axis. When conditions are not exactly stoichiometric, the location of attachment depends on the relative magnitude of the two Lewis numbers. When $l_E > l_D$, the flame is attached to the premixed flame on the interval $y = [1/2, 1]$ and extends downstream to a length $x^* = \frac{1}{r}\ln(a_1/a_0)$. Alternatively, when $l_E < l_D$, the diffusion flame is attached to the premixed flame on the interval $y = [0, 1/2]$ and its length is $x^* = \frac{1}{r}\ln(-a_1/a_0)$. By increasing φ, the diffusion flames shorten and disappear when φ is sufficiently large. The entire structure is shown schematically in Figure 3.1.

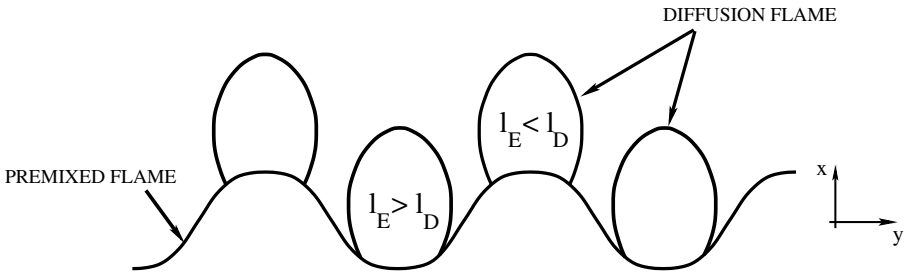


FIG. 3.1. *Schematic of diffusion flames trailing a segment of the perturbed premixed flame.*

Solutions for the temperature and species distribution on either side of the diffusion flame sheet are now found to be

$$\theta = l_D Be^{-rx}\cos(\pi y), \quad z_D = 0, \quad z_E = \nu W \quad \text{for} \quad W > 0, \tag{3.13}$$

and

$$\theta = l_E Be^{-rx}\cos(ky) + \varphi, \quad z_D = -W, \quad z_E = 0 \quad \text{for} \quad W < 0. \tag{3.14}$$

The temperature is continuous across the flame sheet, where it assumes the value

$$\theta^* = -\frac{l_D\varphi}{l_E - l_D}.$$

Note that in the absence of preferential diffusion, $l_E = l_D$, there are no trailing diffusion flames, since the deficient reactant is always depleted at the premixed reaction zone and only the initially excess species leaks through.

To investigate the structure of the diffusion flame, we must perform a local analysis of the reaction zone. Thus we introduce the local coordinate

$$y = F(x) + \beta^{-2/3}\zeta, \qquad F(x) = \frac{1}{\pi}\arccos\left(-\frac{a_0}{a_1}e^{rx}\right).$$

Note that this scaling suggests that the diffusion flame zone is broader than the premixed flame zone. We now introduce the expansions

$$\theta = \theta^* + \beta^{-2/3}\theta_1, \qquad Z_i = \beta^{-2/3}z_i,$$

and insert these into (3.5)–(3.7) to obtain

$$-\frac{\partial^2\theta_1}{\partial\zeta^2} = \frac{G(x)}{\nu Y_{D,u}}z_D z_E, \tag{3.15}$$

$$\frac{\partial^2 z_D}{\partial\zeta^2} = \frac{G(x)}{\nu}z_D z_E, \tag{3.16}$$

$$\frac{\partial^2 z_E}{\partial\zeta^2} = G(x)z_D z_E, \tag{3.17}$$

where

$$G(x) = \frac{e^{\theta^*}}{2(\varphi+2)Y_{D,u}(1+F_x^2)}.$$

Solutions to this system must match the outer solutions (3.13)–(3.14) which provide the conditions

$$z_E \sim \begin{cases} 0, & a_1\zeta > 0, \\ -\nu a_1\zeta\pi\sin(\pi F)e^{-rx}, & a_1\zeta < 0, \end{cases} \tag{3.18}$$

$$z_D \sim \begin{cases} a_1\zeta\pi\sin(\pi F)e^{-rx}, & a_1\zeta > 0, \\ 0, & a_1\zeta < 0. \end{cases} \tag{3.19}$$

To solve the above system, we first sum (3.16) and (3.17) to determine z_D in terms of z_E, i.e.,

$$z_D = \frac{1}{\nu}z_E + \zeta\alpha,$$

where $\alpha = a_1\pi\sin(\pi F(x))e^{-rx}$. Inserting this into (3.17) now results in a single boundary value problem for z_E. It is convenient to introduce the new variable U as $z_E = \nu\alpha U$, which satisfies

$$\frac{\partial^2 U}{\partial\zeta^2} = G(x)\alpha U(U+\zeta), \tag{3.20}$$

$$U \sim \begin{cases} -\zeta, & a_1\zeta \to -\infty, \\ 0, & a_1\zeta \to \infty. \end{cases} \tag{3.21}$$

The transformation

$$U \to (4G\alpha)^{-1/3}V, \quad \zeta = (2/G\alpha)^{1/3}\xi$$

now converts this system to the form

$$\frac{\partial^2 V}{\partial \xi^2} = (V+\xi)(V-\xi), \tag{3.22}$$

$$V \sim \begin{cases} -\xi, & \xi \to -\infty, \\ \xi, & \xi \to \infty. \end{cases} \tag{3.23}$$

This is identical to the diffusion flame structure for the infinite Damköhler number [16, 17], i.e., the Burke–Schumann flame sheets. Solutions of this system indicate that no reactants leak through the flame, and consequently no extinction is possible. We note that these structures resemble tribrachial (triple) flames, which consist of both lean and rich premixed flame segments with a diffusion flame attached to the stoichiometric point and trailing downstream. The present analysis suggests that such structures may form as a result of self-wrinkling of premixed flames when conditions are close to stoichiometric.

4. Flames in counterflow. We now examine the response of a positively stretched flame in a counterflow under near-stoichiometric conditions. As shown in Figure 4.1, consider a flow originating at $x = -\infty$ and impinging against an adiabatic wall located at $x = 0$. This configuration supports a planar flame situated at $x = -d$. Note that this geometry also corresponds to a twin flame configuration with a symmetric flame located in the right half-plane at $x = d$. The flow remains potential flow in the absence of density variations and is given by $\mathbf{V} = K(-x, y)$, where K is the strain rate. This problem was considered previously by Buckmaster [4] using a single-reactant model.

If the flame is planar, the solution is independent of y, and thus we seek steady solutions to the governing equations

$$-Kx\frac{d\tau}{dx} = \frac{d^2\tau}{dx^2}, \quad x < -d, \tag{4.1}$$

$$\tau = 1, \quad x > -d, \tag{4.2}$$

$$-Kx\frac{dh_i}{dx} = \frac{d^2h_i}{dx^2} + l_i\frac{d^2\tau}{dx^2}, \quad x \neq -d. \tag{4.3}$$

These equations are to be solved subject to the following jump conditions evaluated at $x = -d$:

$$[\tau] = 0, \quad [h_i] = 0, \tag{4.4}$$

$$\left[\frac{dh_i}{dx}\right] + l_i\left[\frac{d\tau}{dx}\right] = 0, \tag{4.5}$$

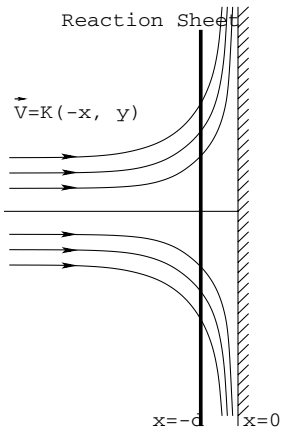


FIG. 4.1. *Schematic of a premixed flame in stagnation point flow against an adiabatic wall. The flame is situated at* $x = -d$.

$$\left[\frac{d\tau}{dx}\right] = \begin{cases} -\sqrt{\frac{2+\varphi+h_E^*-h_D^*}{2+\varphi}}\, e^{h_D^*/2} & \text{if } \varphi + h_E^* - h_D^* > 0, \\ -\sqrt{\frac{2-\varphi+h_D^*-h_E^*}{2+\varphi}}\, e^{(h_E^*+\varphi)/2} & \text{if } \varphi + h_E^* - h_D^* < 0, \end{cases} \tag{4.6}$$

where the boundary conditions are

$$\tau \to 0, \quad h_D \to 0, \quad h_E \to 0, \quad \text{as} \quad x \to -\infty, \tag{4.7}$$

$$\frac{dh_D}{dx}(0-) = 0, \qquad \frac{dh_E}{dx}(0-) = 0. \tag{4.8}$$

The solution to this system is

$$\tau = \begin{cases} \operatorname{erfc}\left(-\sqrt{\frac{K}{2}}x\right) / \operatorname{erfc}\left(\sqrt{\frac{K}{2}}d\right), & x < -d, \\ 1, & x > -d, \end{cases} \tag{4.9}$$

$$h_i = \begin{cases} -l_i \frac{(1+Kd^2)\operatorname{erfc}\left(-x\sqrt{\frac{K}{2}}\right)}{2\operatorname{erfc}\left(d\sqrt{\frac{K}{2}}\right)} - \frac{l_i\sqrt{\frac{K}{2\pi}}}{\operatorname{erfc}\left(d\sqrt{\frac{K}{2}}\right)}\, x\, e^{\frac{-Kx^2}{2}}, & x < -d, \\ -\frac{l_i}{2}(1+Kd^2) + \frac{l_i\sqrt{\frac{K}{2\pi}}}{\operatorname{erfc}\left(d\sqrt{\frac{K}{2}}\right)}\, d\, e^{\frac{-Kd^2}{2}}, & -d < x < 0. \end{cases} \tag{4.10}$$

We note the nonzero enthalpy perturbations behind the flame, and, in particular, the sign of the quantity

$$\varphi + (l_D - l_E)\left(\frac{1}{2}(1+Kd^2) - \sqrt{\frac{K}{2\pi}}\frac{de^{-Kd^2/2}}{\operatorname{erfc}\left(d\sqrt{K/2}\right)}\right)$$

determines which of the two conditions (4.6) is to be used to determine the location of the flame, d. In particular, we find the following.

$\varphi + (l_D - l_E)\Gamma > 0$:

$$\sqrt{K} = \sqrt{\frac{\pi}{2}}\ \mathrm{erfc}(\gamma)\ e^{\gamma^2}\ (1+\mu\Gamma)^{\frac{1}{2}}\ e^{\frac{-l_D}{2}\Gamma}, \tag{4.11}$$

$\varphi + (l_D - l_E)\Gamma < 0$:

$$\sqrt{K} = \sqrt{\frac{\pi}{2}}\ \mathrm{erfc}(\gamma)\ e^{\gamma^2}\left(\frac{2-\varphi}{2+\varphi} - \mu\Gamma\right)^{\frac{1}{2}}\ e^{\frac{-l_E}{2}\Gamma + \frac{\varphi}{2}}, \tag{4.12}$$

where

$$\Gamma = \frac{1}{2} + \gamma^2 - \frac{\gamma e^{-\gamma^2}}{\sqrt{\pi}\ \mathrm{erfc}(\gamma)}, \qquad \gamma = \sqrt{\frac{K}{2}}\ d, \tag{4.13}$$

and

$$\mu = \frac{l_D - l_E}{2+\varphi}. \tag{4.14}$$

For a given mixture, μ and φ are specified, and these expressions determine the flame standoff distance, d, as a function of strain rate, K. Since $\gamma > 0$, the response curve can be constructed by incrementing from $\gamma = 0$ and calculating K from either (4.11) or (4.12), depending on the appropriate inequality that characterizes the mixture. The standoff distance, d, is then found from the second equation in (4.13). As is the case in the single-reactant theory, the dependence of d on K is found to either be a monotonically decreasing function, reaching $d = 0$ at a sufficiently large value of K, or to have a backward C-shape with a turning point at a critical value of K. In the former case, the flame can be pushed all the way to the wall. In the latter case, the turning point corresponds to an extinction point, a maximum strain rate beyond which no steady solution exists.

While constructing the response curves, one must continually monitor the sign of the inequality in (4.11) and (4.12) to decide which condition is the appropriate one to apply. For the uniformly stretched flame considered here, i.e., for K constant, the same equation determines the response along the entire flame surface. In other words, for a given set of parameters, the same species leaks through along the entire flame surface. This is in contrast to the corrugated flame discussed earlier, where it was shown that some portions of the surface burn lean while neighboring regions burn rich.

4.1. Flame response. To discuss the general features of the response curves, we first consider the quantity $\varphi + (l_D - l_E)\Gamma$, whose sign determines whether (4.11) or (4.12) is to be used. A plot of Γ as a function of γ is shown in Figure 4.2. Clearly, Γ is always positive, it assumes its largest value of $1/2$ at $\gamma = 0$, and then decreases monotonically to zero as $\gamma \to \infty$. Since φ, as defined, is also positive, we can conclude that (4.11) determines the response in the following cases, for which, clearly, $\varphi + (l_D - l_E)\Gamma > 0$.

$\varphi \gg \mathbf{1}$:
When conditions are far removed from stoichiometry, the single-reactant theory is recovered. In this limit, $\mu \to 0$ and (4.11) reduces to Buckmaster's result [4] in which the flame response depends only on l_D.

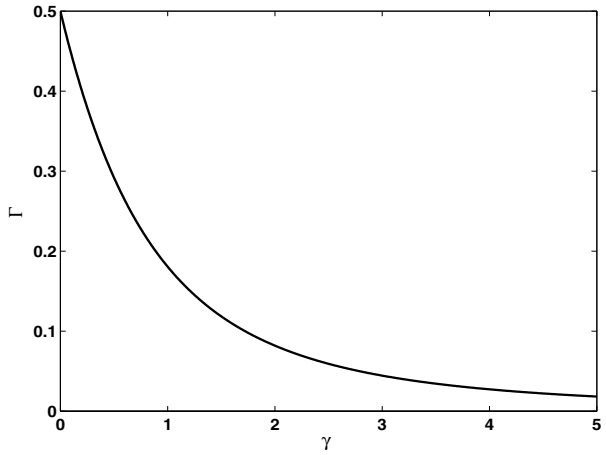


FIG. 4.2. *The quantity* Γ *as a function of* γ*, as determined by* (4.13).

$K \ll 1$:
When the flame is weakly strained, such that $\gamma \gg 1$, the flame retreats and stands far from the stagnation plane, i.e., $d \to \infty$. The structure of these flames is essentially that of a planar flame in a uniform flow field.
$l_D > l_E$:
When the initially deficient reactant is the less mobile of the two, for example, propane in a lean propane/air mixture, it will always be consumed at the reaction sheet.

Equation (4.12) determines the response only when $\varphi + (l_D - l_E)\Gamma < 0$. This occurs for moderately strained flames, in mixtures with $l_E > l_D$, and when conditions are sufficiently close to stoichiometry. In such cases, the initially excess reactant is the less mobile of the two, and its concentration in the reaction zone is consequently low. It is ultimately consumed at the reaction sheet, with the initially deficient species leaking through.

Typical response curves are shown in Figures 4.3–4.6. Figures 4.3 and 4.4 show the standoff distance and flame speed, respectively, as a function of strain rate for several different values of the Lewis number and conditions far removed from stoichiometry, $\varphi \to \infty$. This limit corresponds to the single-reactant theory of Buckmaster [4]. As shown in Figure 4.3, when $l_D < l_D^* = 4$ the response is monotonic, suggesting that the flame can be pushed all the way down to the wall before it is extinguished. When $l_D > l_D^*$, the response curve becomes double-valued. The turning point observed in the corresponding curves is regarded as the extinction point. When the strain rate exceeds the critical value at the turning point, the system can no longer support a planar flame and extinction occurs. The lower portion of the curve, below the extinction point, is presumed to be unstable.

The flame speed, S_f, defined to be the speed of the flame relative to the underlying flow field, is the normal velocity of the incoming flow evaluated at the reaction sheet, i.e., $S_f = Kd$. Curves illustrating the dependence of S_f on K are shown in Figure 4.4. Again we observe that for $\varphi \to \infty$, $l_D^* = 4$ is the critical condition determining the form of the response. We also note that the flame speed of a weakly strained flame can exceed the adiabatic flame speed, $S_f = 1$, when l_D is sufficiently small. This is consistent with previous studies of flame response to straining, cf. [18]. When the flame is weakly strained, the flame retreats far from the stagnation plane and $d \to \infty$.

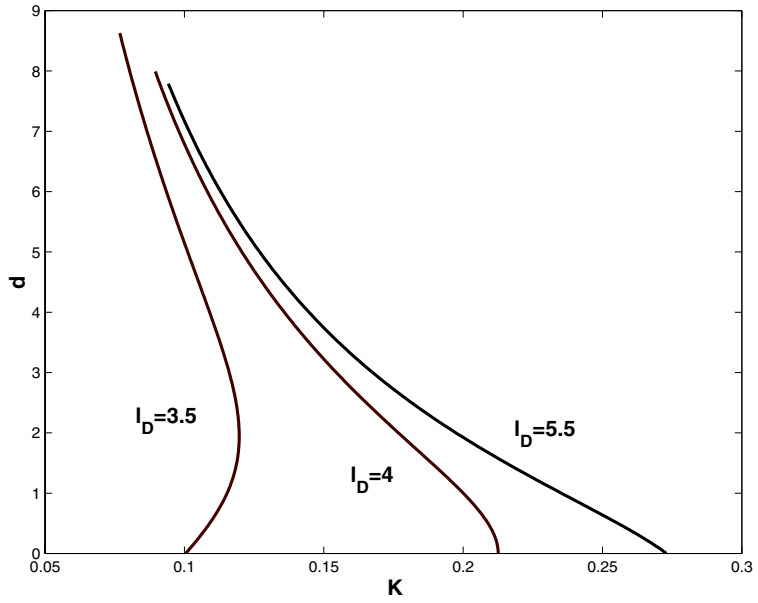


FIG. 4.3. *Standoff distance vs. strain rate for conditions far removed from stoichiometry.*

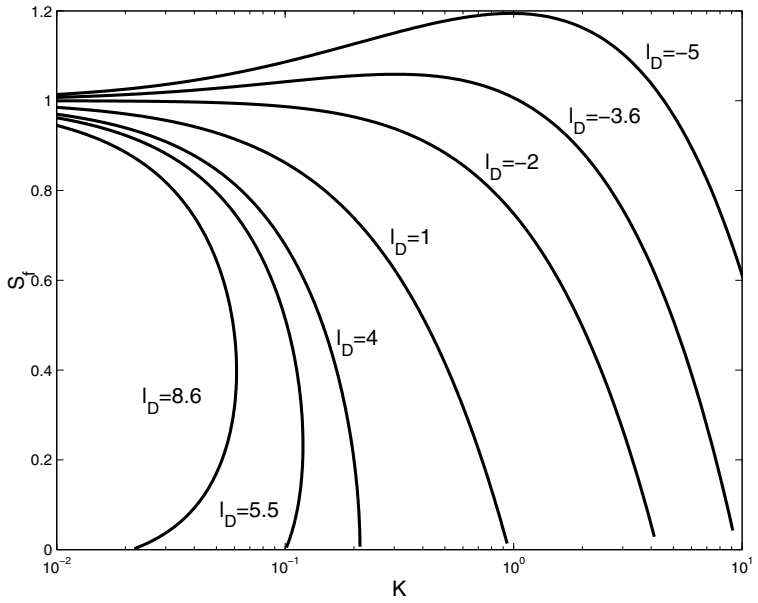


FIG. 4.4. *Standoff distance vs. strain rate for conditions far removed from stoichiometry.*

In this case $\gamma \to \infty$, and thus, again, the inequality $\varphi + (l_D - l_E)\Gamma > 0$ holds for any nonzero φ. Therefore (4.11) is always the appropriate condition to determine flame behavior. For the weak strain rate, this condition determines the flame position to be

$$d = \frac{1}{K} + \left(\frac{\mu - l_D}{2} - 1\right) + o(1),$$

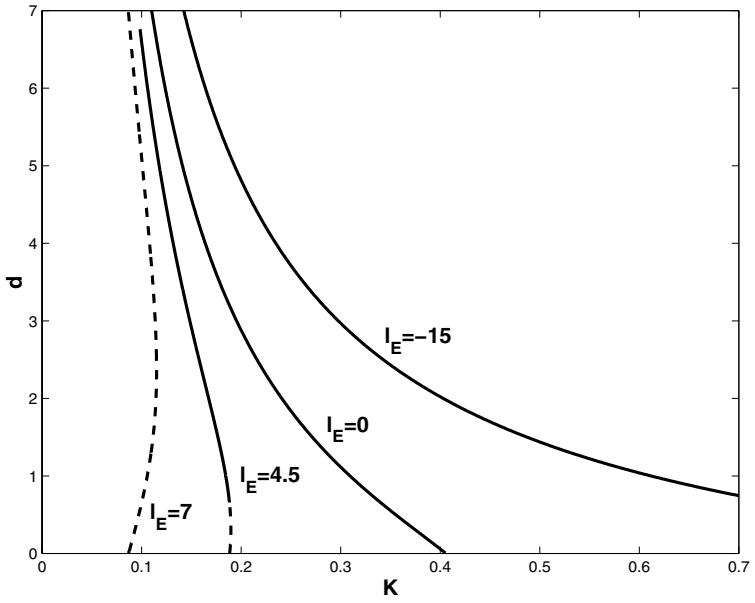


FIG. 4.5. *Standoff distance vs. strain rate for several values of* l_E *with* $l_D = 4.0$ *and* $\varphi = 0.2$. *The solid portions of the curves are determined by* (4.11)*, and the dashed portions by* (4.12).

and the flame speed takes the form

$$S_f = 1 - \left(1 + \frac{l_D - \mu}{2}\right) K + o(K).$$

We conclude that the flame speed will exceed its adiabatic value when $l_D - \mu < -2$, which reduces to the single-reactant theory result, $l_D < -2$, when conditions are far removed from stoichiometry ($\mu \to 0$).

The curves in Figures 4.5 and 4.6 are drawn for selected values of l_E and conditions close to stoichiometry, with $\varphi = 0.2$ and $l_D = 4.0$. The solid portion of each curve represents the segment determined by (4.11), while the dashed portions show where (4.12) is valid.

Since smaller values of d result from larger values of Γ, (4.12) always becomes relevant on the lower portion of the curves when l_E sufficiently exceeds l_D. When the flame is near the wall, the low mobility of the initially excess reactant prevents it from diffusing quickly enough across the strained flow field and it becomes locally deficient, and hence consumed, at the reaction sheet. As the flame moves further away from the wall and the strain weakens, (4.11) will eventually take over.

We note that in situations where the curve becomes double-valued, the transition from a lean to a rich flame may occur either before or after the turning point, and thus extinction conditions may be determined by either (4.11) or (4.12). When (4.11) is valid, the strain rate at which the reaction sheet reaches the wall is given by

$$K = \frac{\pi}{2}(1 + \mu/2)e^{-l_D/2}. \tag{4.15}$$

When a turning point first develops, the slope at this point of intersection will be

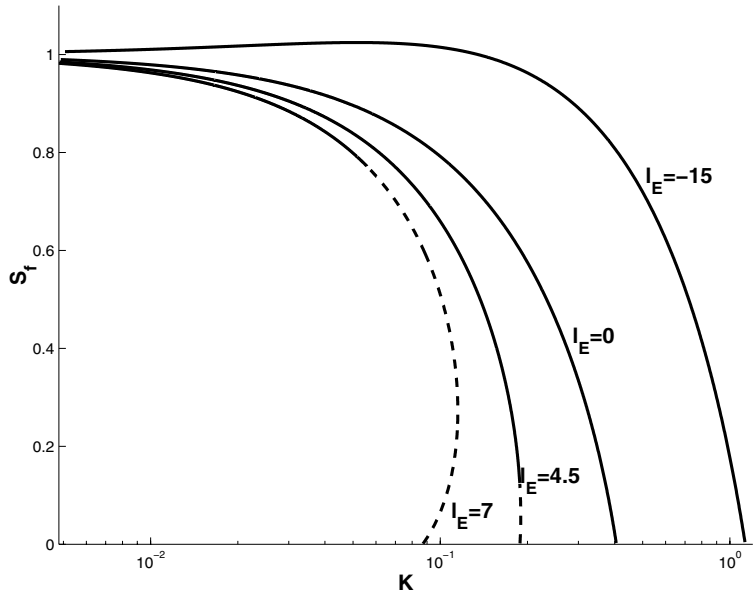


FIG. 4.6. *Flame speed vs. strain rate for several values of* l_E *with* $l_D = 4.0$ *and* $\varphi = 0.2$.

infinite, and upon differentiating (4.11) we find that this occurs at

$$l_D + \frac{2\mu}{2+\mu} = 4.$$

The response is single-valued when $l_D + \frac{2\mu}{2+\mu} < 4$ and double-valued when $l_D + \frac{2\mu}{2+\mu} > 4$. When (4.12) is valid, the value of the strain rate when $d = 0$ is found to be

$$K = \frac{\pi}{2}(p - \mu/2)e^{-l_E/2+\varphi}, \tag{4.16}$$

and the response is single-valued when $l_E + \frac{2\mu}{2p-\mu} < 4$ and double-valued when $l_E + \frac{2\mu}{2p-\mu} > 4$, where $p = (2-\varphi)/(2+\varphi)$.

Comparing Figures 4.3 and 4.5, we observe that the response curves showing d vs. K with $l_D = 4$ and $\varphi = 0.2$ may be either monotonic or double-valued, depending on the values of l_E. The single-reactant model predicts only a monotonic response; see Figure 4.3. Similarly, the flame speed can be significantly modified when conditions are close to stoichiometry, as shown in Figure 4.6. Whereas the single-reactant theory predicts a monotonically decreasing response when $l_D = 4$, the present theory demonstrates that the response may be either single- or double-valued. Furthermore, the flame speed can exceed its adiabatic value for sufficiently small l_E, as can be seen for the curve with $l_E = -15$.

4.2. Effect of the equivalence ratio. Recall that the effects of stoichiometry are measured in terms of φ, which is the deviation from unity of the ratio of the mass fractions of the excess-to-deficient reactants. Also, as defined in (2.6), $\varphi = \beta(\Phi - 1)$ is always positive. Note that the deviation from unity of the equivalence ratio, $\phi_1 = \beta(\phi - 1)$, is related to φ in the following way:

$$\varphi = \begin{cases} \phi_1, & \text{rich mixtures,} \\ -\phi_1, & \text{lean mixtures.} \end{cases}$$

For a given fuel mixture, the roles of l_D and l_E are reversed as stoichiometry is crossed. That is, for lean mixtures l_D and l_E assume the values of l_F and l_O, respectively. However, when the mixture composition is altered to become fuel-rich, $l_D = l_O$ and $l_E = l_F$. It follows from our discussion above that (4.11) will always be valid on one side of stoichiometry while (4.12) will be valid on the other, at least in the immediate vicinity of $\phi_1 \approx 0$. This implies, for example, that a slightly rich mixture can burn lean, with the fuel totally consumed at the reaction sheet, when the Lewis number of the fuel exceeds that of the oxidizer. Of course, as conditions move sufficiently far away from stoichiometry, (4.11) always takes over, indicating that a sufficiently rich mixture will always burn rich. As discussed, Γ assumes its largest value of $1/2$ when $d = 0$, so that in order for (4.12) to play a role along some portion of the response curve, the following inequality must be satisfied:

$$\varphi + (l_D - l_E)/2 < 0.$$

4.2.1. Heavy fuels. To illustrate typical response curves over a range of equivalence ratios, we will first consider a mixture with $l_F = 6.0$, $l_O = 0.0$. These are representative values for mixtures of heavy hydrocarbons in air, for example. For lean mixtures, $l_D = 6.0$, $l_E = 0.0$, and it follows that $\varphi + (l_D - l_E)/2 > 0$. Therefore the entire flame response is determined by (4.11), and the fuel is always consumed. Furthermore, for these parameter values, $l_D + \frac{2\mu}{2+\mu} > 4$ so the response is always double-valued. For rich mixtures, on the other hand, $l_D = 0.0$, $l_E = 6.0$, and we find that (4.12) will determine the lower portion of the response curve when $0 < \phi_1 < 3.0$; (4.11) is valid otherwise. That is, up to $\phi_1 = 3.0$, the flame burns lean when the flame is near the wall and rich when the flame is sufficiently far from the wall. Beyond $\phi_1 = 3.0$ the flame always burns rich. The above inequalities also determine that the response is double-valued when $0 < \phi_1 < 2.0$. Beyond $\phi_1 = 2.0$ the response is single-valued.

In Figure 4.7 we show the standoff distance, strain rate, and flame temperature at extinction as a function of the deviation from unity of the equivalence ratio. The dashed lines on the rich side of stoichiometry indicate that the flame is actually burning lean when it extinguishes. Beyond $\phi_1 = 2.0$ there is no turning point. In this region K_{ext} is taken to be the value of K at which the flame first touches the wall, or when twin flames merge in counterflow. Analysis of the system when $d \approx 0$, presented below, reveals that the flame is indeed extinguished at precisely this value of K, given explicitly by (4.15) and (4.16).

4.2.2. Light fuels. Now we consider a mixture for which the fuel is the lighter species, such as methane/air or hydrogen/air. In Figure 4.8 we show extinction conditions as a function of the equivalence ratio for $l_F = -2.0$, $l_O = 6.0$. The dashed lines on the lean side, but close to stoichiometry, indicate that the flame burns rich at extinction. Note that the trends are opposite of those for heavy fuels, as shown in Figure 4.7.

Similar curves can be readily constructed for various parameter values. In general when the Lewis number of the fuel sufficiently exceeds that of the oxidizer, the response is monotonic when the mixture is sufficiently rich and double-valued otherwise. The response is typical of heavy hydrocarbons such as propane and butane. On the other hand, when the fuel Lewis number is less than that of the oxidizer, the response is monotonic only if the mixture is sufficiently lean. The curves in Figure 4.8 showing d_{ext} as a function of the equivalence ratio are consistent with the theory presented in [12], as well as with the experimental results of Yamaoka and Tsuji [19],

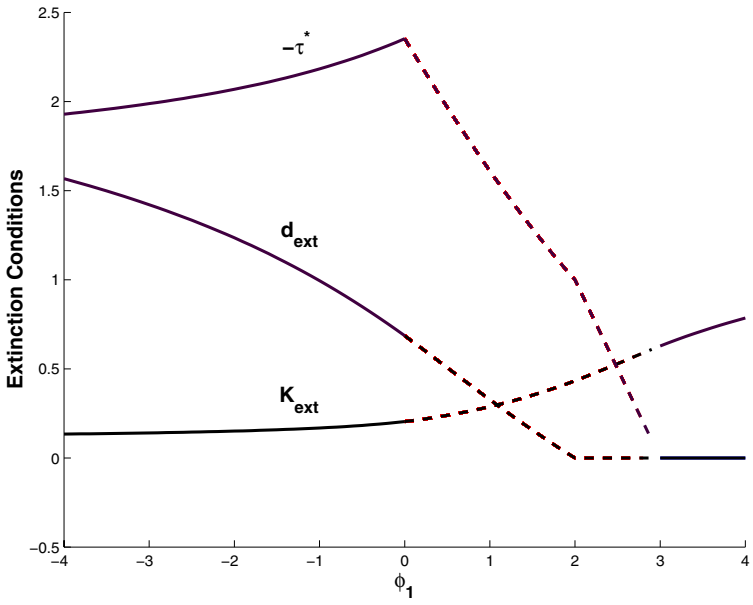


FIG. 4.7. *Standoff distance, strain rate, and (negative) flame temperature perturbation at extinction as a function of mixture strength for* $l_F = 6.0$ *and* $l_O = 0.0$. *The solid portions of the curves indicate extinction is determined by* (4.11) *and dashed portions by* (4.12).

where they measured the standoff distance as a function of the equivalence ratio for methane/air flames. We also note that the flame temperature perturbation levels off on the lean side for lighter fuels and on the rich side for heavier fuels. These regimes correspond to smaller effective Lewis numbers and thus these trends are consistent with the computations of Sato and Tsuji [20], who found that the flame temperature at extinction remains essentially a constant for Lewis numbers below unity.

4.3. The merged flame. In this section, we consider the structure of the flame for the case $d \approx 0$, in order to determine more accurately the extinction criteria after the flames have merged. An analysis of merged flames has been performed previously by Vedarajan, Buckmaster, and Ronney [21] for a single-reactant mixture with unity Lewis number. Here we consider a generalized two-reactant mixture with nonunity Lewis number.

The governing equations are

$$-Kx\frac{d\tau}{dx} = \frac{d^2\tau}{dx^2} + \frac{\beta^2}{2(\varphi+2)\nu Y_{D,u}^2} Y_D Y_E e^{\beta(\tau-1)}, \quad x \le 0, \tag{4.17}$$

$$-Kx\frac{dh_i}{dx} = \frac{d^2h_i}{dx^2} + l_i\frac{d^2\tau}{dx^2}, \quad x \le 0, \tag{4.18}$$

and the mass fractions are given by

$$Y_D/Y_{D,u} = 1 - \tau + \beta^{-1}h_D, \tag{4.19}$$

$$Y_E/\nu Y_{D,u} = 1 - \tau + \beta^{-1}(\varphi + h_E). \tag{4.20}$$

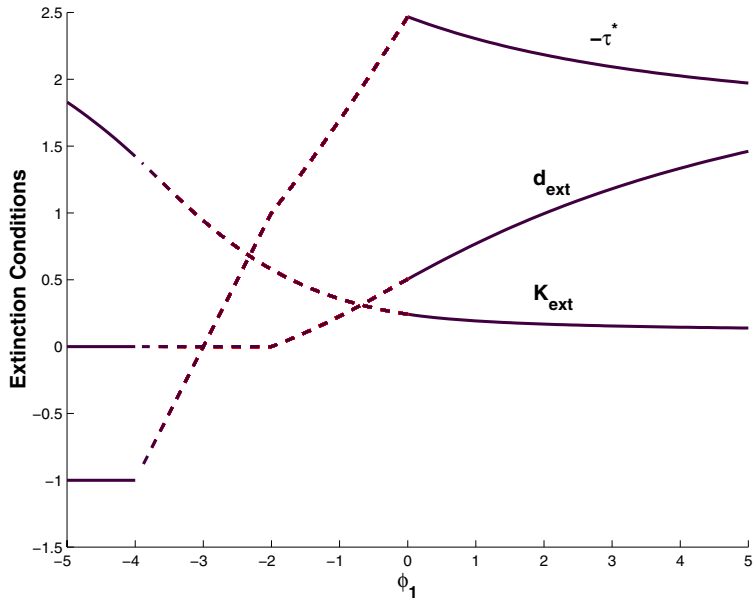


FIG. 4.8. *Standoff distance, strain rate, and (negative) flame temperature perturbation at extinction as a function of mixture strength for* $l_F = -2.0$ *and* $l_O = 6.0$. *The solid portions of the curves indicate extinction is determined by* (4.11) *and dashed portions by* (4.12).

Ahead of the flame, $x < 0$, the solutions are

$$\tau = \operatorname{erfc}(-\sqrt{K/2}x), \tag{4.21}$$

$$h_i = c_i \operatorname{erfc}(-\sqrt{K/2}x) - l_i\sqrt{K/2\pi}\,x e^{-Kx^2/2}, \tag{4.22}$$

where the constants, c_i, are determined by matching to the reaction zone near $x = 0$. The appropriate variables to analyze the reaction zone are

$$x = \beta^{-1}\xi, \quad \tau = 1 + \beta^{-1}\theta, \quad h_i = S_{i,0} + \beta^{-1}S_{i,1}. \tag{4.23}$$

To leading order, $S_{i,0}$ are found to remain constant throughout the reaction zone, and θ satisfies the equation

$$\frac{d^2\theta}{d\xi^2} = -\frac{1}{2(\varphi+2)}(S_{D,0} - \theta)(\varphi + S_{E,0} - \theta)e^{\theta}. \tag{4.24}$$

At the next order, matching the solution of the local enthalpy equation to the outer solution determines a condition for the leading order outer gradients, namely,

$$\frac{dh_i}{dx}(0^-) + l_i\frac{d\tau}{dx}(0^-) = 0.$$

This determines the constant c_i to be $c_i = -l_i/2$, and thus $S_{i,0} = -l_i/2$ throughout the reaction zone. Equation (4.24) can now be integrated once, and matching to the solution upstream yields the following formula for the temperature at the stagnation point θ^*:

$$\frac{K}{\pi} = \left\{\frac{l_D}{2}\left(\frac{l_E}{2} - \varphi\right) + \left(\frac{l_D}{2} + \frac{l_E}{2} - \varphi\right)(\theta^* - 1) + (\theta^* - 1)^2 + 1\right\}\frac{e^{\theta^*}}{2(\varphi+2)}. \tag{4.25}$$

The requirement of nonnegative mass fractions provides the following restriction on θ^*:

$$\theta^* \leq \min[-l_D/2;\ \varphi - l_E/2].$$

A typical plot of θ^* vs. K is shown in Figure 4.9 for the parameter values $\varphi = 1.0$, $l_D = l_E = 2.0$. Only the solid portion of the curve, below $\theta^* = -1$, is valid, to ensure nonnegative mass fractions. Extinction is seen to occur at the lower turning point, where $\theta^* = -1$.

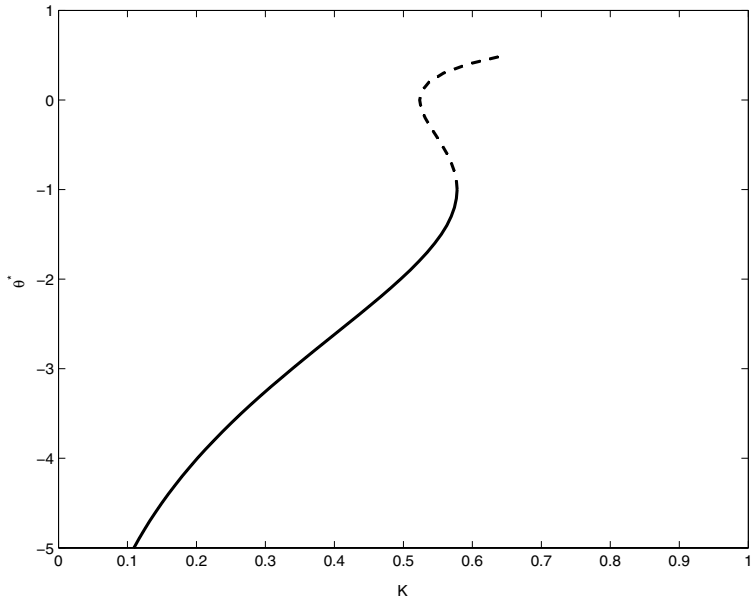


FIG. 4.9. *Flame temperature perturbation vs. strain rate for the merged flame, as determined by* (4.25). *Parameter values are* $l_D = l_E = 2.0$, $\varphi = 1.0$. *Only the solid portion of the curve is relevant, as the dotted portion corresponds to negative mass fractions.*

In general, the resulting curves from (4.25) have turning points at $\theta^* = -l_D/2$ and $\varphi - l_E/2$. However, as we have just discussed, physically relevant solutions exist only for θ^* less than the smaller of these values. Inserting these into (4.25) determines the strain rate at extinction to be

$$K_{\text{ext}} = \frac{\pi}{2}(1 + \mu/2)e^{-l_D/2} \quad \text{when } \theta^* = -l_D/2$$

and

$$K_{\text{ext}} = \frac{\pi}{2}(p - \mu/2)e^{-l_E/2+\varphi} \quad \text{when } \theta^* = \varphi - l_E/2.$$

These are precisely the values of the strain rate at which the flame first touches the stagnation plane, as predicted by the analysis of the separated flames; see (4.15) and (4.16). The present analysis therefore shows that when flames merge, extinction will occur immediately after the flames come into contact with one another (or with an adiabatic wall in the case of stagnation point flow).

5. Conclusions. We have derived a diffusional-thermal model of premixed flames under near-stoichiometric conditions. Our model, (2.18)–(2.23), differs from the single-reactant theory in several ways. First, it consists of an additional coupled equation for the second species, and, consequently, it has an explicit dependence on two Lewis numbers as well as the equivalence ratio. Second, the jump in temperature gradient across the reaction sheet is sensitive not only to the temperature perturbation but also to the local species concentrations. Finally, the derived jump conditions for the gradients across the flame sheet can take one of two forms, depending on which of the two species is ultimately consumed by the reaction. Although the derivation of our model assumes conditions close to stoichiometry, the single-reactant theory is recovered in the appropriate limit, indicating the validity of our model over the complete range of mixture strengths.

In general, the form of the final jump condition (2.23) depends on the sign of the quantity

$$\varphi + h_E^* - h_D^*.$$

This is directly proportional to the difference between the mass fractions of the initially deficient and excess reactants. We have determined that when this quantity is negative, the initially excess reactant is ultimately consumed by the reaction, while a small amount of the initially deficient reactant leaks through. For a weakly strained or curved flame, this quantity takes the form

$$\varphi + (l_D - l_E)(\kappa + K),$$

where κ is the curvature and K the strain rate. Thus, when the two species diffuse at unequal rates, which of the two is ultimately consumed depends on the magnitude and sign of the flame stretch, i.e., the combined effects of curvature and strain. This can have important implications on flame dynamics.

We employed our model to examine the behavior of corrugated flames using a sinusoidal disturbance imposed on a planar flame. Although the stability characteristics were found to be the same as those predicted by the single-reactant theory, albeit with an effective Lewis number properly defined, the overall flame structure is different. Under conditions sufficiently close to stoichiometry, the perturbed flame was found to burn rich along some portions and lean along others. The leakage of fuel and oxidizer along neighboring segments of the premixed flame front resulted in trailing diffusion flame tongues. Analysis of these flames showed them to be Burke–Schumann flame sheets, characterized by complete combustion, that were broader than the premixed reaction zone. Despite these differences, the dynamics of the premixed flame were shown to be unaffected by the local differences in mixture composition along the front.

Our model was also used to investigate the extinction of premixed flames in counterflow. We used our formulas to calculate critical values of strain rate and standoff distance at extinction over a range of equivalence ratios for parameter values typical of both heavy and light fuels. Our theory predicts that rich mixtures of heavy fuels, say, rich propane, are more resistant to strain than lean mixtures; they can withstand larger strain rates and can be pushed closer to the stagnation plane. The opposite trends are predicted for light fuels such as methane or hydrogen. Indeed, our results are consistent with the methane/air experiments reported by Yamaoka and Tsuji [19].

REFERENCES

[1] G. H. Markstein, *Experimental and theoretical studies of flame front stability*, J. Aero. Sci., 18 (1951), pp. 199–209.
[2] A. R. Burgess and L. J. Molero, *Nitrogenous emissions: The role of sulphur and equivalence ratio*, in Proceedings of the American Japanese Flame Research Committees International Symposium, Maui, HI, 1998, pp. 11–15.
[3] G. I. Sivashinsky, *Diffusional-thermal theory of cellular flames*, Combust. Sci. Tech., 15 (1977), pp. 137–145.
[4] J. D. Buckmaster, *The quenching of a deflagration wave held in front of a bluff body*, Proceedings of the Combust. Inst., 17 (1979), pp. 835–842.
[5] Y. D. Kim and M. Matalon, *Propagation and extinction of a flame in a stagnation-point flow*, Combust. Flame, 73 (1988), pp. 303–313.
[6] A. K. Sen and G. S. S. Ludford, *The near-stoichiometric behavior of combustible mixtures. Part* I: *Diffusion of the reactants*, Combust. Sci. Tech., 21 (1979), pp. 15–23.
[7] T. Mitani, *Propagation velocities of two-reactant flames*, Combust. Sci. Tech., 21 (1980), pp. 175–177.
[8] G. Joulin and T. Mitani, *Linear stability analysis of two-reactant flames*, Combust. Flame, 40 (1981), pp. 235–246.
[9] T. L. Jackson, *Effect of thermal expansion of the stability of two-reactant flames*, Combust. Sci. Tech., 53 (1987), pp. 51–54.
[10] G. I. Sivashinsky, *On flame propagation under conditions of stoichiometry*, SIAM J. Appl. Math, 39 (1980), pp. 67–82.
[11] M. Matalon, C. Cui, and J. K. Bechtold, *Hydrodynamic theory of premixed flames: Effects of stoichiometry, variable transport coefficients, and arbitrary reaction orders*, J. Fluid Mech., 487 (2003), pp. 179–210.
[12] J. K. Bechtold and M. Matalon, *Effects of stoichiometry on stretched premixed flames*, Combust. Flame, 119 (1999), pp. 217–232.
[13] K. Seshadri, N. Peters, J. A. van Oijen, and L. P. H. de Goey, *The asymptotic structure of weakly strained moderately rich methane-air flames*, Combust. Theory Model., 5 (2001), pp. 201–215.
[14] B. J. Matkowsky and G. I. Sivashinsky, *An asymptotic derivation of two models in flame theory associated with the constant density approximation*, SIAM J. Appl. Math, 37 (1979), pp. 686–699.
[15] F. A. Williams, *Combustion Theory: The Fundamental Theory of Chemically Reacting Flow Systems*, 2nd ed., Benjamin/Cummings, Menlo Park, CA, 1985.
[16] A. Liñán, *The asymptotic structure of counterflow diffusion flames for large activation energies*, Acta Astronaut., 1 (1974), pp. 1007–1039.
[17] S. Cheatham and M. Matalon, *A general asymptotic theory of diffusion flames with application to cellular instability*, J. Fluid Mech., 414 (2000), pp. 105–144.
[18] M. Matalon and B. J. Matkowsky, *Flames as gasdynamic discontinuities*, J. Fluid Mech., 124 (1982), pp. 239–259.
[19] I. Yamaoka and H. Tsuji, *An anomalous behavior of methane-air and methane-hydrogen-air flames diluted with nitrogen in a stagnation flow*, Proceedings of the Combust. Inst., 24 (1992), pp. 145–152.
[20] J. Sato and H. Tsuji, *Extinction of premixed flames in a stagnation point flow considering general Lewis number*, Combust. Sci. Tech., 33 (1983), pp. 193–205.
[21] T. G. Vedarajan, J. D. Buckmaster, and P. Ronney, *Two-dimensional failure waves and ignition fronts in premixed combustion*, Proceedings of the Combust. Inst., 27 (1998), pp. 537–544.

SIAM J. APPL. MATH.
Vol. 64, No. 4, pp. 1457–1475

A SEMICONTINUOUS FORMULATION OF THE BLOCH–BOLTZMANN–PEIERLS EQUATIONS*

CH. AUER†, F. SCHÜRRER†, AND W. KOLLER‡

Abstract. This paper presents new transport equations to describe the kinetics of coupled electron-phonon systems in semiconductors. The model is based on a semicontinuous version of the Bloch–Boltzmann–Peierls equations for electrons interacting with acoustic and optical phonons. Major aspects of the resulting kinetic equations, such as conservation laws and stability, are investigated. Numerical results of the model applied to a system of phonons and photo-excited electrons in gallium arsenide (GaAs) are compared with data obtained by an ensemble Monte Carlo technique.

Key words. electron-phonon kinetics, Bloch–Boltzmann–Peierls equations, semicontinuous model, far-from-equilibrium systems, laser photo-excitation

AMS subject classifications. 76P05, 82C40, 82D37, 82C22

DOI. 10.1137/S0036139903426101

1. Introduction.

Among various classical and quantum mechanical models [7], quasi-hydrodynamic transport equations are today's standard method for simulating the electrical behavior of semiconductor devices. This semiclassical and macroscopic description is based on the semiconductor Boltzmann equation [4]. One of the main requirements for the derivation of hydrodynamic equations, however, is the assumption that the distribution functions are well approximated by equilibrium distributions with time dependent parameters [13]. Since this condition is rarely met in ultra-integrated devices, i.e., the system is far from equilibrium, the focus of modeling the carrier transport must be shifted to directly solving the Boltzmann equation.

We consider a coupled system of semiclassical Boltzmann equations, the so-called Bloch–Boltzmann–Peierls (BBP) equations, for the interacting system of electrons and phonons. These transport equations have mainly been solved by means of ensemble Monte Carlo simulations [6] to investigate the kinetics of far-from-equilibrium systems [11, 18, 20]. Apart from the popular Monte Carlo technique, the discrete kinetic theory offers an alternative way of treating such problems. For elastically and inelastically interacting classical gases, the semicontinuous version of the Boltzmann equation is a well-established method [2, 8, 9, 10, 15, 16, 17].

Here, we derive a conservative formulation of the BBP equations in the framework of a semicontinuous kinetic theory. This formulation is based on a discretization of the energy variables and a continuum of momentum directions. Since a general class of dispersion relations of carriers and phonons is considered, some assumptions, which are discussed in section 3.1, are made for the aim of a concise treatment. As an application, we study numerically the relaxation of photo-excited electrons interacting with the phonon system in the Γ-valley of gallium arsenide (GaAs). The relaxation process is found to be significantly influenced by reabsorption processes of polar

*Received by the editors April 11, 2003; accepted for publication (in revised form) January 21, 2004; published electronically June 15, 2004. This work was supported by the Fonds zur Förderung der wissenschaftlichen Forschung, Vienna, under contract P14669-TPH.
http://www.siam.org/journals/siap/64-4/42610.html
†Institute of Theoretical and Computational Physics, Graz University of Technology, Petersgasse 16, A-8010 Graz, Austria (auer@itp.tu-graz.ac.at, schuerrer@itp.tu-graz.ac.at).
‡Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2AZ, United Kingdom (w.koller@imperial.ac.uk).

optical phonons.

Our approach (combined with a P_N-expansion) offers the possibility of solving the BBP equations for general dispersion relations of electrons, holes, and phonons with much less consumption of CPU-time than with Monte Carlo methods. It has the advantage of providing a posteriori error estimates. Moreover, the results are independent of random number generators. The derived model equations reflect all major aspects of the continuous description, such as conservation laws, equilibrium states, and stability theorems. A detailed mathematical proof of these properties is deferred to a subsequent publication.

The derived semicontinuous equations are generally applicable to problems without external fields. Such problems are of great interest in semiconductor device modeling. The relaxation after laser photo-excitation, recombination, and the effect of far-from-equilibrium phonon distributions on the electronic system can be studied and provide important insight into the dynamical behavior of electron-phonon systems.

This paper is organized as follows. In section 2 we review the continuous transport equations and discuss their interesting mathematical and physical features. The semicontinuous formulation of the BBP equations for a general class of dispersion relations for electrons and phonons is developed in section 3. In section 4 we summarize the conservation laws and provide an H-theorem concerning the stability of the equilibrium solution. Finally, we present some numerical results concerning hot electrons in GaAs in section 5.

2. Kinetic model. We set up our kinetic model with the standard BBP equations [21],

$$\frac{\partial f}{\partial t} + \boldsymbol{v}(\boldsymbol{k}) \cdot \nabla_r f = \sum_{i=1}^{n_{ph}} C_i^{el}(f, g_i), \tag{2.1}$$

$$\frac{\partial g_i}{\partial t} + \boldsymbol{c}_i(\boldsymbol{q}) \cdot \nabla_r g_i = C_i^{ph}(f, g_i), \tag{2.2}$$

with the phase space distribution functions $f(\boldsymbol{k}, \boldsymbol{r}, t)$ for electrons and $g_i(\boldsymbol{q}, \boldsymbol{r}, t)$ for different types $i = 1, \dots, n_{ph}$ of phonons. The wave vectors $\boldsymbol{k} \in \mathcal{B}, \boldsymbol{q} \in \mathcal{B}$ are elements of the first Brillouin zone $\mathcal{B}$, while $\boldsymbol{r} \in V \subset \mathbb{R}^3$ indicates the position and $t \in \mathbb{R}^+$ denotes the time. Further, we introduce the mean velocity of electrons $\boldsymbol{v}(\boldsymbol{k})$ and $\boldsymbol{c}_i(\boldsymbol{q})$ for the phonon type i by

$$\boldsymbol{v}(\boldsymbol{k}) = \frac{1}{\hbar} \nabla_k E(\boldsymbol{k}), \qquad \boldsymbol{c}_i(\boldsymbol{q}) = \nabla_q \omega_i(\boldsymbol{q}), \tag{2.3}$$

with $\hbar = h/2\pi$, the reduced Planck constant, the electron energy $E(\boldsymbol{k})$ in the considered band, and the dispersion relation $\omega_i(\boldsymbol{q})$ of the ith phonon branch. The collision operators

$$C_i^{el}(f, g_i) = \frac{V}{8\pi^3} \int_{\mathcal{B}} d\boldsymbol{k}' \Big[w_i^+(\boldsymbol{k}, \boldsymbol{k}', \boldsymbol{q}^+) \kappa_1(g_i^+, f, f') + w_i^-(\boldsymbol{k}, \boldsymbol{k}', \boldsymbol{q}^-) \kappa_2(g_i^-, f, f') \Big], \tag{2.4}$$

$$C_i^{ph}(f, g_i) = \frac{V}{4\pi^3} \int_{\mathcal{B}} d\boldsymbol{k}' \, w_i^+(\boldsymbol{k}^-, \boldsymbol{k}', \boldsymbol{q}) \kappa_1(g_i, f^-, f') \tag{2.5}$$

with the transition rates

$$w_i^{\pm}(\boldsymbol{k}, \boldsymbol{k}', \boldsymbol{q}) = s_i(\boldsymbol{q}) \, \delta[E(\boldsymbol{k}) - E(\boldsymbol{k}') \pm \hbar \omega_i(\boldsymbol{q})] \tag{2.6}$$

and the functions

$$\kappa_1(g,f,f') = (g+1)f'(1-f) - gf(1-f') = g(f'-f) + f'(1-f), \tag{2.7}$$

$$\kappa_2(g,f,f') = gf'(1-f) - (g+1)f(1-f') = g(f'-f) - f(1-f') \tag{2.8}$$

couple the electron equation with the phonon equations. We have omitted the arguments $\boldsymbol{r}, t$ of the distribution functions and have used the abbreviations $\boldsymbol{q}^{\pm} = \pm(\boldsymbol{k}'-\boldsymbol{k})$, $\boldsymbol{k}^{-} = \boldsymbol{k}' - \boldsymbol{q}$, $f = f(\boldsymbol{k})$, $f' = f(\boldsymbol{k}')$, $f^{-} = f(\boldsymbol{k}^{-})$, $g_i = g_i(\boldsymbol{q})$, $g_i^{\pm} = g_i(\boldsymbol{q}^{\pm})$ in (2.4) and (2.5) for brevity. The collision operators (2.4) and (2.5) with the transition rates (2.6) take into account only normal processes due to the electron-phonon interaction. The function $s_i(\boldsymbol{q})$ introduced in (2.6) is determined by the absolute value of the matrix element of the electron-phonon interaction Hamilton operator in the states $|\boldsymbol{k}\rangle$, $|\boldsymbol{k}'\rangle$ and, therefore, depends on the type of interaction. The delta distribution in the transition rates ensures conservation of the total energy of the affected electron and phonon in every scattering event. It should be noted that the wave vectors $\boldsymbol{k}, \boldsymbol{k}', \boldsymbol{q}$ do not appear independently from each other in the operators (2.4) and (2.5). If we consider an electron that is scattered from the initial state $\boldsymbol{k}$ to the final state $\boldsymbol{k}'$ involving a phonon of the branch i with wave vector $\boldsymbol{q}$, then the plus sign of $w_i^{\pm}$ represents an absorption process and the minus sign represents an emission process of the phonon.

The major aspects of the transport equations (2.1) and (2.2) with the collision operators (2.4) and (2.5), such as conservation laws, equilibrium solutions and their stability, as well as the initial value problem for the space homogeneous problem, have been analyzed by Majorana [12] in the case of a constant phonon frequency, i.e., $\omega_i(\boldsymbol{q}) = \omega_0$. A generalized kinetic theory for electrons and phonons, based on the BBP equations, is presented in [19].

3. Semicontinuous model. In this section, we derive a semicontinuous version of the kinetic equations (2.1) and (2.2) with the collision operators (2.4) and (2.5). For this purpose, it is necessary to distinguish between two different types of phonons: acoustic phonons, labeled with index $i = 1$, obeying the dispersion relation $e_1(\boldsymbol{q}) = \hbar\omega(\boldsymbol{q})$, and optical phonons, $i = 2$, with constant energy $e_2(\boldsymbol{q}) = e_0 = \hbar\omega_0$ (Einstein model).

In the case of electrons interacting with acoustic phonons, we first introduce the electron and phonon energy, E, $e = \hbar\omega$, as relevant variables of the distribution functions f, g_1 and reformulate the collision operators. This reformulation is possible under mild assumptions on the dispersion relations. The energy variables are then discretized in a second step.

Since optical phonons in the Einstein model cannot be characterized by their energy, we split up the wave vector into its modulus and direction, i.e., $\boldsymbol{q} = q\boldsymbol{\Omega}_{\boldsymbol{q}}$, and discretize the modulus q. After applying a special approximation procedure to functions of the discretized variables, we obtain the semicontinuous BBP equations.

3.1. Interaction geometry and assumptions on the dispersion relations. The semicontinuous model is based on an energy dependent formulation of the kinetic equations. To introduce the electron and phonon energies as relevant variables, we assume that the unique functions

$$|\boldsymbol{k}| = k(E, \boldsymbol{\Omega}_k), \qquad |\boldsymbol{q}| = q(\hbar\omega, \boldsymbol{\Omega}_q) \tag{3.1}$$

exist for the given dispersion relations $E(\boldsymbol{k}), \omega(\boldsymbol{k})$ if $\boldsymbol{\Omega}_k = \boldsymbol{k}/|\boldsymbol{k}| \in \mathbb{S}^2$, $\boldsymbol{\Omega}_q = \boldsymbol{q}/|\boldsymbol{q}| \in \mathbb{S}^2$ and $E \in I^1 \subset \mathbb{R}^+$, $\hbar\omega \in I^2 \subset \mathbb{R}^+$, respectively. The energy intervals I^1 and I^2 are

chosen in such a way that the distribution functions $f(\boldsymbol{k})$ and $g_1(\boldsymbol{k})$ vanish for energies $E(\boldsymbol{k}), \hbar\omega(\boldsymbol{k})$ outside I^1 and I^2, respectively. It should be mentioned that for common semiconductors the functions (3.1) usually exist in that region of the $\boldsymbol{k}$-space, which is relevant for the transport properties. In some cases, however, this region consists of several isolated parts located around band extrema. Such situations can be treated with multivalley models, where different species, one for each part of the Brillouin zone, are introduced and described separately by appropriate distribution functions. In the following we consider only one kind of electron.

In order to derive the semicontinuous BBP equations, we introduce suitable coordinate systems. This often turns out to be very useful when approximating Boltzmann collision operators [14]. We consider the orthonormal transformation matrix

$$\mathrm{T}(\boldsymbol{\Omega}) = \begin{pmatrix} \cos\vartheta\cos\varphi & -\sin\varphi & \sin\vartheta\cos\varphi \\ \cos\vartheta\sin\varphi & \cos\varphi & \sin\vartheta\sin\varphi \\ -\sin\vartheta & 0 & \cos\vartheta \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} \sin\vartheta\cos\varphi \\ \sin\vartheta\sin\varphi \\ \cos\vartheta \end{pmatrix} \tag{3.2}$$

with $\vartheta \in [0,\pi]$, $\varphi \in [0,2\pi]$, and two coordinate systems $\mathcal{S}_j$, $j = 1,2$, whose basis vectors $\boldsymbol{a}_i^{(j)} \in \mathbb{R}^3, i = 1,2,3$, should be given by

$$\boldsymbol{a}_i = \mathrm{T}^{-1}(\boldsymbol{\Omega}_k)\boldsymbol{a}_i^{(1)} = \mathrm{T}^{-1}(\boldsymbol{\Omega}_q)\boldsymbol{a}_i^{(2)}, \quad i = 1,2,3, \tag{3.3}$$

where $\boldsymbol{a}_i$ are the basis vectors of an arbitrary but fixed basic coordinate system. It should be pointed out that the third axis of the system $\mathcal{S}_1$ ($\mathcal{S}_2$) is aligned with the vector $\boldsymbol{k}$ ($\boldsymbol{q}$). If we represent the unit vector $\boldsymbol{\Omega}_{k'} = \boldsymbol{k}'/|\boldsymbol{k}'|$ by

$$\boldsymbol{\Omega}_{k'}^{(j)} = \begin{pmatrix} \sin\vartheta_j\cos\varphi_j \\ \sin\vartheta_j\sin\varphi_j \\ \cos\vartheta_j \end{pmatrix}, \qquad j = 1,2, \tag{3.4}$$

the quantity ϑ_j is the angle between $\boldsymbol{k}'$ and $\boldsymbol{k}$ for $j = 1$ and between $\boldsymbol{k}'$ and $\boldsymbol{q}$ for $j = 2$, which is advantageous for the integration of the collision operators (2.4) and (2.5).

The electron-phonon interaction processes considered in our kinetic model are governed by the energy and momentum relations

$$\boldsymbol{k}' = \boldsymbol{k} \pm \boldsymbol{q}, \qquad E(\boldsymbol{k}') = E(\boldsymbol{k}) \pm e_i(\boldsymbol{q}) \tag{3.5}$$

of the affected electron and phonon of type $i = 1,2$. Due to (3.4) the direction $\boldsymbol{\Omega}_{k'}$ of $\boldsymbol{k}'$ is parameterized by the angular variables ϑ_1 and φ_1. Now, we concentrate on interaction processes between electrons and acoustic phonons. We consider fixed values for the initial quasi momentum $\boldsymbol{k} = k(E,\boldsymbol{\Omega}_k)\boldsymbol{\Omega}_k$, the electron energy after the interaction E', and the polar angle φ_1. The solutions of (3.5), for the fixed parameters E, $\boldsymbol{\Omega}_k$, E', and φ_1, determine the scattering angle ϑ_1 as well as the phonon wave vector $\boldsymbol{q}$ for allowed scattering events. For the following treatment, we assume that such solutions, if they exist, are unique. In the case of optical phonons, we impose the uniqueness of solutions to (3.5) for fixed parameters E, $\boldsymbol{\Omega}_k$, q, and φ_1. Concerning the phonon collision operators, we additionally require that solutions to (3.5) with respect to ϑ_2 are unique for a fixed phonon wave vector $\boldsymbol{q}$ and fixed parameters E' and φ_2. These conditions are satisfied by dispersion relations which depend weakly on the direction of the wave vector. It should be pointed out that the above mentioned assumptions are made in order to ensure a clear formalism and notation.

3.2. Integration of the collision operators. We reformulate the collision operators by carrying out the variable transformation $(k'_1, k'_2, k'_3) \to (E', \vartheta_1, \varphi_1)$ in the integral of the electron collision operator (2.4) with the relation

$$\boldsymbol{k}'(E', \xi_1, \varphi_1, \boldsymbol{\Omega}_k) = k[E', \boldsymbol{u}(\xi_1, \varphi_1, \boldsymbol{\Omega}_k)]\, \boldsymbol{u}(\xi_1, \varphi_1, \boldsymbol{\Omega}_k), \tag{3.6}$$

$$\boldsymbol{u}(\xi_j, \varphi_j, \boldsymbol{\Omega}_k) = \mathrm{T}(\boldsymbol{\Omega}_k)\, \boldsymbol{\Omega}_{k'}^{(j)}, \quad j = 1, 2, \tag{3.7}$$

where we split up the vector $\boldsymbol{k}'$ into its modulus and direction unit vector. The latter is expressed as a function of ϑ_1, φ_1 through the transformation (3.2) and we introduce the abbreviation $\xi_1 = \cos\vartheta_1$. The absolute value of the Jacobi determinant results in

$$|J(E', \xi_1, \varphi_1, \boldsymbol{\Omega}_k)| = \left|\det\left[\frac{\partial(k'_1, k'_2, k'_3)}{\partial(E', \xi_1, \varphi_1)}\right]\right| = k'^2 \left|\frac{\partial k'}{\partial E'}\right| \tag{3.8}$$

with $k' = k[E', \boldsymbol{u}(\xi_1, \varphi_1, \boldsymbol{\Omega}_k)]$. Concerning the integration limits, we can assume that the relevant $\boldsymbol{k}$-region, i.e., in which $E(\boldsymbol{k}) \in I^1$, is a subset of the first Brillouin zone $\mathcal{B}$ and we obtain

$$\boldsymbol{k}' \in \mathcal{B} \longrightarrow E' \in I^1, \quad \xi_1 \in [-1, 1], \quad \varphi_1 \in [0, 2\pi]. \tag{3.9}$$

Taking into account (2.6), the electron collision operators (2.4) can then be written as

$$C_i^{el}(f, g_i) = \frac{V}{8\pi^3} \int_{I^1} dE' \int_0^{2\pi} d\varphi_1 \int_{-1}^{1} d\xi_1\, |J| \left\{ s_i^+ \kappa_1(g_i^+, f, f') \delta(F_i^+) + s_i^- \kappa_2(g_i^-, f, f') \delta(F_i^-) \right\} \tag{3.10}$$

for $i = 1, 2$ with the functions

$$F_i^{\pm}(E, E', \xi_1, \varphi_1, \boldsymbol{\Omega}_k) = E - E' \pm e_i[\boldsymbol{q}^{\pm}(E, E', \xi_1, \varphi_1, \boldsymbol{\Omega}_k)], \tag{3.11}$$

$$\boldsymbol{q}^{\pm}(E, E', \xi_1, \varphi_1, \boldsymbol{\Omega}_k) = \pm\{k[E', \boldsymbol{u}(\xi_1, \varphi_1, \boldsymbol{\Omega}_k)]\, \boldsymbol{u}(\xi_1, \varphi_1, \boldsymbol{\Omega}_k) - k(E, \boldsymbol{\Omega}_k)\, \boldsymbol{\Omega}_k\} \tag{3.12}$$

and the abbreviation $s_i^{\pm} = s_i(\boldsymbol{q}^{\pm})$. At first sight the representation (3.6) seems to result in needlessly complicated expressions for the integrand of the collision operator (3.10). However, it has the advantage that we obtain convenient functions $q^{\pm}(E, E', \xi_1, \varphi_1, \boldsymbol{\Omega}_k)$, representing the modulus of the phonon wave vector. For example, if we consider dispersion relations, which depend on only the wave vector moduli, then $q^{\pm}$ depends on only ξ_1, i.e., the cosine of the angle between $\boldsymbol{k}$ and $\boldsymbol{k}'$ for given energies E, E'.

An analogous treatment of the phonon collision operators (2.5) leads to

$$C_i^{ph}(f, g_i) = \frac{V}{4\pi^3} s_i \int_{I^1} dE' \int_0^{2\pi} d\varphi_2 \int_{-1}^{1} d\xi_2\, |J| \kappa_1(g_i, f^-, f') \delta(G_i) \tag{3.13}$$

for $i = 1, 2$, where we have introduced the functions

$$G_i(E', \xi_2, \varphi_2, \boldsymbol{q}) = E[\boldsymbol{k}^-(E', \xi_2, \varphi_2, \boldsymbol{q})] - E' + e_i(\boldsymbol{q}), \tag{3.14}$$

$$\boldsymbol{k}^-(E', \xi_2, \varphi_2, \boldsymbol{q}) = k[E', \boldsymbol{u}(\xi_2, \varphi_2, \boldsymbol{\Omega}_q)]\, \boldsymbol{u}(\xi_2, \varphi_2, \boldsymbol{\Omega}_q) - \boldsymbol{q}. \tag{3.15}$$

We continue with the investigation of the poles $\xi_1^{\pm}$ of the delta distribution $\delta(F_1^{\pm})$ defined by

$$F_1^{\pm}(E, E', \xi_1, \varphi_1, \boldsymbol{\Omega}_k) = 0. \tag{3.16}$$

We consider solutions to (3.16) with respect to ξ_1 for arbitrary, but fixed, parameters E, E', φ_1, and $\mathbf{\Omega}_k$. According to section 3.1 there exist unique solutions, and we denote them with

$$\xi_1^{\pm} = \zeta_1^{\pm}(E, E', \varphi_1, \mathbf{\Omega}_k). \tag{3.17}$$

It must be remembered that the functions (3.17) are subject to the condition

$$|\zeta_1^{\pm}(E, E', \varphi_1, \mathbf{\Omega}_k)| \leq 1 \tag{3.18}$$

because of $\xi_1 = \cos\vartheta_1$. The arguments of the functions (3.17) will be omitted in the following for brevity. If the solutions to (3.16) do not exist for the given parameters E, E', φ_1, $\mathbf{\Omega}_k$ or if the solutions do not satisfy the condition (3.18), the collision term (3.10) vanishes. Hence, the energies E' can be restricted to the range

$$\mathcal{D}^{a\pm}(E, \varphi_1, \mathbf{\Omega}_k) = \left\{ E' \in I^1 \middle| \, \exists \zeta_1^{\pm} \wedge |\zeta_1^{\pm}| \leq 1 \right\}, \tag{3.19}$$

which can be seen as a selection rule for the scattering processes. Now, we carry out the integration with respect to ξ_1 in the operator (3.10) to obtain

$$C_1^{el}(f, g_1) = \int_0^{2\pi} d\varphi_1 \int_{I^1} dE' \left[\sigma_1^{+} \kappa_1(g_1^{+}, f, f'^{+}) + \sigma_1^{-} \kappa_2(g_1{}^{-}, f, f'^{-}) \right], \tag{3.20}$$

where we use $f'^{\pm} = f[k(E', \boldsymbol{u}_1^{\pm})\boldsymbol{u}_1^{\pm}]$ and the functions $\boldsymbol{u}_1^{\pm} = \boldsymbol{u}(\zeta_1^{\pm}, \varphi_1, \mathbf{\Omega}_k)$, $g_1^{\pm} = g_1[\boldsymbol{q}^{\pm}(E, E', \zeta_1^{\pm}, \varphi_1, \mathbf{\Omega}_k)]$ are evaluated at $\xi_1 = \zeta_1^{\pm}$. The scattering rate is given by

$$\sigma_1^{\pm}(E, E', \varphi_1, \mathbf{\Omega}_k) = \frac{V}{8\pi^3} \left[|J| s_1^{\pm} \left(\left| \frac{\partial F_1^{\pm}}{\partial \xi_1} \right| \right)^{-1} \right]_{\xi_1 = \zeta_1^{\pm}} \chi_{\mathcal{D}^{a\pm}}(E'), \tag{3.21}$$

with the characteristic function $\chi_{\mathcal{M}}$ of the set $\mathcal{M}$. Further, we use (3.8), (3.11), (3.19), and $s_1^{\pm} = s_1(\boldsymbol{q}^{\pm})$, where $\boldsymbol{q}^{\pm}$ is given in (3.12). It should be mentioned that $e[\boldsymbol{q}^{\pm}(\xi_1 = \zeta_1^{\pm})] = \pm(E' - E)$ and, therefore, we can write

$$\boldsymbol{q}^{\pm} = q[\pm(E' - E), \hat{\boldsymbol{u}}^{\pm}(E, E', \zeta_1^{\pm}, \varphi_1, \mathbf{\Omega}_k)] \hat{\boldsymbol{u}}^{\pm}(E, E', \zeta_1^{\pm}, \varphi_1, \mathbf{\Omega}_k), \tag{3.22}$$

$$\hat{\boldsymbol{u}}^{\pm}(E, E', \xi_j, \varphi_j, \mathbf{\Omega}_k) = \frac{\boldsymbol{q}^{\pm}(E, E', \xi_j, \varphi_j, \mathbf{\Omega}_k)}{|\boldsymbol{q}^{\pm}(E, E', \xi_j, \varphi_j, \mathbf{\Omega}_k)|}, \quad j = 1, 2. \tag{3.23}$$

Concerning the integration of the phonon collision operator (3.13) with respect to ξ_2, the pole ξ_2^{+} is defined by

$$G_1[e, E', \xi_2^{+}, \varphi_2, \mathbf{\Omega}_q] = E[\boldsymbol{k}^{-}(E', \xi_2^{+}, \varphi_2, q(e, \mathbf{\Omega}_q)\mathbf{\Omega}_q)] - E' + e = 0, \tag{3.24}$$

where we have introduced the phonon energy $e = e_1(\boldsymbol{q})$. The uniqueness of an existing solution to (3.24) with respect to ξ_2 for fixed e, E', φ_2, $\mathbf{\Omega}_q$ as postulated in section 3.1, i.e.,

$$\xi_2^{+} = \zeta_2(e, E', \varphi_2, \mathbf{\Omega}_q), \tag{3.25}$$

leads us to the relevant energy ranges

$$\mathcal{D}^{a}(e, \varphi_2, \mathbf{\Omega}_q) = \left\{ E' \in I^1 \middle| \, \exists \zeta_2 \wedge |\zeta_2| \leq 1 \right\} \tag{3.26}$$

by additionally exploiting the relation $\xi_2 = \cos\vartheta_2$. Hence, the phonon collision operator can be written as

$$C_1^{ph}(f, g_1) = \int_0^{2\pi} d\varphi_2 \int_{I^1} dE' \sigma_1 \kappa_1(g_1, f^-, f'), \tag{3.27}$$

with $f' = f[k(E', \boldsymbol{u}_2)\boldsymbol{u}_2]$ and $f^- = f(\boldsymbol{k}_2^-)$. The functions $\boldsymbol{u}_2 = \boldsymbol{u}[\zeta_2, \varphi_2, q(e, \boldsymbol{\Omega}_q)\boldsymbol{\Omega}_q]$ and $\boldsymbol{k}_2^- = \boldsymbol{k}^-(E', \zeta_2, \varphi_2, \boldsymbol{\Omega}_q)$ are defined by (3.7), (3.15) and evaluated at $\xi_2 = \zeta_2$. The new scattering rate is given by

$$\sigma_1(e, E', \varphi_2, \boldsymbol{\Omega}_q) = \frac{V}{4\pi^3} s_1 \left[|J| \left(\left| \frac{\partial G_1}{\partial \xi_2} \right| \right)^{-1} \right]_{\xi_2 = \zeta_2} \chi_{\mathcal{D}^a}(E') \tag{3.28}$$

with the relations (3.8), (3.24), and (3.26).

Next, we treat the collision operators in the case of optical phonons. This means that we assume a constant phonon energy $e_2(\boldsymbol{q}) = e_0 = \hbar\omega_0$, which simplifies the argument $F_2^{\pm}$ of the delta distribution (3.11) in the collision operator (3.10). Performing the integration with respect to the electron energy E' leads to

$$\begin{aligned} C_2^{el}(f, g_2) = \frac{V}{8\pi^3} \int_0^{2\pi} d\varphi_1 \int_{-1}^{1} d\xi_1 \Big[& |J|^+ s_2^+ \kappa_1(g_2^+, f, f'^+_*)\, \chi_{\mathcal{D}^{o+}}(E) \\ & + |J|^- s_2^- \kappa_2(g_2^-, f, f'^-_*)\, \chi_{\mathcal{D}^{o-}}(E) \Big]. \end{aligned} \tag{3.29}$$

The functions $|J|^{\pm}, s_2^{\pm}, g_2^{\pm}, f'^{\pm}_*$ are evaluated at $E' = E \pm e_0$. The insertion of the characteristic functions $\chi_{\mathcal{D}^{o\pm}}(E)$ of the sets

$$\mathcal{D}^{o+} = [\min(I^1), \max(I^1) - e_0], \quad \mathcal{D}^{o-} = [\min(I^1) + e_0, \max(I^1)] \tag{3.30}$$

ensures that $E \pm e_0 = E' \in I^1$. We introduce the modulus of the phonon wave vector $q^{o\pm} = |\boldsymbol{q}^{\pm}|$ as the relevant variable instead of ξ_1 with the relations

$$q^{o\pm}(E, \xi_1, \varphi_1, \boldsymbol{\Omega}_k) = \left[k^2 + (k'^{\pm})^2 - 2kk'^{\pm}\xi_1 \right]^{\frac{1}{2}} \tag{3.31}$$

for the triangle formed by $\boldsymbol{k}, \boldsymbol{k'}^{\pm}, \boldsymbol{q}^{\pm}$. The vector

$$\boldsymbol{k'}^{\pm}(E, \xi_1, \varphi_1, \boldsymbol{\Omega}_k) = k[E \pm e_0, \boldsymbol{u}(\xi_1, \varphi_1, \boldsymbol{\Omega}_k)]\, \boldsymbol{u}(\xi_1, \varphi_1, \boldsymbol{\Omega}_k) \tag{3.32}$$

corresponds to energies $E' = E \pm e_0$. In section 3.1 we have imposed unique solutions to the energy and momentum relations (3.5) for given parameters E, q, φ_1, and $\boldsymbol{\Omega}_k$. In this case the inverse functions of $q = q^{o\pm}$ with respect to ξ_1 exist, and we denote them with

$$\xi_1 = \zeta_1^{o\pm}(E, q, \varphi_1, \boldsymbol{\Omega}_k). \tag{3.33}$$

If we substitute q for ξ_1 in the integral of (3.29), the collision operator results in

$$C_2^{el}(f, g_2) = \int_0^{2\pi} d\varphi_1 \int_{I^3} dq \left[\sigma_2^+ \kappa_1(g_2^+, f, \hat{f}'^+) + \sigma_2^- \kappa_2(g_2^-, f, \hat{f}'^-) \right]. \tag{3.34}$$

Here $g_2^{\pm} = g_2[q\hat{\boldsymbol{u}}^{\pm}(E, E \pm e_0, \zeta_1^{o\pm}, \varphi_1, \boldsymbol{\Omega}_k)]$ and $\hat{f}'^{\pm} = f(\boldsymbol{k'}^{\pm})$ with the vectors $\hat{\boldsymbol{u}}$ introduced in (3.23) and $\boldsymbol{k'}^{\pm}$ given in (3.32), which are computed at $\xi_1 = \zeta_1^{o\pm}$. The

interval I^3 is defined in such a way that the phonon distribution function $g(\boldsymbol{q})$ vanishes for q-values outside I^3. The equations (3.31), (3.33) and the sets

$$\mathcal{D}^{\pm}(E,\varphi_1,\boldsymbol{\Omega}_k)=\left\{q\in I^3\middle|\,|k-k'^{\pm}(\xi_1=1)|\le q\le k+k'^{\pm}(\xi_1=-1)\right\} \tag{3.35}$$

together with $|J|^{\pm}=|J(E'=E\pm e_0)|$, where (3.8) is used, yield the transition rates

$$\sigma_2^{\pm}(E,q,\varphi_1,\boldsymbol{\Omega}_k)=\frac{V}{8\pi^3}\left[|J|^{\pm}s_2^{\pm}\left(\frac{\partial q^{o\pm}}{\partial\xi_1}\right)^{-1}\right]_{\xi_1=\zeta_1^{o\pm}}\chi_{\mathcal{D}^{\pm}}(q)\chi_{\mathcal{D}^{o\pm}}(E). \tag{3.36}$$

The phonon collision operator for optical phonons can be obtained from the acoustic phonon operator (3.27) by setting $e=e_0$, i.e.,

$$C_2^{ph}(f,g_2)=\int_0^{2\pi}d\varphi_2\int_{I^1}dE'\sigma_2\kappa_1(g_2,\hat{f}^-,\hat{f}') \tag{3.37}$$

with $\hat{f}'=f[k(E',\boldsymbol{u}_2^o)\boldsymbol{u}_2^o]$ and $\hat{f}^-=f(\boldsymbol{k}_2^{o-})$. The functions $\boldsymbol{u}_2^o=\boldsymbol{u}(\zeta_2^o,\varphi_2,\boldsymbol{\Omega}_q)$ and $\boldsymbol{k}_2^{o-}=\boldsymbol{k}^-(E',\zeta_2^o,\varphi_2,\boldsymbol{\Omega}_q)$ are defined by (3.7), (3.15) and evaluated at $\xi_2=\zeta_2^o$, which is determined by

$$G_2(E',\xi_2^+,\varphi_2,q\boldsymbol{\Omega}_q)=0\Rightarrow\xi_2^+=\zeta_2^o(q,E',\varphi_2,\boldsymbol{\Omega}_q). \tag{3.38}$$

The scattering rate results in

$$\sigma_2(q,E',\varphi_2,\boldsymbol{\Omega}_{\boldsymbol{q}})=\frac{V}{4\pi^3}s_2\left[|J|\left(\left|\frac{\partial G_2}{\partial\xi_2}\right|\right)^{-1}\right]_{\xi_2=\zeta_2^o}\chi_{\mathcal{D}^o}(E'), \tag{3.39}$$

with (3.8), (3.24), and the energy ranges

$$\mathcal{D}^o(q,\varphi_2,\boldsymbol{\Omega}_q)=\left\{E'\in I^1\middle|\,\exists\zeta_2^o\wedge|\zeta_2^o|\le 1\right\}. \tag{3.40}$$

It should be pointed out that at this stage we have performed only a reformulation of the collision operators and therefore no approximations have been made.

3.3. Discretization. The following discretization and approximation procedure is similar to the discretizations presented in [17]. Here, we introduce discretizations of the electron energy, the phonon energy, and the modulus of the optical phonon wave vector. Starting with the electron energy variable, we choose $n+1$ values $E_i\in I^1$, which form an arithmetic series $E_i=E_0+i\Delta^1$, $i\in S_1=[0,n]\subset\mathbb{N}$, with the energy distance $\Delta^1>0$. Furthermore, we introduce a partition I_i^1, $i\in S_1$, of the interval I^1, so that E_i belongs to the interior of I_i^1. It is convenient to use equidistant intervals I_i^1 with length Δ^1. Due to the energy conserving transition rates (2.6), the acoustic phonon energy can always be expressed as electron energy difference, $e=E'-E$, in the collision operators (3.20) and (3.27). Thus, the discretization of the electron energy determines the set of phonon energies $e_i=i\Delta^2$, with $\Delta^2=\Delta^1=\Delta$ and $i\in S_2=[1,n]\subset\mathbb{N}$, if we assume that $e>0$. As in the case of electrons, we define a partition I_i^2, $i\in S_2$, of the interval I^2, where e_i is an element of the interior of I_i^2. Again, it is reasonable to assume equidistant intervals I_i^2 with length Δ. This fragmentation of I^1, I^2 is possible if the basic phonon energy range is defined by $I^2=[E_0-\min(I^1),E_n-\min(I^1)]$.

The state of the particles is characterized by a discrete set of energies E_i, e_i and arbitrary unit vectors $\mathbf{\Omega}$. The discretization has been constructed to ensure that $E_i \pm e_j = E_k$ with $k = i \pm j$ if $k = (i \pm j) \in S_1$. Hence, the relevant energy and momentum relations are exactly fulfilled on the level of individual microscopic collision processes. This is the main feature of a conservative semicontinuous model.

In the case of optical phonons, we discretize the modulus q of the wave vector $\boldsymbol{q}$. Since the phonon energy does not depend on the wave vector, the energy momentum relations decouple and we can freely choose discrete values for q. To satisfy the relation $E_i \pm e_0 = E_j$ with discrete energy values E_i, E_j, the phonon energy must be an integer multiple of the discretization length, i.e., $e_0 = \alpha\Delta$, $\alpha \in S_1$. We introduce a partition of I^3 with m subintervals I_i^3, $i \in S_3 = [1, m] \subset \mathbb{N}$, and a set of m values q_i, $i \in S_3$, in such a way that q_i is an element of I_i^3. The length of the interval I_i^3 is denoted with Δ_i^3.

Next, we approximate functions $\Phi(x^j)$, $j = 1, 2, 3$, of the discretized variables $x^1 = E$, $x^2 = e$, $x^3 = q$ by

$$\Phi(x^j) \approx \sum_{i \in S_j} \Theta_{ij}(x^j)\Phi_i^j, \qquad \Phi_i^j = \Phi(x_i^j), \quad j = 1, 2, 3, \tag{3.41}$$

with the shape function $\Theta_{ij}(x^j)$, which is equal to 1 for $x^j \in I_i^j$ and zero otherwise for $j = 1, 2, 3$. The approximation method applied to integrals of functions $\Phi(x^j)$ yields

$$\int_{I^j} \Phi(x^j)dx^j \approx \sum_{i \in S_j} \Delta_i^j \Phi_i^j, \quad j = 1, 2, 3. \tag{3.42}$$

The approximation of functions of electron energy differences, e.g.,

$$\Phi(E - E') = \sum_{i,j \in S_1,\, i>j} \Theta_{i1}(E)\Theta_{j1}(E')\Phi(E_i - E_j) = \sum_{i \in S_1} \sum_{k \in S_2} \Theta_{i1}(E)\Theta_{k2}(e)\Phi(e_k) \tag{3.43}$$

with $e = (E - E') \in I^2$ and $k = (i - j) \in S_2$, constitutes a special case.

3.4. Semicontinuous kinetic equations. We start to discretize the electron equation (2.1) by integrating the equation over the interval I_i^1 with respect to the variable E. The left-hand side can then be approximated by

$$\int_{I_i^1} \left[\frac{\partial f}{\partial t} + \boldsymbol{v}(\boldsymbol{k}) \cdot \nabla_r f\right] dE \approx \Delta \left[\frac{\partial f_i}{\partial t} + \boldsymbol{v}_i \cdot \nabla_r f_i\right] \tag{3.44}$$

with the abbreviations $f_i = f[k(E_i, \mathbf{\Omega}_k)\mathbf{\Omega}_k, \boldsymbol{r}, t]$ and $\boldsymbol{v}_i = \boldsymbol{v}[k(E_i, \mathbf{\Omega}_k)\mathbf{\Omega}_k]$. The application of the approximation procedure (3.41) to the integral of C_1^{el}, (3.20), yields

$$\int_{I_i^1} C_1^{el}(f, g_1)dE \approx \Delta C_{1i}^{el} = \Delta \int_0^{2\pi} d\varphi_1 \sum_{j \in S_1} \left[A_{1ij}^{+} \kappa_1(g_{1ij}^{+}, f_i, f'^{+}_{ij}) + A_{1ij}^{-} \kappa_2(g_{1ij}^{-}, f_i, f'^{-}_{ij})\right] \tag{3.45}$$

for $i \in S_1$ with the abbreviations

$$f'^{\pm}_{ij} = f[k(E_j, \boldsymbol{u}^{\pm}_{1ij})\boldsymbol{u}^{\pm}_{1ij}, \boldsymbol{r}, t], \tag{3.46}$$

$$g^{\pm}_{1ij} = g^{\pm}_1[\boldsymbol{q}^{\pm}(E_i, E_j, \zeta^{\pm}_{1ij}, \varphi_1, \boldsymbol{\Omega}_k), \boldsymbol{r}, t], \tag{3.47}$$

$$A^{\pm}_{1ij} = \Delta\sigma^{\pm}_1(E_i, E_j, \varphi_1, \boldsymbol{\Omega}_k), \tag{3.48}$$

$$\boldsymbol{u}^{\pm}_{1ij} = \boldsymbol{u}(\zeta^{\pm}_{1ij}, \varphi_1, \boldsymbol{\Omega}_k), \quad \zeta^{\pm}_{1ij} = \zeta^{\pm}_1(E_i, E_j, \varphi_1, \boldsymbol{\Omega}_k). \tag{3.49}$$

We use the unit vector $\boldsymbol{u}^{\pm}$ introduced in (3.7), and the function $\zeta^{\pm}_1$ defined in (3.17) is evaluated at E_i, E_j.

Now, we treat the second part of the electron collision operator involving the optical phonons. We integrate the operator (3.34) over I^1_i with respect to the electron energy variable E. The approximation (3.41) of functions of E and q results in

$$\int_{I^1_i} C^{el}_2(f, g_2)dE \approx \Delta C^{el}_{2i} = \Delta \int_0^{2\pi} d\varphi_1 \sum_{j\in S_3} \left[A^{+}_{2ij}\kappa_1(g^{+}_{2ij}, f_i, \hat{f}'^{+}_{ij}) + A^{-}_{2ij}\kappa_2(g^{-}_{2ij}, f_i, \hat{f}'^{-}_{ij})\right] \tag{3.50}$$

for $i \in S_1$ with

$$\hat{f}'^{\pm}_{ij} = f[k(E_{i\pm\alpha}, \boldsymbol{u}^{o\pm}_{1ij})\boldsymbol{u}^{o\pm}_{1ij}, \boldsymbol{r}, t], \tag{3.51}$$

$$g^{\pm}_{2ij} = g^{\pm}_2[q_j\hat{\boldsymbol{u}}^{\pm}(E_i, E_{i\pm\alpha}, \zeta^{o\pm}_{1ij}, \varphi_1, \boldsymbol{\Omega}_k), \boldsymbol{r}, t], \tag{3.52}$$

$$A^{\pm}_{2ij} = \Delta^3_j\sigma^{\pm}_2(E_i, q_j, \varphi_1, \boldsymbol{\Omega}_k), \tag{3.53}$$

$$\boldsymbol{u}^{o\pm}_{1ij} = \boldsymbol{u}(\zeta^{o\pm}_{1ij}, \varphi_1, \boldsymbol{\Omega}_k), \quad \zeta^{o\pm}_{1ij} = \zeta^{o\pm}_1(E_i, q_j, \varphi_1, \boldsymbol{\Omega}_k). \tag{3.54}$$

We point out that $e_0 = \alpha\Delta$ holds as imposed in section 3.3. The function $\zeta^{o\pm}_{1ij}$ has been introduced in (3.33), and concerning the vectors $\boldsymbol{u}$ and $\hat{\boldsymbol{u}}^{\pm}$ we refer to (3.7) and (3.23). Adding up the two scattering operators (3.45), (3.50) and equating this sum with the approximation of the streaming part (3.44) gives a discretized version of the electron Boltzmann equation (2.1).

In the next step, we approximate the integral of the left-hand sides of the phonon equations (2.2). Starting with acoustic phonons, we obtain

$$\int_{I^2_i} \left[\frac{\partial g_1}{\partial t} + \boldsymbol{c}_1(\boldsymbol{q}) \cdot \nabla_r g_1\right] de \approx \Delta \left[\frac{\partial g_{1i}}{\partial t} + \boldsymbol{c}_{1i} \cdot \nabla_r g_{1i}\right] \tag{3.55}$$

for $i \in S_2$ and with $g_{1i} = g[q(e_i, \boldsymbol{\Omega}_q)\boldsymbol{\Omega}_q, \boldsymbol{r}, t]$, $\boldsymbol{c}_{1i} = \boldsymbol{c}_1[q(e_i, \boldsymbol{\Omega}_q)\boldsymbol{\Omega}_q]$. The approximation of the integrated optical phonon streaming operator yields

$$\int_{I^3_i} \frac{\partial g_2}{\partial t} dq \approx \Delta^3_i \frac{\partial g_{2i}}{\partial t} \tag{3.56}$$

for $i \in S_3$ if we use the abbreviation $g_{2i} = g(q_i\boldsymbol{\Omega}_q, \boldsymbol{r}, t)$ and take into account that the mean velocity $\boldsymbol{c}_2(\boldsymbol{q})$ defined in (2.3) vanishes for the Einstein dispersion relation $e_2(\boldsymbol{q}) = e_0$. The collision operators (3.27) and (3.37) are discretized in the same way as the streaming operators above. In the case of acoustic phonons, the approximation (3.41) results in

$$\int_{I^2_i} C^{ph}_1(f, g_1)de \approx \Delta C^{ph}_{1i} = \Delta \int_0^{2\pi} d\varphi_2 \sum_{j\in S_1} B_{1ij}\kappa_1(g_{1i}, f^{-}_{ij}, f'_{ij}) \tag{3.57}$$

for $i \in S_2$ with

$$f'_{ij} = f[k(E_j, \boldsymbol{u}_{2ij})\boldsymbol{u}_{2ij}, \boldsymbol{r}, t], \tag{3.58}$$

$$f^-_{ij} = f[\boldsymbol{k}^-(E_j, \zeta_{2ij}, \varphi_2, q(e_i, \boldsymbol{\Omega}_q)\,\boldsymbol{\Omega}_q), \boldsymbol{r}, t], \tag{3.59}$$

$$B_{1ij} = \Delta\sigma_1(e_i, E_j, \varphi_2, \boldsymbol{\Omega}_q). \tag{3.60}$$

The vectors $\boldsymbol{u}_{2ij} = \boldsymbol{u}(\zeta_{2ij}, \varphi_2, \boldsymbol{\Omega}_q)$, $\boldsymbol{k}^-$ are defined in (3.7), (3.15), and the function $\zeta_{2ij} = \zeta_2(e_i, E_j, \varphi_2, \boldsymbol{\Omega}_q)$ is introduced in (3.25). The discretization of the modulus of the phonon wave vector and the approximation of the optical phonon collision operator yield

$$\int_{I_i^3} C_2^{ph}(f, g_2)dq \approx \Delta_i^3 C_{2i}^{ph} = \Delta_i^3 \int_0^{2\pi} d\varphi_2 \sum_{j\in S_1} B_{2ij}\kappa_1(g_{2i}, \hat{f}^-_{ij}, \hat{f}'_{ij}) \tag{3.61}$$

for $i \in S_3$ with

$$\hat{f}'_{ij} = f[k(E_j, \boldsymbol{u}^o_{2ij})\boldsymbol{u}^o_{2ij}, \boldsymbol{r}, t], \tag{3.62}$$

$$\hat{f}^-_{ij} = f[\boldsymbol{k}^-(E_j, \zeta^o_{2ij}, \varphi_2, q_i\,\boldsymbol{\Omega}_q), \boldsymbol{r}, t], \tag{3.63}$$

$$B_{2ij} = \Delta\sigma_2(q_i, E_j, \varphi_2, \boldsymbol{\Omega}_q). \tag{3.64}$$

We define the unit vector $\boldsymbol{u}^o_{2ij} = \boldsymbol{u}(\zeta^o_{2ij}, \varphi_2, \boldsymbol{\Omega}_q)$ by using (3.7) and the function $\zeta^o_{2ij} = \zeta^o_2(q_i, E_j, \varphi_2, \boldsymbol{\Omega_q})$, where we evaluate (3.38) at $q = q_i$, $E' = E_j$. The phonon energy is given by $e_0 = \alpha\Delta$.

Equating the approximations (3.55) and (3.56) of the left-hand side of (2.2) to the collision operators (3.57) and (3.61) yields discretized phonon Boltzmann equations. Together with the streaming operator (3.44) and the collision operators of the electron equation, (3.45) and (3.50), we obtain the semicontinuous BBP equations

$$\frac{\partial f_i}{\partial t} + \boldsymbol{v}_i \cdot \nabla_r f_i = C_{1i}^{el} + C_{2i}^{el}, \quad i \in S_1, \tag{3.65}$$

$$\frac{\partial g_{1i}}{\partial t} + \boldsymbol{c}_{1i} \cdot \nabla_r g_{1i} = C_{1i}^{ph}, \quad i \in S_2, \tag{3.66}$$

$$\frac{\partial g_{2i}}{\partial t} = C_{2i}^{ph}, \quad i \in S_3. \tag{3.67}$$

3.5. Macroscopic quantities. For the introduction of macroscopic quantities in the semicontinuous model, it is useful to define the vector function

$$\boldsymbol{\Phi} = [\phi_0^1(\boldsymbol{\Omega}), \dots, \phi_n^1(\boldsymbol{\Omega}), \phi_1^2(\boldsymbol{\Omega}), \dots, \phi_n^2(\boldsymbol{\Omega}), \phi_1^3(\boldsymbol{\Omega}), \dots, \phi_m^3(\boldsymbol{\Omega})] \tag{3.68}$$

for $\boldsymbol{\Omega} \in \mathbb{S}^2$. Its components are $\phi_i^1(\boldsymbol{\Omega}) = \Phi^1[k(E_i, \boldsymbol{\Omega})\boldsymbol{\Omega}]$, $\phi_i^2(\boldsymbol{\Omega}) = \Phi^2[q(e_i, \boldsymbol{\Omega})\boldsymbol{\Omega}]$, and $\phi_i^3(\boldsymbol{\Omega}) = \Phi^3(q_i\boldsymbol{\Omega})$, where $\Phi^1(\boldsymbol{k})$, $\Phi^2(\boldsymbol{q})$, and $\Phi^3(\boldsymbol{q})$ are arbitrary functions of $\boldsymbol{k}$, $\boldsymbol{q} \in \mathcal{B}$. In addition, we introduce the vector

$$\boldsymbol{f} = (f_0^1, \dots, f_n^1, f_1^2, \dots, f_n^2, f_1^3, \dots, f_m^3) = (f_0, \dots, f_n, g_{11}, \dots, g_{1n}, g_{21}, \dots, g_{2m}), \tag{3.69}$$

consisting of the discretized distribution functions. Then, we consider the functional

$$\langle \boldsymbol{\Phi}, \boldsymbol{f} \rangle = \sum_{i=1}^{3} \sum_{j\in S_i} \Delta_j^i \int_{\mathbb{S}^2} d\boldsymbol{\Omega} D_j^i(\boldsymbol{\Omega})\phi_j^i(\boldsymbol{\Omega})f_j^i(\boldsymbol{\Omega}), \tag{3.70}$$

where $\Delta_j^i = \Delta$ for $i = 1, 2$, $j \in S_i$, holds, and we use

$$D_i^1(\mathbf{\Omega}_k) = D^1(E_i, \mathbf{\Omega}_k) = \frac{1}{4\pi^3} \left| \det \left[\frac{\partial(k_1, k_2, k_3)}{\partial(E, \xi_k, \varphi_k)} \right] \right|_{E=E_i}, \quad i \in S_1, \tag{3.71}$$

$$D_i^2(\mathbf{\Omega}_q) = D^2(e_i, \mathbf{\Omega}_q) = \frac{1}{8\pi^3} \left| \det \left[\frac{\partial(q_1, q_2, q_3)}{\partial(e, \xi_q, \varphi_q)} \right] \right|_{e=e_i}, \quad i \in S_2, \tag{3.72}$$

$$D_i^3(\mathbf{\Omega}_q) = D^3(q_i, \mathbf{\Omega}_q) = \frac{1}{8\pi^3} q_i^2, \quad i \in S_3. \tag{3.73}$$

The angular variables $\xi_k = \cos\vartheta_k, \varphi_k$ and $\xi_q = \cos\vartheta_q, \varphi_q$ represent the unit vectors $\mathbf{\Omega}_k$ and $\mathbf{\Omega}_q$ as defined in (3.2). With the latter functional we can express the macroscopic quantities in the semicontinuous formalism. The densities of the electron particle number, the total quasi momentum, and the total energy are given by

$$n^{el}(\boldsymbol{r}, t) = \langle \mathbf{\Phi}_n, \boldsymbol{f} \rangle, \quad \Phi_n^1 = 1, \Phi_n^2 = \Phi_n^3 = 0, \tag{3.74}$$

$$k_l^{tot}(\boldsymbol{r}, t) = \langle \mathbf{\Phi}_{k_l}, \boldsymbol{f} \rangle, \quad \Phi_{k_l}^1 = k_l, \Phi_{k_l}^2 = \Phi_{k_l}^3 = q_l, \ l = 1, 2, 3, \tag{3.75}$$

$$e^{tot}(\boldsymbol{r}, t) = \langle \mathbf{\Phi}_e, \boldsymbol{f} \rangle, \quad \Phi_e^1 = E(\boldsymbol{k}), \Phi_e^2 = e(\boldsymbol{q}), \Phi_e^3 = e_0. \tag{3.76}$$

By defining the vectors

$$\boldsymbol{\nu}_\beta = (v_{0\beta} f_0, \dots, v_{n\beta} f_n, c_{11\beta} g_{11}, \dots, c_{1n\beta} g_{1n}, c_{21\beta} g_{21}, \dots, c_{2m\beta} g_{2m}) \tag{3.77}$$

for $\beta = 1, 2, 3$, we obtain the corresponding current densities of the particle number, the total quasi momentum, and the total energy, $u_\beta^{el}(\boldsymbol{r}, t) = \langle \mathbf{\Phi}_n, \boldsymbol{\nu}_\beta \rangle$, $K_{\beta\gamma}^{tot}(\boldsymbol{r}, t) = \langle \mathbf{\Phi}_{k_\beta}, \boldsymbol{\nu}_\gamma \rangle$, $Q_\beta^{tot}(\boldsymbol{r}, t) = \langle \mathbf{\Phi}_e, \boldsymbol{\nu}_\beta \rangle$ with $\beta = 1, 2, 3$ and $\gamma = 1, 2, 3$.

4. Conservation laws, stability, and equilibrium of the discretized model. Here, we briefly summarize the interesting mathematical aspects of the semicontinuous BBP equations and sketch their derivation, while the rigorous proofs are deferred to a subsequent paper. To this end, we introduce the vector $\boldsymbol{J}$ formed by the semicontinuous collision operators (3.45), (3.50), (3.57), and (3.61), i.e., $J_i^1 = C_{1i}^{el} + C_{2i}^{el}$ for $i \in S_1$, $J_i^2 = C_{1i}^{ph}$ for $i \in S_2$, and $J_i^3 = C_{2i}^{ph}$ for $i \in S_3$, and define the functional

$$\langle \mathbf{\Phi}, \boldsymbol{J} \rangle = \sum_{i=1}^{3} \sum_{j \in S_i} \Delta_j^i \int_{\mathbb{S}^2} d\mathbf{\Omega} D_j^i(\mathbf{\Omega}) \phi_j^i(\mathbf{\Omega}) J_j^i(\mathbf{\Omega}). \tag{4.1}$$

In addition, we consider the space of collisional invariants $\mathcal{C}$, which contains all triples $[\Phi^1(\boldsymbol{k}), \Phi^2(\boldsymbol{q}), \Phi^3(\boldsymbol{q})]$ satisfying $\langle \mathbf{\Phi}, \boldsymbol{J} \rangle = 0$. Moreover, we introduce the functions

$$\Phi_H^1(\boldsymbol{k}) = \log\left[\frac{f(\boldsymbol{k})}{1 - f(\boldsymbol{k})}\right], \ \Phi_H^2(\boldsymbol{q}) = \log\left[\frac{g_1(\boldsymbol{q})}{1 + g_1(\boldsymbol{q})}\right], \ \Phi_H^3(\boldsymbol{q}) = \log\left[\frac{g_2(\boldsymbol{q})}{1 + g_2(\boldsymbol{q})}\right]. \tag{4.2}$$

With the definitions above and the sets

$$S_i^+(\varphi, \mathbf{\Omega}) = \left\{ j \in S_1 \middle| E_j \in \mathcal{D}^{a+}(E_i, \varphi, \mathbf{\Omega}) \right\}, \tag{4.3}$$

$$S_i^q(\varphi, \mathbf{\Omega}) = \left\{ j \in S_3 \middle| q_j \in \mathcal{D}^{+}(E_i, \varphi, \mathbf{\Omega}) \right\}, \tag{4.4}$$

we can state the basic theorem.

THEOREM 4.1. *The presented semicontinuous BBP equations* (3.65)–(3.67) *have the following properties.*

1. *They provide the continuity equations*

$$\begin{aligned}
&\frac{\partial n^{el}(\boldsymbol{r},t)}{\partial t} + \nabla_r \cdot \boldsymbol{u}^{el}(\boldsymbol{r},t) = 0,\\
&\frac{\partial \boldsymbol{k}^{tot}(\boldsymbol{r},t)}{\partial t} + \nabla_r \cdot \mathrm{K}^{tot}(\boldsymbol{r},t) = \boldsymbol{0},\\
&\frac{\partial e^{tot}(\boldsymbol{r},t)}{\partial t} + \nabla_r \cdot \boldsymbol{Q}^{tot}(\boldsymbol{r},t) = 0
\end{aligned} \tag{4.5}$$

for the macroscopic quantities defined in section (3.5).

2. *In the equilibrium state, the relation*

$$\boldsymbol{J} = 0 \Leftrightarrow [\Phi_H^1(\boldsymbol{k}), \Phi_H^2(\boldsymbol{q}), \Phi_H^3(\boldsymbol{q})] \in \mathcal{C} \tag{4.6}$$

holds, and these conditions are equivalent to

$$\begin{aligned}
f_i(\boldsymbol{\Omega}) &= \left[\exp\left(\frac{E_i - \mu}{k_B T}\right) + 1\right]^{-1},\\
g_{1ij}^{+}(\varphi_1, \boldsymbol{\Omega}) &= \left[\exp\left(\frac{e_{j-i}}{k_B T}\right) - 1\right]^{-1},\\
g_{2ik}^{+}(\varphi_1, \boldsymbol{\Omega}) &= \left[\exp\left(\frac{e_0}{k_B T}\right) - 1\right]^{-1}
\end{aligned} \tag{4.7}$$

with $i \in S_1$, $j \in S_i^+(\varphi_1, \boldsymbol{\Omega})$, $k \in S_i^q(\varphi_1, \boldsymbol{\Omega})$, $\varphi_1 \in [0, 2\pi]$, $\boldsymbol{\Omega} \in \mathbb{S}^2$, *if the total quasi momentum* $\boldsymbol{k}^{tot}(\boldsymbol{r},t)$ *vanishes identically.*

3. *A Lyapunov functional for the considered kinetic equations is given by*

$$\begin{aligned}
H &= \sum_{i=1}^{3} \sum_{j \in S_i} \Delta_j^i \int_{\mathbb{S}^2} d\boldsymbol{\Omega}\, D_j^i(\boldsymbol{\Omega}) h_j^i(\boldsymbol{\Omega}),\\
h_i^1 &= f_i \log(f_i) + (1 - f_i)\log(1 - f_i),\\
h_i^{j+1} &= g_{ji}\log(g_{ji}) - (1 + g_{ji})\log(1 + g_{ji}), \quad j = 1, 2,
\end{aligned} \tag{4.8}$$

with the distribution functions $f_i = f[k(E_i, \boldsymbol{\Omega})\boldsymbol{\Omega}]$, $g_{1i} = g_1[q(e_i, \boldsymbol{\Omega})\boldsymbol{\Omega}]$, $g_{2i} = g_2(q_i\boldsymbol{\Omega})$, *and* $D_j^i(\boldsymbol{\Omega})$ *defined in* (3.71)–(3.73).

The integration of the continuity equations (4.5) over the crystal volume V with respect to the local variable $\boldsymbol{r}$ shows that the electron particle number, the total quasi momentum, and the total energy are preserved. It should be mentioned that the conservation of the total quasi momentum is a consequence of neglecting umklapp processes in the scattering operators.

From (4.7), we infer that the equilibrium solutions of the model equations are discretized versions of the Fermi–Dirac distribution concerning the electrons and Bose–Einstein distribution for the phonons. The restriction of the indices j, k of the equilibrium phonon distributions results from the fact that the scattering rates A_{1ij}^{+}, A_{2ik}^{+} appearing in the electron collision operators (3.45) and (3.50) vanish for $j \notin S_i^+(\varphi_1, \boldsymbol{\Omega})$ and $k \notin S_i^q(\varphi_1, \boldsymbol{\Omega})$.

The Boltzmann H-functional presented in Theorem 4.1 satisfies the relations

$$\frac{dH}{dt} \le 0, \qquad H - H_* \ge 0, \tag{4.9}$$

where H_* denotes the functional (4.8) evaluated with the equilibrium distributions (4.7) and the equality signs hold if and only if the system is in the equilibrium state.

4.1. Sketch of the proof. In order to derive the mathematical properties presented in Theorem 4.1, we reformulate the functional (4.1) by exploiting symmetries of the integrated semicontinuous collision operators (3.45), (3.50), (3.57), and (3.61). By introducing the abbreviations $\Phi_i^1 = \Phi^1[k(E_i, \boldsymbol{\Omega}_k)\boldsymbol{\Omega}_k]$, $\Phi'^1_{ij} = \Phi^1[k(E_j, \boldsymbol{u}^+_{1ij})\boldsymbol{u}^+_{1ij}]$, and $\hat{\Phi}'^1_{ij} = \Phi^1[k(E_{i+\alpha}, \boldsymbol{u}^{o+}_{1ij})\boldsymbol{u}^{o+}_{1ij}]$, as well as $\Phi^2_{ij} = \Phi^2[\boldsymbol{q}^+(E_i, E_j, \zeta^+_{1ij}, \varphi_1, \boldsymbol{\Omega}_k)]$ and $\Phi^3_{ij} = \Phi^3[q_j\hat{\boldsymbol{u}}^+(E_i, E_{i+\alpha}, \zeta^{o+}_{1ij}, \varphi_1, \boldsymbol{\Omega}_k)]$, this reformulation results in

$$\begin{aligned}\langle \boldsymbol{\Phi}, \boldsymbol{J}\rangle = \Delta \int_{\mathbb{S}^2} d\boldsymbol{\Omega}_k \int_0^{2\pi} d\varphi_1 \sum_{i\in S_1}\sum_{j\in S_1} \left[\Phi_i^1 - \Phi'^1_{ij} + \Phi^2_{ij}\right] D_i^1 A^+_{1ij}\kappa_1(g^+_{1ij}, f_i, f'^+_{ij}) \\ + \Delta \int_{\mathbb{S}^2} d\boldsymbol{\Omega}_k \int_0^{2\pi} d\varphi_1 \sum_{i\in S_1}\sum_{j\in S_3} \left[\Phi_i^1 - \hat{\Phi}'^1_{ij} + \Phi^3_{ij}\right] D_i^1 A^+_{2ij}\kappa_1(g^+_{2ij}, f_i, \hat{f}'^+_{ij}).\end{aligned} \tag{4.10}$$

Starting from this expression, we can show that the macroscopic quantities (3.74)–(3.76) are collisional invariants, and therefore, we can state the continuity equations (4.5). The derivation of (4.10) also allows us to prove the equivalence (4.6). Expanding the triple (4.2) evaluated at equilibrium $[\Phi_H^{1*}, \Phi_H^{2*}, \Phi_H^{3*}]$ in terms of the natural basis $\{\Phi^i_n, \Phi^i_{k_1}, \Phi^i_{k_2}, \Phi^i_{k_3}, \Phi^i_e\}$, $i = 1, 2, 3$, of the space of collisional invariants $\mathcal{C}$, leads us to the equilibrium distribution functions (4.7). Moreover, we find the relations (4.9) by inserting (4.2) into (4.10) and by considering the statement (4.6).

5. Numerical results. In this section, we present numerical results concerning the relaxation of an interacting hot-electron hot-phonon system in GaAs. More precisely, we investigate the temporal evolution of the distribution functions of conduction band electrons, longitudinal acoustic phonons, and polar optical phonons during and after a photo-excitation with a laser pulse [5, 11].

We consider spherically symmetric dispersion relations and scattering probabilities, i.e., $E(\boldsymbol{k}) = E(|\boldsymbol{k}|)$, $e_1(\boldsymbol{q}) = e_1(|\boldsymbol{q}|)$, and $s_i(\boldsymbol{q}) = s_i(|\boldsymbol{q}|)$. This implies that the distribution functions $f_i(\boldsymbol{\Omega}_k, \boldsymbol{r}, t)$, $i \in S_1$, and $g_{ij}(\boldsymbol{\Omega}_q, \boldsymbol{r}, t)$, $i = 1, 2$, $j \in S_{i+1}$, do not depend on the angular variables $\boldsymbol{\Omega}_k$, $\boldsymbol{\Omega}_q$ during the time evolution determined by the semicontinuous BBP equations (3.65)–(3.67) if the initial distributions $f_i^I(\boldsymbol{\Omega}_k, \boldsymbol{r})$, $g^I_{ij}(\boldsymbol{\Omega}_q, \boldsymbol{r})$ are independent of $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Omega}_q$. Consequently, the integrations with respect to φ_1 and φ_2 in the scattering operators (3.45), (3.50) and (3.57), (3.61) contribute only the trivial factor 2π to the corresponding scattering rates. When we further restrict our attention to space-homogeneous problems, the derived model equations (3.65)–(3.67) simplify to a coupled system of first order ordinary differential equations, which can be solved with standard numerical methods.

Since we choose excitation energies that are not too high, the relevant $\boldsymbol{k}$-region of the first Brillouin zone can be restricted to a small area near the Γ-point where the conduction band shows an absolute minimum. Hence, we use a parabolic electron dispersion relation $E(\boldsymbol{k}) = \hbar^2|\boldsymbol{k}|^2/2m^*$ with the effective mass m^* and apply the Debey model to the acoustic phonons, i.e., $e_1(\boldsymbol{q}) = c|\boldsymbol{q}|$, where c denotes the sound velocity (see [1]). The polar optical phonons are treated in the Einstein approximation, i.e., $\hbar\omega(\boldsymbol{q}) = e_0$ as introduced in section 3. The scattering rate corresponding to the electron-phonon interaction is given by (2.6) with

$$s_1(\boldsymbol{q}) = \frac{2\pi}{\hbar}\frac{\hbar^2\Xi_1^2 q^2}{2V\rho e_1(\boldsymbol{q})}, \tag{5.1}$$

$$s_2(\boldsymbol{q}) = \frac{2\pi}{\hbar}\frac{\hbar^2\varepsilon_0^2}{2V\gamma e_0}\frac{1}{q^2}, \qquad \frac{1}{\gamma} = \frac{e_0^2}{\hbar^2}\left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0}\right) \tag{5.2}$$

for acoustic and optical phonons. Here, ρ denotes the mass density, ε_0 denotes the elementary charge, and the quantities ϵ_∞, ϵ_0 represent the high-frequency and electrostatic dielectric functions. The interaction with acoustic phonons is modeled as first order scattering [4] with the deformation potential Ξ_1. Concerning the interaction with the polar optical phonons, screening effects [4] are not taken into account. The values of all relevant parameters are taken from [11] and summarized in Table 5.1.

TABLE 5.1
Material parameters of GaAs used in the calculations.

Effective electron mass:	m^*/m_0	0.063
Mass density:	ρ_m (g/cm^3)	5.32
Sound velocity:	c (km/s)	5.12
Optical phonon energy:	e_0 (meV)	35.5
Static dielectric constant:	ϵ_0	12.9
High-frequency dielectric constant:	ϵ_∞	10.9
Acoustic deformation potential:	Ξ_1 (eV)	7.0

As an initial condition we consider a thermal equilibrium at 77 K. Therefore, the phonons are governed by Bose–Einstein distributions with the same temperature, and we assume that the conduction band is empty at the beginning. As soon as the laser pulse sets in, electrons from the valence bands are transferred to the conduction band. We suppose that the transfer rate of electrons with energy E_i is given by

$$\frac{dn_i^{el}(t)}{dt} = \alpha_n \left\{ \left[1 + \left(\frac{E_i - E_*}{\alpha_E} \right)^2 \right] \cosh\left(\frac{t - t_0}{\alpha_t} \right) \right\}^{-1}, \tag{5.3}$$

where $n_i^{el}(t)\Delta^1$ represents the particle density of electrons with energies in the interval I_i^1, E_* denotes the averaged injection energy, and at $t = t_0$, the maximum of the transfer rate is reached. Furthermore, the parameter α_E characterizes the energetic and α_t the temporal width of the laser pulse, while α_n determines the total number of excited electrons n^{el} or, equivalently, the intensity of the laser. We use $E_* = 250$ meV, $\alpha_E = 10$ meV, $\alpha_t = 0.305$ ps, where the energy scale is fixed in such a way that $E = 0$ at the conduction band minimum. The parameter t_0 is chosen high enough so that we can neglect the contributions of (5.3) for negative times t, and α_n is determined by fixing the total electron density to $n^{el} = 5 \times 10^{15}\text{cm}^{-3}$. The increase of the electron density through the photo-excitation is modeled by an appropriate source term on the right-hand side of (3.65) according to the rate (5.3).

It should be pointed out that the collision operators (3.45), (3.50), (3.57), and (3.61) take into account only normal processes caused by the electron-phonon interaction. Since the electron density is very small, the electron-electron interaction plays a minor role and is neglected. Due to the selection rules of the phonon-phonon interaction, the only important interaction process for the present problem is the decay of optical phonons into two acoustic phonons and vice versa (see [21]). The energy of the optical phonons is very high compared to the electronically active acoustic phonons. Hence, the acoustic phonons resulting from a decay of an optical phonon do not contribute to the electron-phonon interaction for the considered electron energy range. The collision operator for the phonon-phonon interaction can then be treated in the relaxation time approximation (see [11]), i.e.,

$$\left[\frac{\partial g_{2i}}{\partial t} \right]_{pp} = C_2^{pp} = -\frac{g_{2i} - g_{*2}(e_0, T_0)}{\tau_0} \tag{5.4}$$

with the relaxation time τ_0 and the Bose–Einstein distribution $g_{*2}(e_0, T_0)$. If the temperature T_0 is fixed, we realize the coupling of the optical phonon system to a heat bath, represented by acoustic phonons in equilibrium at T_0, which do not affect the electrons. The operator (5.4) is added to the right-hand side of (3.61), and we set $\tau = 3.5$ ps, $T_0 = 77$ K.

Regarding the discretization parameters, we choose 800 discrete electron energy values between $E_0 = 0.1$ meV and $E_{799} = 400.1$ meV with the constant stepsize $\Delta = 0.5$ meV. Since the optical phonon energy is given by $e_0 = 35$ meV, we take $\alpha = 71$. According to the choice of the electron energies, we obtain 11 relevant values for the acoustic phonon energies in the range from $e_1 = 0.5$ meV to $e_{11} = 5.5$ meV. Furthermore, we use an equidistant discretization for the modulus of the optical phonon wave vector q with a discretization length $\Delta^3 = 0.2 \times 10^5$ cm^{-1}. The 800 discrete q-values belong to the interval $[1.1, 160.9] \times 10^5$ cm^{-1}.

The integration of the model equations (3.65)–(3.67) supplemented by the terms for the laser excitation and the phonon-phonon interaction is performed with a standard solver based on an explicit Runge–Kutta scheme [3]. The results of the calculations are discussed in the following.

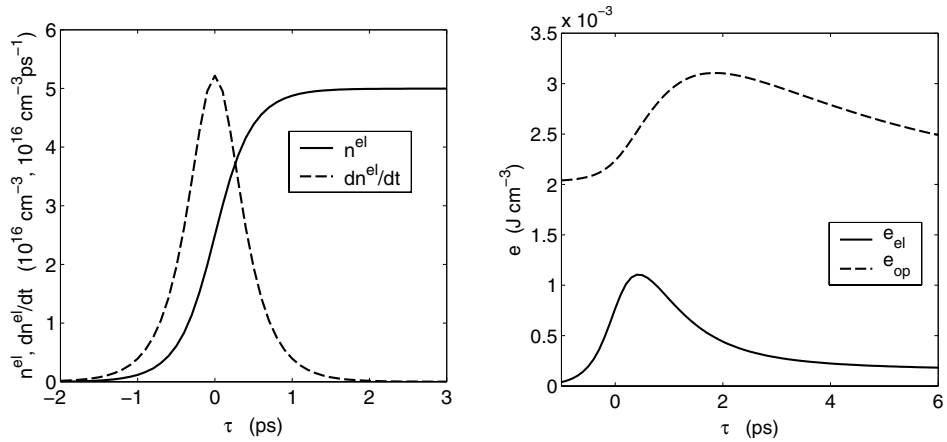


FIG. 5.1. *Electron density and its time derivative as a function of time on the left-hand side. Time evolution of the electron energy and the optical phonon energy in the right-hand plot.*

The left-hand plot of Figure 5.1 depicts the time dependence of the electron particle number and the transfer rate due to the laser photo-excitation. We observe that the electron density in the conduction band tends to the constant value $n^{el} = 5 \times 10^{15}$cm^{-3}, as soon as the laser intensity vanishes. Obviously, the electron density is conserved during the subsequent relaxation process. The time evolution of the energy densities of electrons and optical phonons is shown in the right-hand plot of Figure 5.1. For $t < 0$, the electron energy rises in concert with the laser intensity while the energy of the optical phonons remains largely unaffected. Although the laser pulse then fades away, the electron energy continues to increase. At the same time the emission of optical phonons transfers energy to the phonon system. For times $t \gtrsim 3$, both electrons and optical phonons lose energy to the heat bath of acoustic phonons.

Our semicontinuous model provides further insight into the details of the relaxation process. The left-hand plot of Figure 5.2 shows n_i^{el}, the electron particle density per energy interval Δ according to the energies E_i for different times $\tau = t - t_0$. The mean energy E_* of the photo-excited electrons is bigger than that of the equilib-

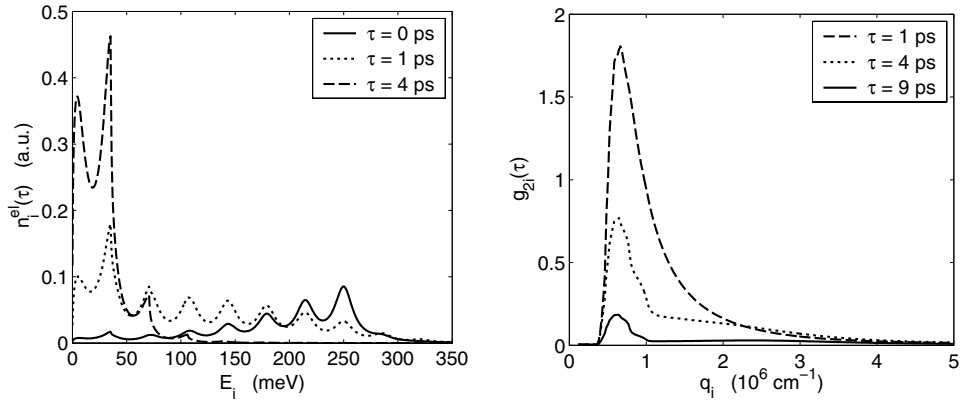


FIG. 5.2. *Particle density of electrons with energies E_i (left-hand plot) and phonon distribution function (right-hand plot) for different times during and after the laser excitation.*

rium distribution with the temperature of the phonon system. Therefore, we observe a relaxation of the hot electron gas accompanied by the emission of mainly optical phonons in accord with the much stronger interaction of electrons with polar optical phonons compared to that with acoustic phonons. Since the emission of an optical phonon always reduces the electron energy by the same constant phonon energy e_0, the electron density as a function of E_i shows a quasi-periodic structure with period e_0. The increase of optical phonons strongly depends on the modulus of the wave vector and leads to far-from-equilibrium distributions as can be seen in the right-hand plot of Figure 5.2. The dependence of $g_{2i}(\tau)$ on q_i is mainly determined by the function $s_2(q)$ and the sets $D_q^{\pm}(E, \varphi_1, \mathbf{\Omega}_k)$ defined in (3.35), which reflect the conservation laws for the involved collision processes.

Figure 5.3 displays the optical phonon distribution function for fixed values q_i and presents a comparison with data obtained from a Monte Carlo simulation [11]. Initially, the distribution functions increase due to the emission of phonons. Af-

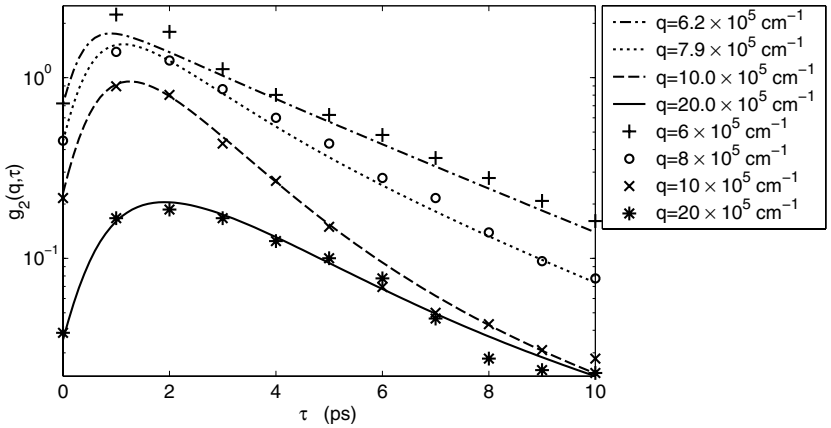


FIG. 5.3. *Distribution functions of the polar optical phonons as a function of time for different moduli of the wave vector q. The lines represent the results from the semicontinuous model equations, and the markers $+, \circ, \times, *$ display the Monte Carlo data taken from* [11].

terward, they decrease due to the phonon-phonon interaction and the reabsorption of polar optical phonons. The distribution functions for $q_i \approx 6 \times 10^5 \mathrm{cm}^{-1}$, $q_i \approx 20 \times 10^5 \mathrm{cm}^{-1}$ show an exponential decay with the time constant τ_0 as a result of the relaxation time approximation of the phonon-phonon interaction. In the case of $q_i \approx 10 \times 10^5 \mathrm{cm}^{-1}$ the reabsorption is very efficient and gives rise to a reinforced decrease of $g_{2i}(\tau)$ as compared to that for smaller and higher q-values. A more detailed discussion of the Monte Carlo results can be found in [11]. The results of our calculation are in good agreement with the Monte Carlo data. The dominant kinetic effect, i.e., the strong reabsorption of polar optical phonons, shows up in both treatments.

As a consequence of the weak interaction of electrons with acoustic phonons, the influence of such phonons on the electrons is negligible in the considered time interval. Taking into account the acoustic phonons, however, prevents the occurrence of decoupling effects of the kinetic equations (2.1) and (2.2) as treated in [12].

To summarize, we find that the semicontinuous model equations are able to accurately describe the kinetics of far-from-equilibrium systems. Numerical results on relaxation processes are obtained by standard methods. The calculations can be performed on a PC with a small consumption of CPU-time. We find that the physical conservation laws are satisfied within numerical roundoff errors. Our formulation is quite general regarding dispersion relations and flexible concerning the modeling of collision processes. It can be extended, e.g., by taking into account electron-electron interaction, phonon-phonon interaction, or impurity scattering.

REFERENCES

[1] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing, Fort Worth, TX, 1976.

[2] N. Bellomo and R. Gatignol, eds., *Lecture Notes on the Discretization of the Boltzmann Equation*, Series on Advances in Mathematics for Applied Sciences 63, World Scientific, River Edge, NJ, 2003.

[3] D. R. Dormand and P. J. Prince, *A family of embedded Runge–Kutta formulae*, J. Comput. Appl. Math., 6 (1980), pp. 19–26.

[4] D. K. Ferry, *Semiconductors*, MacMillan, New York, 1991.

[5] E. D. Grann, K. T. Tsen, and D. K. Ferry, *Nonequilibrium phonon dynamics and electron distribution functions in InP and InAs*, Phys. Rev. B, 53 (1996), pp. 9847–9851.

[6] C. Jacoboni and L. Reggiani, *The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials*, Rev. Modern Phys., 55 (1983), pp. 645–705.

[7] A. Jüngel, *Quasi-hydrodynamic Semiconductor Equations*, Birkhäuser, Basel, Switzerland, 2001.

[8] W. Koller, F. Hanser, and F. Schürrer, *A semi-continuous extended kinetic model*, J. Phys. A, 33 (2000), pp. 3417–3430.

[9] W. Koller and F. Schürrer, *P_N approximation of the nonlinear semi-discrete Boltzmann equation*, Transport Theory Statist. Phys., 30 (2001), pp. 471–489.

[10] W. Koller and F. Schürrer, *Conservative solution methods for extended Boltzmann equations*, Riv. Mat. Univ. Parma (6), 4* (2001), pp. 109–169.

[11] P. Lugli, P. Bordone, L. Reggiani, M. Rieger, P. Kocevar, and S. M. Goodnick, *Monte Carlo studies of nonequilibrium phonon effects in polar semiconductors and quantum wells.* I. *Laser photoexcitation*, Phys. Rev. B, 39 (1989), pp. 7852–7865.

[12] A. Majorana, *Space homogeneous solutions of the Boltzmann equation describing electron-phonon interactions in semiconductors*, Transport Theory Statist. Phys., 20 (1991), pp. 261–279.

[13] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, Vienna, Austria, 1990.

[14] L. Pareschi and G. Russo, *Numerical solution of the Boltzmann equation* I: *Spectrally accurate approximation of the collision operator*, SIAM J. Numer. Anal., 37 (2000), pp. 1217–1245.

[15] L. Preziosi, *The semicontinuous Boltzmann equation for gas mixtures*, Math. Models Methods Appl. Sci., 3 (1993), pp. 665–680.
[16] L. Preziosi and E. Longo, *On a conservative polar discretization of the Boltzmann equation*, Japan J. Indust. Appl. Math., 14 (1997), pp. 399–435.
[17] L. Preziosi and L. Rondoni, *Conservative energy discretization of the Boltzmann operator*, Quart. Appl. Math., 57 (1999), pp. 699–721.
[18] M. Rieger, P. Kocevar, P. Lugli, P. Bordone, L. Reggiani, and S. M. Goodnick, *Monte Carlo studies of nonequilibrium phonon effects in polar semiconductors and quantum wells.* II. *Non-ohmic transport in n-type gallium arsenide*, Phys. Rev. B, 39 (1989), pp. 7866–7875.
[19] A. Rossani, *Generalized kinetic theory of electrons and phonons*, Phys. A, 305 (2002), pp. 323–329.
[20] P. Supancic, U. Hohenester, P. Kocevar, D. Snoke, R. M. Hannak, and W. W. Rühle, *Transport analysis of the thermalization and energy relaxation of photoexcited hot electrons in Ge-doped GaAs*, Phys. Rev. B, 53 (1996), pp. 7785–7791.
[21] J. M. Ziman, *Electrons and Phonons*, Oxford University Press, London, 1960.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1477–1508

STEREOGRAPHIC COMBING A PORCUPINE OR STUDIES ON DIRECTION DIFFUSION IN IMAGE PROCESSING*

NIR A. SOCHEN†, CHEN SAGIV†, AND RON KIMMEL‡

Abstract. This paper addresses the problem of feature enhancement in noisy images when the feature is known to be constrained to a manifold. As an example, we approach the direction denoising problem in a general dimension via the geometric Beltrami framework for image processing. The spatial-direction space is a fiber bundle in which the spatial part is the base manifold and the direction space is the fiber. The feature (direction) field is represented accordingly as a section of the spatial-feature fiber bundle. The resulting Beltrami flow is a selective smoothing process that respects the bundle's structure, i.e., the feature constraint. Direction diffusion is treated as a canonical example of a non-Euclidean feature space. The structures of the fiber spaces of interest in this paper are the unit circle S^1, the unit sphere S^2, and the unit hypersphere S^n. Applications to color analysis are discussed, and numerical experiments demonstrate again the benefits of the Beltrami framework in comparison to other feature enhancement schemes for nontrivial geometries in image processing.

Key words. anisotropic diffusion, constrained optimization, orientation diffusion, Beltrami framework

AMS subject classifications. 58J35, 58J90

DOI. 10.1137/S0036139902415518

1. Introduction.

Many objects of low-level vision are vector fields of various types. This is the case for gray-value images, color images, movies, 3D (three-dimensional) volumetric images, and disparity in stereo vision, to name just a few examples. These vector fields are traditionally considered as taking values in $\mathbb{R}^n$. Operations on these fields such as denoising, enhancement, sharpening, and segmentation are done using a variety of algorithms. Several types of vector fields are constrained in a nontrivial way. When the constraint can be expressed via the vanishing of a smooth function, e.g., a polynomial, the vector fields take their values in a non-Euclidean space. One notable example is the direction vector field which assigns a local direction to each pixel in the image. These directions are unit length vectors that span the unit n-dimensional sphere S^n. Other classes of non-Euclidean vector fields are perceptually treated color images [20] and the regularization of frames [23]. We study in this paper the n-dimensional direction vector fields and spherically constrained color models via the Beltrami framework [19].

The basic objects in the Beltrami framework are embedding maps of Riemannian manifolds. These maps embed the image manifold (a surface for a 2D image) in a fiber bundle whose base is the spatial manifold, e.g., $\mathbb{R}^2$, and the fiber is the feature manifold, e.g., $\mathbb{R}^1$, for the intensity feature alone. If we denote by F the feature manifold and assume that the image is given on a flat surface, then the spatial-feature manifold M is given locally as $M = \mathbb{R}^2 \otimes F$. In all the examples below, the

*Received by the editors October 2, 2002; accepted for publication (in revised form) November 26, 2003; published electronically June 22, 2004. We acknowledge grants from the Israel Academy of Science, Israel Ministry of Science, Research fund of the University of Tel-Aviv, and the Adamas super-center for brain research.

http://www.siam.org/journals/siap/64-5/41551.html

†Department of Applied Mathematics, University of Tel Aviv, Ramat-Aviv, Tel-Aviv 69978, Israel (sochen@math.tau.ac.il, chensagi@post.tau.ac.il).

‡Department of Computer Science, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel (ron@cs.technion.ac.il).

fiber bundle is trivial, yet our local treatment extends to nontrivial bundles as well. Global issues of nontrivial fiber bundles are beyond the scope of this paper.

Another important ingredient of the Beltrami framework is a geometrical functional, known as the Polyakov action (or harmonic energy [1]), which is defined over the space of embedding maps. The minimization of the Polyakov action is done by an Euler–Lagrange operator that drives, through a gradient descent equation, the initial noisy feature vector fields towards a minimum of the Polyakov action. The special form of this functional favors piecewise smooth images. Jumps in the feature space (feature edges) are consequently preserved [4, 5].

Almost all works that try to minimize a functional with respect to a constraint quantity embed the constrained feature in a higher-dimensional Euclidean space and perform the minimization for the coordinates of this unconstrained space. The common wisdom is to combine a minimization of an unconstrained function and a projection on the constraint variety/manifold. The treatment of direction diffusion was recently addressed along these lines in the low-level vision community. These studies follow the well established literature in the liquid crystal community [3]. The harmonic energy functional and its minimization are subjects to intensive mathematical study as well [6, 7]. Two approaches for this problem are known: in a paper that first directly addresses this issue, Perona [13] uses a single parameter θ as an internal coordinate in S^1. The second approach [21, 22, 2] embeds the unit circle S^1 in $\mathbb{R}^2$ (the sphere S^2 in $\mathbb{R}^3$) and works with the external coordinates; see also [24] for a related effort. The first approach is problematic because of the periodicity of S^1. Averaging small angles around zero such as $\theta = \epsilon$ and $\theta = 2\pi - \epsilon$ leads to the erroneous conclusion that the average angle is $\theta = \pi$. Perona solved this problem by exponentiating the angle so that $V = e^{i\theta}$. This is actually the embedding of S^1 in $\mathbb{C}$ which is isometric to $\mathbb{R}^2$. This method is specific to a 2D embedding space where complex numbers can be used. The problem in using only one internal coordinate manifests itself in the numerical implementation of the PDE through the breaking of rotation invariance. In the second approach we have to make sure that we always stay on S^1 along the flow. This problem is known as the projection problem. It is solved in the continuum by adding a projection term. Tang, Sapiro, and Caselles [21, 22] propose the formalism of p-harmonic maps applied to the case of direction and color diffusion, and present experiments in the case $p = 2$, which corresponds to the Dirichlet integral. Moreover, they also present experiments for the case $p = 1$ as the immediate extension of the Rudin–Osher–Fatemi total variation (TV) denoising algorithm [14] to the case of general maps with values on manifolds. Nevertheless, they did not study in detail the algorithm for the $p = 1$ case. The algorithmic study for the case $p = 1$ was done by Chan and Shen [2], who also use external coordinates with a projection term and a TV measure in order to better preserve discontinuities in the vector field. This works well for the case where the codimension is one, like a circle. Yet it is difficult to generalize this approach to higher codimensions like the sphere. Moreover, the flow of the external coordinates is difficult to control numerically since numerical errors should be projected onto S^1 and since no well-defined projection exists. Recently an implicit way to define manifolds has been used in this context [1]. We concentrate in this paper on the explicit methods. A comparison between the implicit harmonic energy method and the implicit Beltrami framework can be found in [16].

We propose to work directly on the constrained manifold and to avoid the projection problem altogether. Our solution produces an adaptive smoothing process, which preserves direction discontinuities. The proposed solution works for all dimensions and codimensions, and overcomes possible parameterization singularities by introducing

several internal coordinates on different patches (charts) such that the union of the patches is the feature manifold, i.e., S^n. Adaptive smoothness is achieved by the description of the vector field as a 2D section of the $(n+2)$-dimensional spatial-feature fiber bundle manifold with S^n fibers.

The problem is formulated, in the Beltrami framework [19, 9], in terms of the embedding map

$$Y : (\Sigma, g) \to (M, h),$$

where Σ is the 2D image manifold and M, in this case, is $\mathbb{R}^n \otimes S^1$ with $n = 2$ ($n = 4$) for gray-level (color) images. The key point is the choice of *local coordinate systems* for *both* manifolds,[1] the image manifold Σ (with metric g) and the embedding manifold M (with metric h). At the same time we should verify that the geometric filter (i.e., the denoising PDE) does not depend on the specific choice of coordinates we make.

Once a local coordinate system is chosen for the embedding space and the optimization is done directly in this local coordinate system, we can never leave M and avoid the problem of projection. The difficulty represented in the problem of projection is transformed into the problem of the choice of a local coordinate system, as we describe below. Other examples of enhancement by the Beltrami framework of nonflat feature spaces, like the color perceptual space and the derivatives vector field, can be found in [20, 17].

An important issue in this approach is the numerical consideration in the choice of local coordinates. While all coordinates are equally good from analytic and geometric points of view, they are different from a numeric standpoint. A comparative study on the numerical and algorithmic accuracies of different schemes is presented here and shows that, for a range of parameters, one can get a better numerical accuracy while maintaining the edge preserving quality of the anisotropic diffusions.

This paper is organized as follows. We review the Beltrami framework and point to the relation with harmonic maps in section 2. We analyze the case of the general n-dimensional direction diffusion with hemispheric coordinate system in section 3. A stereographic coordinate system is introduced in section 4, and the appropriate equations are derived. Section 5 deals with the numerical implementation of the ideas presented in the previous sections for color image processing. Section 6 presents results on various vector fields and color images. We compare in section 7 different direction diffusion schemes from numerical and algorithmic points of view. We summarize and conclude in section 8.

2. The Beltrami framework. Let us briefly review the Beltrami geometric framework for nonlinear diffusion in computer vision [19].

2.1. Representation and Riemannian structure. An image, and many other quantities of interest in computer vision, are naturally represented via the concept of a fiber bundle. The image domain is the base manifold. In the present study it is taken as a subset of $\mathbb{R}^2$ with the canonical Cartesian coordinate system (Y^1, Y^2). It is denoted by Ω. At each point in the base manifold we attach a feature space—the fiber. The fibers at different points of the base manifold are isomorphic. The fiber space is denoted by F. The feature space, or fiber, may be a linear vector space or more interestingly a Riemannian manifold. An image (or other quantity of interest)

[1] Note the difference between this approach and the one presented in [21, 22, 2], where the image metric is flat.

is a choice of a particular point in the fiber for every point in the base manifold. Such a particular choice is called *a section* of the (trivial) fiber bundle $\Omega \otimes F$.

In general an n-dimensional (Riemannian) manifold is defined by a collection of maps from charts of the manifold to $\mathbb{R}^n$. Each chart covers part of the manifold. Their union covers the whole manifold, and the transformation of the coordinates on the intersection between any two charts is smooth. The Riemannian structure transforms in a proper way (as a tensor) under any change of the coordinate system. We denote the coordinates on the 2D section by (x^1, x^2), the coordinates on a chart of the embedding space (the fiber bundle) by $(Y^1, \ldots, Y^n)$. The embedding space is a hybrid spatial-feature space. The first two coordinates (Y^1, Y^2) are the spatial coordinates on Ω (the base manifold), and the rest $(Y^3, \ldots, Y^n)$ are the feature coordinates (the fiber's coordinates). The simplest example is a gray-value image which is represented as a 2D surface embedded in $\mathbb{R}^3$. We denote the map by $Y : \Sigma \to \mathbb{R}^3$, where Σ is a 2D section. The map Y is given in our example by $(Y^1 = x^1,\ Y^2 = x^2,\ Y^3 = I(x^1, x^2))$. We choose on this surface a Riemannian structure, namely a metric. Note that this differs from the harmonic energy functional, where the metric is taken from the base manifold and not from the section. The metric is a positive definite and a symmetric 2-tensor that may be defined through the local distance measurements

$$ds^2 = g_{11}(dx^1)^2 + 2g_{12}dx^1dx^2 + g_{22}(dx^2)^2. \tag{2.1}$$

We use the Einstein summation convention in which the above equation reads as $ds^2 = g_{\mu\nu}dx^\mu dx^\nu$, where repeated indices are summed over. We denote the inverse of the metric by $g^{\mu\nu}$.

2.2. Image metric selection: The induced metric. A reasonable assumption is that distances measured in the embedding spatial-feature fiber bundle, such as distances between pixels and differences between gray-levels, correspond directly to distances measured on the image manifold, i.e., the section. This is the assumption of isometric embedding under which we can calculate the image metric in terms of the embedding maps Y^i and the embedding space metric h_{ij}. This follows directly from the fact that the length of infinitesimal distances on the manifold can be calculated on the manifold and on the embedding space with the same result. Formally, $ds^2 = g_{\mu\nu}dx^\mu dx^\nu = h_{ij}dY^i dY^j$. By the chain rule, $dY^i = \partial_\mu Y^i dx^\mu$, we get $ds^2 = g_{\mu\nu}dx^\mu dx^\nu = h_{ij}\partial_\mu Y^i \partial_\nu Y^i dx^\mu dx^\nu$, from which we have

$$g_{\mu\nu} = h_{ij}\partial_\mu Y^i \partial_\nu Y^j. \tag{2.2}$$

As an example we take the gray-level image as a 2D image manifold embedded in the 3D Euclidean space $\mathbb{R}^3$. The embedding maps are

$$(Y^1(x^1, x^2) = x^1,\ Y^2(x^1, x^2) = x^2,\ Y^3(x^1, x^2) = \beta I(x^1, x^2)). \tag{2.3}$$

The scaling factor β defines the ratio between distances in gray-values and distances in the spatial space. It is a free parameter of the framework that interpolates between the Euclidean L_2 and L_1 types of flows, as we will see below. We choose to parameterize the image manifold by the canonical coordinate system $x^1 = x$ and $x^2 = y$. The embedding, by abuse of notation, is $(x, y, \beta I(x, y))$. The induced metric element g_{11} is calculated as follows:

$$g_{11} = h_{ij}\partial_{x^1} Y^i \partial_{x^1} Y^j = \delta_{ij}\partial_x Y^i \partial_x Y^j = \partial_x x \partial_x x + \partial_x y \partial_x y + \partial_x \beta I \partial_x \beta I = 1 + \beta^2 I_x^2. \tag{2.4}$$

Other elements are calculated in the same manner. The result is

$$G = (g_{\mu\nu}) = \begin{pmatrix} 1+\beta^2 I_x^2 & \beta^2 I_x I_y \\ \beta^2 I_x I_y & 1+\beta^2 I_y^2 \end{pmatrix}. \tag{2.5}$$

2.3. Polyakov action: A measure on the space of embedding maps. Denote by (Σ, g) the image manifold and its metric, and by (M, h) the space-feature manifold and its metric. Then the functional $S[\cdot,\cdot,\cdot]$ attaches a real number to a map $Y : \Sigma \to M$,

$$S[Y^i, g_{\mu\nu}, h_{ij}] = \int dV ||\nabla \vec{Y}||^2_{h,g}, \tag{2.6}$$

where dV is a volume element and the integration is over the Riemannian Frobenius norm[2] of the tangent map dY. In a local coordinate system the volume element is expressed by $dV = dx^1 dx^2 \sqrt{g}$ and $||\nabla \vec{Y}||^2_{h,g} = \langle \nabla Y^i, \nabla Y^j \rangle_g h_{ij} = g^{\mu\nu} \partial_\mu Y^i \partial_\nu Y^j h_{ij}$. The Polyakov action is expressed in this local system of coordinates as

$$S[Y^i, g_{\mu\nu}, h_{ij}] = \int dx^1 dx^2 \sqrt{g} g^{\mu\nu} \partial_\mu Y^i \partial_\nu Y^j h_{ij}. \tag{2.7}$$

This functional, for $m = 2$ (a 2D image manifold) and $h_{ij} = \delta_{ij}$, was proposed by Polyakov [12] in the context of high energy physics and the theory known as *string theory*. It is important to note that the image metric and the feature coordinates—i.e., intensity, color, direction, etc.—are independent variables. This functional is the natural generalization of the L_2 norm from Euclidean domains to Riemannian manifolds. The minimization of the functional with respect to the image metric can be solved analytically in the 2D case (see, for example, [18]). The minimizer is the induced metric. If we choose, a priori, the image metric induced from the metric of the embedding spatial-feature space M, then the Polyakov action is reduced to the area (volume) of the image manifold:

$$S[Y^i, h_{ij}] = 2\int dV = 2\int dx^1 dx^2 \sqrt{g} = 2\int dx^1 dx^2 \sqrt{\det(\partial_\mu Y^i \partial_\nu Y^j h_{ij})}. \tag{2.8}$$

This follows from the form of the induced metric,

$$\langle \nabla Y^i, \nabla Y^j \rangle_g h_{ij} = g^{\mu\nu} \partial_\mu Y^i \partial_\nu Y^j h_{ij} = g^{\mu\nu} g_{\mu\nu}$$

and the identity

$$g^{\mu\nu} g_{\mu\nu} = \mathrm{Tr}(G^{-1} G^T) = \mathrm{Tr}(G^{-1} G) = \mathrm{Tr}(\mathrm{Id}) = 2, \tag{2.9}$$

where $\mathrm{Tr}(X)$ denotes the trace of the matrix X.

Using standard methods in the calculus of variations (see [18]), the Euler–Lagrange equations with respect to the embedding are

$$-\frac{1}{2\sqrt{g}} h^{il} \frac{\delta S}{\delta Y^l} = \frac{1}{\sqrt{g}} \partial_\mu (\sqrt{g} g^{\mu\nu} \partial_\nu Y^i) + \Gamma^i_{jk} \langle \nabla Y^j, \nabla Y^k \rangle_g. \tag{2.10}$$

Since $(g_{\mu\nu})$ is positive definite, $g \equiv \det(g_{\mu\nu}) > 0$ for all x^μ. This factor is the simplest one that does not change the minimization solution while giving a reparameterization

[2]By Riemannian Frobenius norm we mean that the square of the elements is with respect to the Riemannian structures of the corresponding Riemannian manifolds.

invariant expression. The operator that is acting on Y^i in the first term is the natural generalization of the Laplacian from flat spaces to manifolds and is called *the second order differential operator of Beltrami* [10], or the *Beltrami operator*, and is denoted by Δ_g. The second term involves the Levi–Civita connection whose coefficients are the Christoffel symbols. The coefficients are given in terms of the metric of the embedding space

$$\Gamma^i_{jk} = \frac{1}{2} h^{il} \left(\partial_j h_{lk} + \partial_k h_{jl} - \partial_l h_{jk} \right). \tag{2.11}$$

This is the term that takes into account the fact that the image surface flows in a non-Euclidean manifold and not in $\mathbb{R}^n$.

A map that satisfies the Euler–Lagrange equations $-\frac{1}{2\sqrt{g}} h^{il} \frac{\delta S}{\delta Y^l} = 0$ is a *harmonic map*. The 1D and 2D examples are a geodesic curve on a manifold and a minimal surface.

The nonlinear diffusion or scale-space equation emerges as the gradient descent minimization flow

$$Y^i_t = \frac{\partial}{\partial t} Y^i = -\frac{1}{2\sqrt{g}} h^{il} \frac{\delta S}{\delta Y^l} = \Delta_g Y^i + \Gamma^i_{jk} \langle \nabla Y^j, \nabla Y^k \rangle_g. \tag{2.12}$$

This flow evolves a given surface towards a minimal surface, and in general it continuously changes a map towards a harmonic map.

Before closing this review of the Beltrami framework, we would like to point out a few similarities and differences between this flow and those suggested in [14, 13, 21, 2]:

1. For flat fibers:
   - We use the induced metric, while in other flows the image metric is flat. The difference comes from the fact that in our framework the image manifold is a *section* of the fiber bundle, while in the harmonic map formulation it is the *base* manifold.
   - In the case of flat and 1D fibers we get the "regularized total variation" functional. In the limit of large β the evolution equation is identical (up to $\sqrt{g}$) to the TV one. In the limit $\beta \to 0$ we get the linear diffusion case. In intermediate values we find a good compromise such that over-smoothing, on the one hand, and stair-casing, on the other hand, can be avoided. The Beltrami framework, in this case, is a one-parameter generalization of the TV scheme.
   - The multichannel functional, in the Beltrami framework, is another generalization of the TV functional. A term that depends on the direction of the gradients is added to the term that depends on their magnitude only. This provides a better adaptation of the process to the image features.
   - The Beltrami flow is degenerate (at $\nabla I \to \infty$). One can prove that discontinuities are preserved for a finite time [5].
2. For nonflat fibers:
   - The coordinates Y^i are the local coordinates of the feature space, while in the above-mentioned flows they are coordinates of a *third* manifold, i.e., $\mathbb{R}^{n+1}$, in which the feature space S^n is embedded. In other words, the fiber in the harmonic map approach is embedded in $\mathbb{R}^{n+1}$. This is not possible in general (see the Nash embedding theorem [11]).
   - The Polyakov functional is different in this case from the TV functional due to the different weighting of the magnitude of the gradients.

- The flow equation (2.12) has a clear geometric meaning. It is a mean curvature flow projected (analytically) on the fiber. This projection is an edge preserving operation [19]. It depends on ∇I in the general multichannel case and not on $|\nabla I|$ as in the harmonic map approach.

3. Hemispheric direction diffusion.

3.1. Fiber geometry. We are interested in the case where the fiber feature space is the hypersurface S^n. We choose to represent the hypersphere S^n as an n-dimensional manifold embedded in $\mathbb{R}^{n+1}$, with Cartesian coordinate system $\{U^i\}_{i=3}^{n+3}$, as the constrained hypersurface

$$\sum_{i=3}^{n+3}(U^i)^2 = 1. \tag{3.1}$$

We work in the chart, where $\{Y^i\}_{i=3}^{n+2}$ are local coordinates. On this chart, $U^i = Y^i$, $i = 3, \ldots, n+2$, and

$$U^{n+3} = \sqrt{1 - \sum_{i=3}^{n+2}(Y^i)^2}.$$

Denote the metric elements for the feature space only by $\tilde{h}_{ij}$. The metric elements and the inverse metric elements are given by

$$\tilde{h}_{ij} = \delta^{ij} + \frac{Y^i Y^j}{1 - \sum_{k=3}^{n+2}(Y^k)^2},$$

$$(\tilde{h}^{-1})_{ij} = \delta^{ij} - Y^i Y^j. \tag{3.2}$$

3.2. The induced metric. The induced metric and its inverse are accordingly

$$\begin{aligned} g_{\mu\nu} &= \delta_{\mu\nu} + \sum_{i,j=3}^{n+2} \tilde{h}_{ij}\partial_\mu Y^i \partial_\nu Y^j, \\ g^{\mu\nu} &= \frac{1}{g}\left(\delta^{\mu\nu} + \epsilon^{\mu\sigma}\epsilon^{\nu\rho}\sum_{i,j=3}^{n+3}\tilde{h}_{ij}\partial_\sigma Y^i \partial_\rho Y^j\right), \\ g &= \det(g_{\mu\nu}), \\ &= 1 + \sum_{i,j=3}^{n+3}\tilde{h}_{ij}(Y_x^i Y_x^j + Y_y^i Y_y^j), \\ &\quad + \frac{1}{2}\epsilon^{\mu\sigma}\epsilon^{\nu\rho}\sum_{i,j,k,l=3}^{n+3}\tilde{h}_{ij}\tilde{h}_{kl}\partial_\mu Y^i \partial_\rho Y^j \partial_\nu Y^k \partial_\sigma Y^l, \end{aligned} \tag{3.3}$$

where $(g^{\mu\nu})$ is the inverse of $(g_{\mu\nu})$, g is the determinant, and $\epsilon^{\mu\nu}$ is the 2D antisymmetric tensor

$$(\epsilon^{\mu\nu}) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

An implicit summation on all repeated Greek indices is assumed.

3.3. The flow equations. The Levi–Civita coefficients are calculated in Appendix B with the simple result

$$\Gamma^i_{jk} = Y^i \tilde{h}_{jk}. \tag{3.4}$$

The minimization of the Polyakov action leads to the following evolution equations:

$$Y^i_t = \Delta_g Y^i + 2Y^i - Y^i \mathrm{Tr}(g^{\mu\nu}), \qquad i = 1, \ldots, n. \tag{3.5}$$

3.4. The 1D hemispheric direction diffusion.

3.4.1. The S^1 Beltrami operator. The S^1 manifold can be described as the solution to $U^2 + V^2 = 1$. We will work with two charts. One is ($Y^1 = x$, $Y^2 = y$, $Y^3 = \beta U$), and the other is ($Y^1 = x$, $Y^2 = y$, $Y^3 = \beta V$). By abuse of notation we denote the map by $(x, y, \beta Y)$. The parameter β is a scaling factor. Each one of the charts will be used in the range $Y^2 \leq 1/2$. The line element on each of the charts of the image manifold is

$$ds^2 = ds^2_{\mathbb{R}^2} + ds^2_{S^1} = dx^2 + dy^2 + \frac{\beta^2}{1 - Y^2} dY^2. \tag{3.6}$$

By using the chain rule we find

$$ds^2 = (1 + A(Y)Y_x^2)dx^2 + 2A(Y)Y_x Y_y dxdy + (1 + A(Y)Y_y^2)dy^2, \tag{3.7}$$

where $A(Y) = \frac{\beta^2}{1-Y^2}$.

The induced metric is therefore

$$(g_{\mu\nu}) = \begin{pmatrix} 1 + A(Y)Y_x^2 & A(Y)Y_x Y_y \\ A(Y)Y_x Y_y & 1 + A(Y)Y_y^2 \end{pmatrix}, \tag{3.8}$$

and the Beltrami operator acting on Y is $\Delta_g Y = \frac{1}{\sqrt{g}} \partial_\mu (\sqrt{g} g^{\mu\nu} \partial_\nu Y)$, where $g = 1 + A(Y)(Y_x^2 + Y_y^2)$ is the determinant of $(g_{\mu\nu})$, and $(g^{\mu\nu})$ is the inverse matrix of $(g_{\mu\nu})$.

3.4.2. The Levi–Civita connection. Since the embedding space is non-Euclidean, we have to calculate the Levi–Civita connection. Remember that the metric of the embedding space is

$$(h_{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & A(Y) \end{pmatrix}. \tag{3.9}$$

The Levi–Civita connection coefficients are given by the fundamental theorem of Riemannian geometry in the following formula: $\Gamma^i_{jk} = \frac{1}{2} h^{il} (\partial_j h_{lk} + \partial_k h_{jl} - \partial_l h_{jk})$, where the derivatives are taken with respect to Y^i for $i = 1, 2, 3$.

The only nonvanishing term is Γ^3_{33}, which reads

$$\Gamma^3_{33} = \frac{1}{2A(Y)} \partial_Y (A(Y)) = \frac{Y}{1 - Y^2} = Y h_{33}. \tag{3.10}$$

The second term in the Euler–Lagrange equations in this case reads $Y h_{33} ||\nabla Y||^2_g$. We can rewrite this expression using the following identities:

$$\begin{aligned} h_{33}||\nabla Y||^2_g &= (h_{11}g^{11} + h_{22}g^{22} + h_{33}\partial_\mu Y \partial_\nu Y g^{\mu\nu}) - (h_{11}g^{11} + h_{22}g^{22}) \\ &= g_{\mu\nu}g^{\mu\nu} - (g^{11} + g^{22}) = 2 - \frac{1}{g}(g_{11} + g_{22}) = 2 - \frac{1}{g}(1 + g), \end{aligned} \tag{3.11}$$

where we used the induced metric identity (2.2), and the identity (2.9), in order to rewrite

$$2 = \mathrm{Tr}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = g_{\mu\nu}g^{\mu\nu} = h_{11}g^{11} + h_{22}g^{22} + h_{33}\partial_\mu Y \partial_\nu Y g^{\mu\nu}. \tag{3.12}$$

3.4.3. The flow and the switches. The Beltrami flow is

$$Y^i_t = \Delta_g Y^i + \Gamma^i_{jk}(Y^1, Y^2, Y^3)\langle \nabla Y^j, \nabla Y^k \rangle_g \tag{3.13}$$

for $i = 3$. Only modifying the fiber values while keeping the case manifold constant is a *projection* in the direction of the fiber. This projection slows the diffusion around edges. The Beltrami flow on the two charts reads finally as

$$U_t = \Delta_g U + U\frac{g-1}{g},$$

$$V_t = \Delta_g V + V\frac{g-1}{g}. \tag{3.14}$$

In the implementation we compute the diffusion for U and V simultaneously and take the values $(U, \mathrm{sign}(V)\sqrt{1-U^2})$ for the range $U^2 \le V^2$, and $(\mathrm{sign}(U)\sqrt{1-V^2}, V)$ for the range $V^2 \le U^2$.

4. Stereographic direction diffusion.

4.1. Fiber geometry. The hemispheric parameterization requires more charts as n increases. As a result we have to work closer and closer to the singularity. As a cure for that we switch to stereographic parameterization, which demands only two charts independent of the dimension of the sphere. Moreover, we always work on the furthest point from the singularity, that is, on the equator.

Every hypersphere S^n can be isometrically embedded in $\mathbb{R}^{n+1}$. The hypersphere is realized as the place of all the points in $\mathbb{R}^{n+1}$ that satisfy the constraint $\sum_{i=1}^{n+1} U^i U^i = 1$. We denote by Y^i for $i = 1, \ldots, n$ the Cartesian coordinate system on the subspace $\mathbb{R}^n$ that passes through the equator of S^n, i.e., $\{\vec{U} \in \mathbb{R}^{n+1} | U^{n+1} = 0\}$. The stereographic transformation gives the values of Y^i as functions of the points on the north (south) hemispheres of the hypersphere. Explicitly it is given (after shifting the indices by two for notation consistent with the next sections) as

$$Y^i = \frac{U^i}{1 - U^{n+3}}, \quad i = 3, \ldots, n+2.$$

Inverting these relations, we find

$$U^i = \frac{2Y^i}{1 + \sum_{i=1}^n Y^i}, \quad i = 3, \ldots, n+2,$$

$$U^{n+3} = \frac{-1 + \sum_{i=3}^{n+2} Y^i}{1 + \sum_{i=3}^{n+2} Y^i}. \tag{4.1}$$

4.2. The induced metric. Now we can compute the induced metric of our feature space

$$h_{ij} = \sum_{k=3}^{n+3} \frac{\partial U^k \partial U^k}{\partial Y^i \partial Y^j} = \frac{4}{(1+A)^2}\delta_{ij}, \quad i, j = 3, \ldots, n+2, \tag{4.2}$$

where $A = \sum_{k=3}^{n+2} (Y^k)^2$.

4.3. The flow equations. The Levi–Civita connection can be obtained using (2.11) and (4.2). The result is

$$\Gamma^i_{jk} = \frac{4}{1+A}\left(Y^i\delta_{jk} - Y^k\delta_{ij} - Y^j\delta_{ki}\right).$$

The resulting diffusion equations are

$$Y^i_t = \Delta_g Y^i + \sum_{jk} \frac{4}{1+A}\left(Y^i\delta_{jk} - Y^k\delta_{ij} - Y^j\delta_{ki}\right)\partial_\mu Y^j \partial_\nu Y^k g^{\mu\nu}, \tag{4.3}$$

where $i = 3, \ldots, n+2$. This can be rearranged to

$$Y^i_t = \Delta_g Y^i - 4g^{\mu\nu}(\partial_\mu \log(1+A))(\partial_\nu Y^i) + (1+A)(2 - g^{11} - g^{22})Y^i. \tag{4.4}$$

4.4. 1D and 2D directions. We denote our coordinate system by the subscripts s (for south) and n (for north). The equations for the 1D case read

$$(Y_s)_t = \Delta_g Y_s - 4g^{\mu\nu}(\partial_\mu \log(1+A))(\partial_\nu Y_s) + (1+A)(2 - g^{11} - g^{22})Y_s, \tag{4.5}$$

where $A = Y_s^2$ and the induced metric is a function of Y_s. A parallel equation is written for Y_n. We solve the north and south equations simultaneously for values *smaller* than 1. At each iteration we update the values which are greater than 1 by the simple relation $Y_s = 1/Y_n$. Note that the problematic zone(s), i.e., ± 1, are as far as possible from the singularities, i.e., the poles.

The 2D case is managed similarly via

$$\begin{aligned}(Y^1_s)_t &= \Delta_g Y^1_s - 4g^{\mu\nu}(\partial_\mu \log(1+A_s))(\partial_\nu Y^1_s) + (1+A_s)(2 - g^{11} - g^{22})Y^1_s,\\ (Y^2_s)_t &= \Delta_g Y^2_s - 4g^{\mu\nu}(\partial_\mu \log(1+A_s))(\partial_\nu Y^2_s) + (1+A_s)(2 - g^{11} - g^{22})Y^2_s,\end{aligned} \tag{4.6}$$

where $A_s = (Y^1_s)^2 + (Y^2_s)^2$ and the induced metric depends on Y^1_s and Y^2_s. As in the 1D case, we solve simultaneously for the south and north patches and work with Y^i's which are smaller than 1. The update for values that are greater than 1 after the diffusion (in each iteration) is done by $Y^i_s = A_s Y^i_n$. Again the decision zone, i.e., the equator, is the most numerically stable region since it is the furthest from the poles, where singularities may appear.

5. Color diffusion. There are many coordinate systems and models of color space which try to be as close as possible to human color perception. One of the popular coordinate systems is the HSV system [15]. In this system, color is characterized by hue, saturation, and value. The saturation and value take their values in $\mathbb{R}^+$, while the hue is an angle that parameterizes S^1.

In order to denoise and enhance color images by a nonlinear diffusion process which is more adapted to human perception, we use here the HSV system. We need a special treatment of the hue coordinate in section 3.

Let us represent the image as a mapping $\mathbf{Y} : \Sigma \to \mathbb{R}^4 \times S^1$, where Σ is the 2D image surface and $\mathbb{R}^4 \times S^1$ is parameterized by the coordinates (x, y, H, S, V). As mentioned above, a diffusion process in this coordinate system is problematic. We define therefore two coordinates,

$$U = \cos H \qquad \text{and} \qquad W = \sin H,$$

and continue in a way similar to section 3. The metric of $\mathbb{R}^4 \times S^1$ on the patch where U parameterizes S^1 and $W(U)$ is nonsingular is

$$h_{ij} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & A(U) & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{5.1}$$

where $A(U) = 1/(1 - U^2)$.

The induced metric is therefore

$$\begin{aligned} ds^2 &= dx^2 + dy^2 + A(U)dU^2 + dS^2 + dV^2 \\ &= dx^2 + dy^2 + A(U)(U_x dx + U_y dy)^2 + (S_x dx + S_y dy)^2 + (V_x dx + V_y dy)^2 \\ &= (1 + A(U)U_x^2 + S_x^2 + V_x^2)dx^2 \\ &\quad + 2(A(U)U_x U_y + S_x S_y + V_x V_y)dxdy + (1 + A(U)U_y^2 + S_y^2 + V_y^2)dy^2. \end{aligned} \tag{5.2}$$

Similar expressions are obtained on the other dual patch.

The only nonvanishing Levi–Civita connection coefficient is $\Gamma^3_{33} = U h_{33}$. The resulting flow is

$$\begin{aligned} U_t &= \Delta_g U + 2U - U(g^{11} + g^{22}), \\ W_t &= \Delta_g W + 2W - W(g^{11} + g^{22}), \\ S_t &= \Delta_g S, \\ V_t &= \Delta_g V. \end{aligned} \tag{5.3}$$

Note that the switch between U and W should be applied not only to the U and W equations but also to the S and V evolution equations where, at each point, one needs to work with the metric that is defined on one of the patches.

6. Experimental results. Our first example deals with the gradient direction flow via the Beltrami framework. Figure 6.1 shows a vector field before and after the application of the flow for a given evolution time. The normalized gradient vector field extracted from the image is presented before and after the flow and shows the way the field flows into a new smooth direction transactions field.

Our second example deals with color diffusion using different color spaces. We use machine color space as our spectral model, where we first restrict the colors to one quarter of the upper hemisphere defined around the black point in the RGB space, as shown in Figure 6.2. In this example we use the hemispheric direction diffusion. The intensity, or more accurately the magnitude, is handled separately. This is a simple example since a single chart can be used as a parameterization, and indeed this simplified version was often used by others as an example.

Next, we explore a popular model that captures some of our color perception. The HSV (hue, saturation, value) model proposed in [15] is often used as a "user-oriented" color model, rather than the RGB "machine-oriented" model.

Figure 6.3 shows the classical representation of the HSV color space, in which the hue is measured as an angle, while the value (sometimes referred to as brightness) and the color saturation are mapped onto finite nonperiodic intervals. This model lands itself into a filter that operates on the spatial x, y coordinates, the value and saturation coordinates, and the hue periodic variable. Our image is now embedded in $\mathbb{R}^4 \times S^1$.

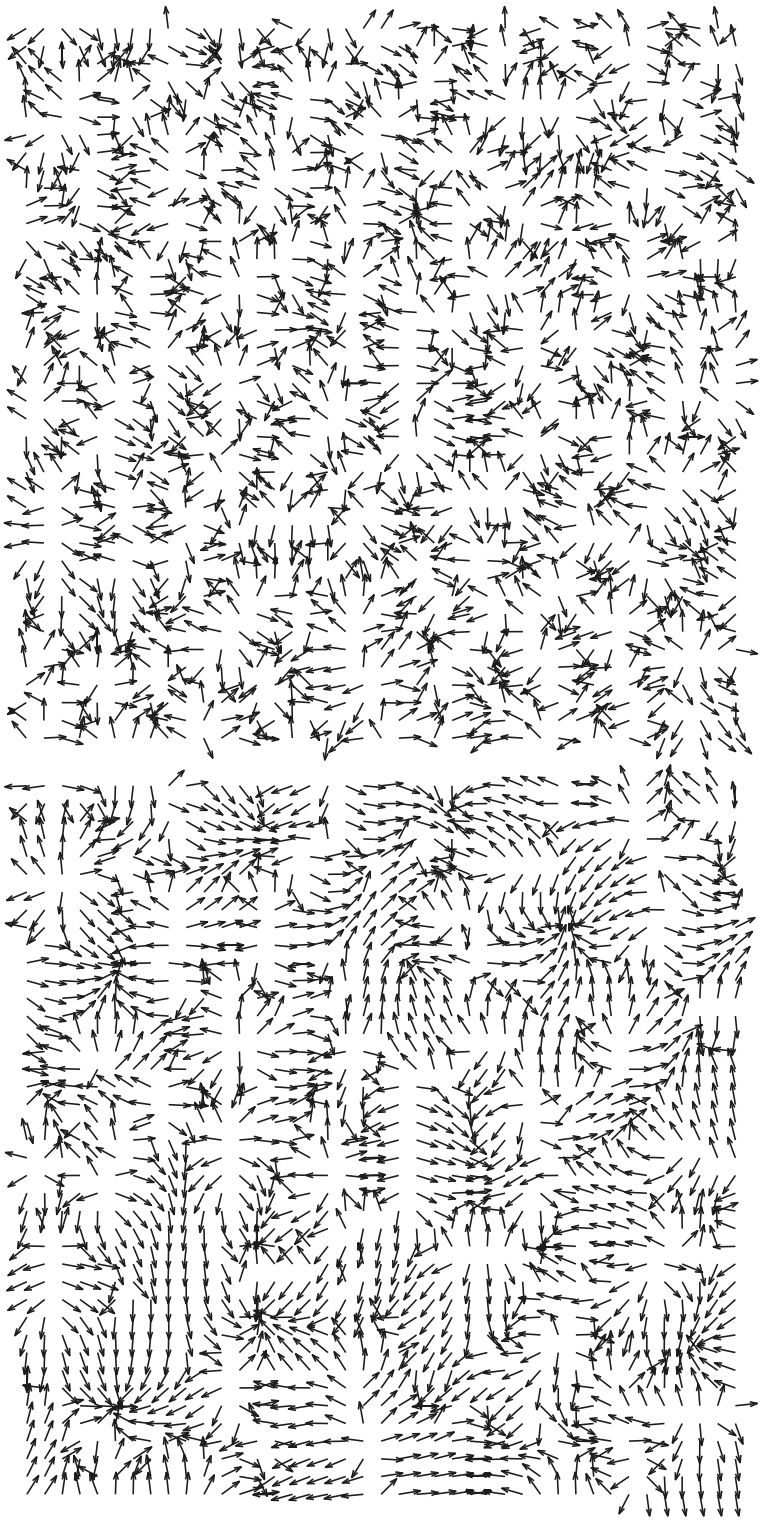


FIG. 6.1. *Two vector fields before (upper) and after (lower) the flow on* S^1.

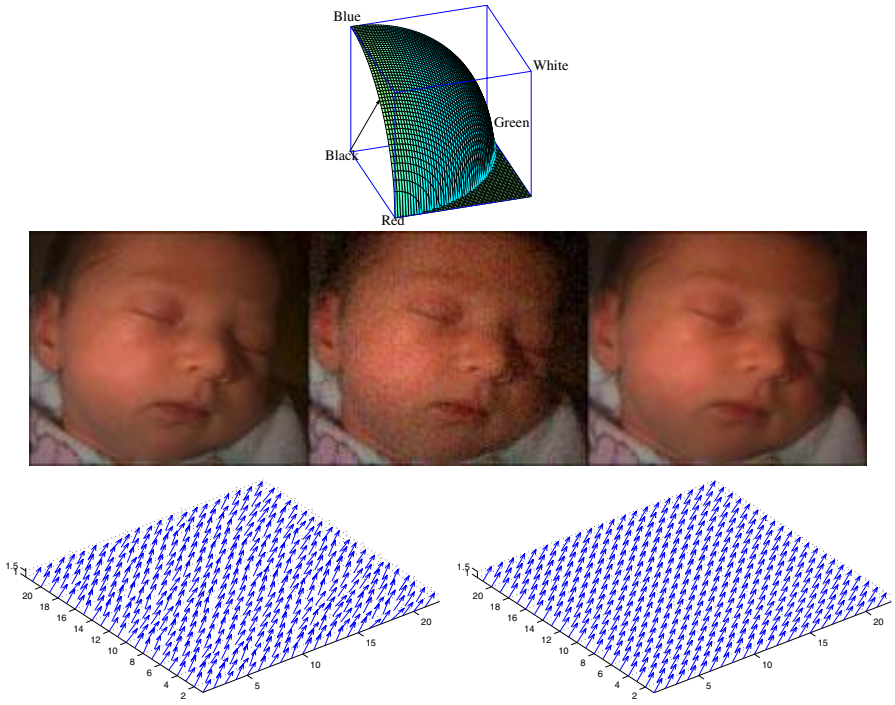


FIG. 6.2. *The colors are restricted to one quarter of the upper hemisphere defined around the black point in the RGB space.*

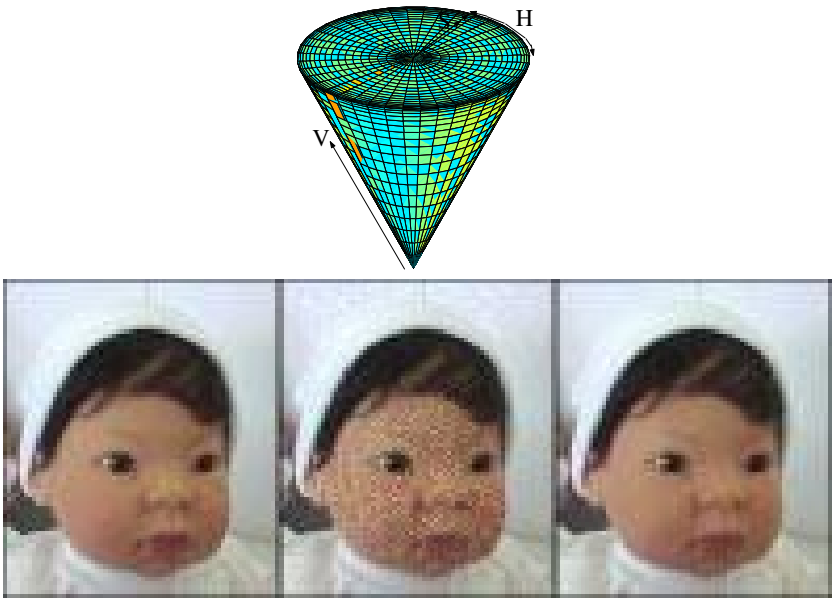


FIG. 6.3. *The HSV color model captures human color perception better than the RGB model which is the common way our machines represent colors. The original image (left), the noisy image (middle), and the filtered image (right) demonstrate the effect of the flow as a denoising filter in the HSV color space when using hemispheric coordinates.*

FIG. 6.4. *An example of stereographic direction diffusion used in the HSV color space. The original image (left), the noisy image (middle), and the filtered image (right) demonstrate the effect of the flow as a denoising filter in the HSV color space when using stereographic coordinates.*

We use the hemispheric direction diffusion for the results shown in Figure 6.3 and the stereographic direction diffusion for the results shown in Figure 6.4. For the complete set of full-size color images see http://www.math.tau.ac.il/∼sochen/Porcupine/porcupine.html.

7. Comparison to other schemes. Several schemes have been suggested to handle direction diffusion. The first to directly address this issue was Perona [13], who uses a single parameter θ as an internal coordinate. However, the periodicity of S^1 leads to erroneous values of θ. Another approach, the linear approach, was offered by Tang, Sapiro, and Caselles [21], in which the unit circle S^1 is embedded in R^2 and external coordinates are used. However, in this flow we have to actively keep our coordinates on S^1, which means that we have to project the results on the unit circle. Chan and Shen [2] studied in detail another scheme in which the evolution equation is derived according to the TV measure.

Kimmel and Sochen [8] have proposed an adaptive hemispheric smoothing scheme, which is edge preserving, based on the Beltrami framework [19]. Throughout this section this scheme is referred to as HP (hemispheric porcupine). The direction vector field is described as a 2D manifold embedded in a higher-dimensional space $M = R^2 \times S^1$. The key point in the HP scheme is the selection of local coordinate systems on the manifold, so that their union is S^1. On the other hand, the local coordinates selection is done so that the numerical error is minimized. The advantage of this scheme is that throughout the flow the coordinates are constrained to S^1. Thus, there is no need for a supplementary projection stage. We address in this work the issue of selecting the right charts to cover S^1, and an alternative stereographic coordinate system is proposed. In this paper we refer to this scheme as SP (stereographic porcupine).

In this study we compare the numerical behavior of the above-mentioned schemes, evaluate their algorithmic performance, and examine their edge preserving quality.

7.1. The evolution equations. In this subsection we mention the evolution equations for each scheme. The interested reader is referred to the original articles.

As a first step, the direction θ is embedded in R^2 via the map $\theta \to \omega = [\cos(\theta), \sin(\theta)]$. The plane is then diffused for some time t, and the result is projected back to the unit circle via the map $\omega_t = [x, y] \to \arctan(\frac{y}{x})$. This is if (x, y) is still a one unit vector. If not, then the phase of the vector is used to determine the appropriate projection; see Figure 7.1.

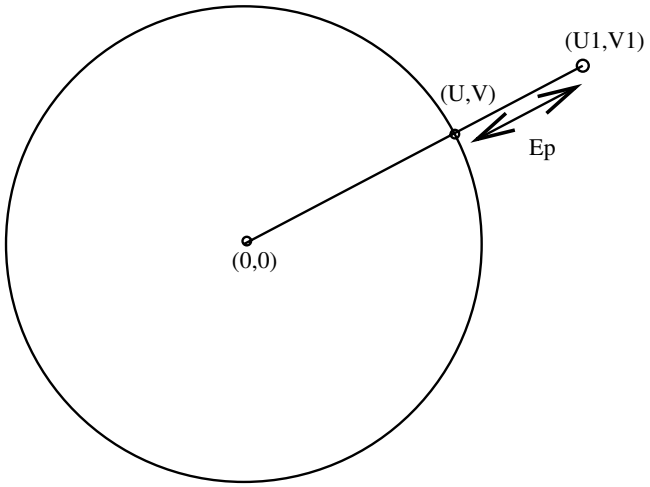


FIG. 7.1. *The projection error in the linear and TV schemes.*

Tang, Sapiro, and Caselles [21] use the following flow for an L_2 energy (which results in a linear scheme),

$$f_t = \Delta f + \| \bigtriangledown f \|^2 f, \tag{7.1}$$

where f stands for the pair (U, V).

Chan and Shen [2] use the following flow for the TV energy:

$$f_t = \text{div}\left(\frac{\bigtriangledown f}{\|f\|}\right) + \| \bigtriangledown f \| f, \tag{7.2}$$

where f stands for the pair (U, V).

Kimmel and Sochen [8] use the following equation for the HP scheme:

$$u_t = \Delta_g U + U \cdot \frac{g-1}{g}, \tag{7.3}$$

$$v_t = \Delta_g V + V \cdot \frac{g-1}{g}, \tag{7.4}$$

where $g = 1 + A(U)((U_x)^2 + (U_y)^2)$ and $A(U) = \frac{1}{1-U^2}$. The SP scheme is given by the following equation:

$$Z_t = \Delta_g Z - 4g^{\mu\nu}(\partial_\mu log(1+A))(\partial_\nu Z) + (1+A)(2 - g^{11} - g^{22})Z, \tag{7.5}$$

where a stereographic coordinate system is used. Here $A = Z^2$, and Z stands for both north and south coordinates.

We remark that in the HP and SP schemes, according to the Beltrami framework, images are considered as surfaces rather than functions. The related diffusion scheme minimizes the area of the image surface. Thus, a basic concept in the Beltrami framework is the manifold's metric. In order to construct a valuable geometric measure for a direction image we have to combine the spatial coordinates with the direction

information. The simplest combination is done by introducing a scaling parameter β, so that

$$ds^2 = dx^2 + dy^2 + \beta^2 \frac{1}{1-U^2} dU^2.$$

The parameter β has dimensions $[\frac{distance}{direction}]$, and it fixes the relative scale between the size of direction information and spatial distances. The parameter β plays an important role in this study. It is a measure of the degree of coupling between the different channels in the diffusion flow. Higher values of β draw the scheme to a behavior similar to that of the TV scheme [2], and smaller values of β cause a behavior similar to that of the linear scheme [21].

Therefore, we expect both HP and SP schemes to have a numerical error and an edge preserving quality which depend on this parameter β.

7.2. Evaluation of the direction diffusion schemes. The evaluation of the different schemes offered for direction diffusion is based on two main attributes of these schemes. The first is their numerical and algorithmic accuracy, which is presented by their degree of error. The second is the edge preserving quality of the scheme. We use direction information which is synthetic. Then, random noise chosen from a uniform distribution on a predefined interval is added to the direction data, and each scheme is used to denoise the image. The numerical error of each scheme is calculated. The algorithmic error is also defined, as the deviation of the resultant direction from the original noise-free direction data. The edge preserving quality of each algorithm is examined on an artificial image which is composed of two different directions, and also on an image which combines a slowly varying direction and a large direction edge.

7.3. Definition of the numerical error. The numerical error is differently defined and calculated for each scheme. In the linear and TV schemes, the numerical error is defined as the amount of the projection needed, so that the direction information is on the unit circle. Thus, if the flow has resulted in some coordinates $(U1, V1)$ which are not necessarily on the unit circle, we take as the projected coordinates the intersection of the unit circle with the line connecting $(U1, V1)$ to the origin of axes; see Figure 7.1. The point (U, V) is given by

$$U = \frac{U1}{\sqrt{U1^2 + V1^2}}, \quad V = \frac{V1}{\sqrt{U1^2 + V1^2}}. \tag{7.6}$$

Thus, the error is clearly

$$error = \sqrt{(U1 - U)^2 + (V1 - V)^2}. \tag{7.7}$$

In the HP and SP schemes, the evaluation of the error is not straightforward, as there is no projection error; the evolving coordinates never leave the unit circle. The numerical error is therefore defined relative to the results of a similar flow in which there is no selection of a local coordinate system; thus, the coordinates (u, v) are not coupled and are not constrained to the unit circle. For the HP we denote this error by $HE_{U1,V1}$ and expect it to obtain a sharp maximum at $(-\pi, \frac{-\pi}{2}, 0, \frac{\pi}{2}, \pi)$ because one of the internal coordinates approaches 1 there and the denominator approaches infinity (see Figure 7.2). It is important to notice that *it is not an error of the hemispheric scheme*. In its minimum value, obtained between the sharp maximum points, it provides a maximum bound on the error in the HP scheme.

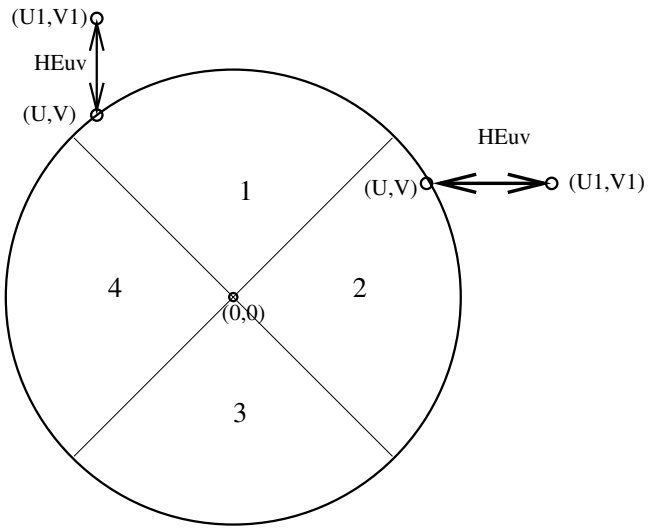


FIG. 7.2. *Artificial error in the HP scheme. In regions* 1 *and* 3 *the* U *coordinate is selected, and therefore the numerical error results from the difference between* $V1$*, which is independently calculated, and* V*, which is derived from the coordinate* U*. In regions* 2 *and* 4 *the* V *coordinate is selected, and therefore the numerical error results from the difference between* $U1$*, which is independently calculated, and* U*, which is derived from the coordinate* V*.*

For SP the definition of an error is even more complicated. Not only is there no projection error, but there are more variables for which an error term may be defined. First, u and v are obtained using the embedding $\theta \to (u, v) = [\cos(\theta), \sin(\theta)]$. Next, the stereographic coordinates Z_n and Z_s are derived, as the intersection of the line between the north (south) pole and the south (north) hemisphere. Thus, we may look at the error in Z_s and Z_n as well as in u and v. Following are the error terms used:

- $\boldsymbol{SE_{zn}}$ **and** $\boldsymbol{SE_{zs}}$**—Error terms for the stereographic coordinates.** We let Z_n and Z_s evolve independently. Then, we compare the stand-alone Z_n to the one calculated using the coupled Z_n and Z_s (where we select the local appropriate chart according to the direction). We do the same for Z_s (see Figure 7.3). We expect the error for Z_n to have a singularity at $\frac{\pi}{2}$ and the error for Z_s to have a singularity at $\frac{-\pi}{2}$. Note that SE_{zn} as defined is expected to be zero in the range $[-\pi, 0]$, and SE_{zs} as defined is expected to be zero in the range $[0, \pi]$. Since this is an error for the values of Z_n and Z_s, we need another error definition which measures the degree of error in the (u, v) coordinates.
- $\boldsymbol{SE_{UV}}$**—Error terms for** $\boldsymbol{U}$ **and** $\boldsymbol{V}$**.** It is important to evaluate the error for the (U, V) variables. We define the error term as the distance between the vector (U, V) when evaluated using the coupled Z_n and Z_s, and the vector (U, V) when using the independently calculated Z_n and Z_s (see Figure 7.4).

It is important to note that SE_z and SE_{UV} *are not errors* of the Beltrami porcupine methods. They give an indication of the actual error by noticing that *the minimum* of SE_z and SE_{UV} is *the upper bound for the Beltrami porcupine algorithm*. This is so since the most unreliable numeric regions are exactly the regions where the minimum in the $SE_{Z,UV}$ is obtained. The actual error in other areas is smaller since we do not trust one of the components that leads to a greater error. Thus, a small value of an error may indicate that using the appropriate local chart

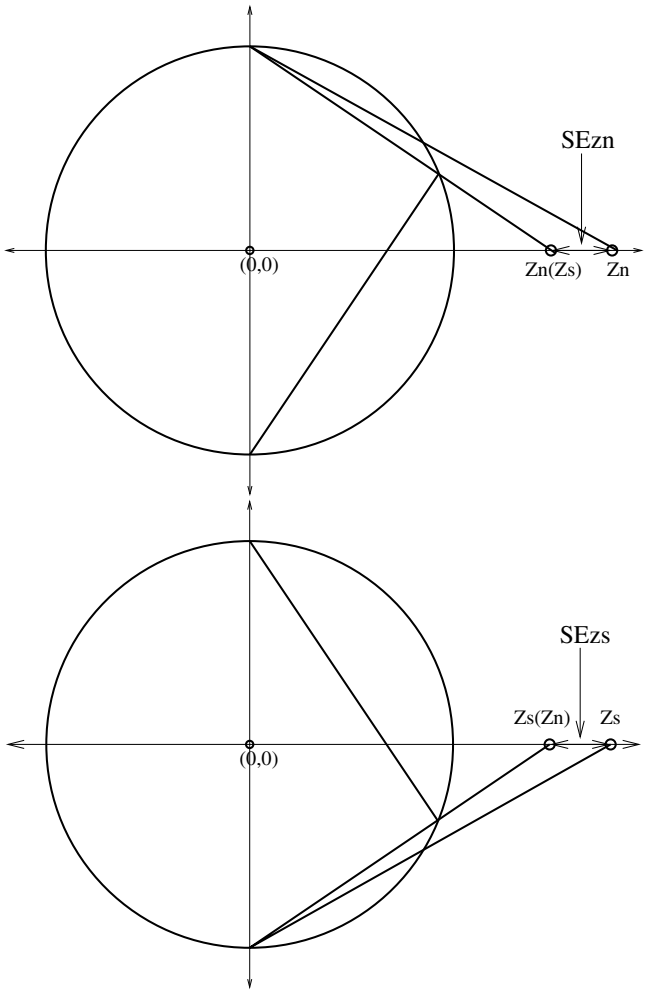


FIG. 7.3. *Artificial error in the Z_n (upper) and Z_s (lower) variables in the SP scheme.*

is not as important as it is when the error is larger. The higher the error, the more important it is to use the right local chart.

7.3.1. Definition of the algorithmic error. The definition of the algorithmic error is the same for all schemes. It is simply the deviation of the direction following diffusion from the noise-free direction, which is originally given. While the numerical error gives an indication of the stability of the method, the algorithmic error deals with performance: how close the resultant direction is to the actual one. The algorithmic error is defined as follows:

$$E = \sqrt{(\cos(\theta) - \cos(\theta_1))^2 + (\sin(\theta) - \sin(\theta_1))^2},$$

where θ is the original noise-free angle and θ_1 is the resultant angle following the diffusion scheme.

7.3.2. Definition of an edge preservation quality. An important quality of any diffusion scheme is its edge preserving ability. The first test image used to examine

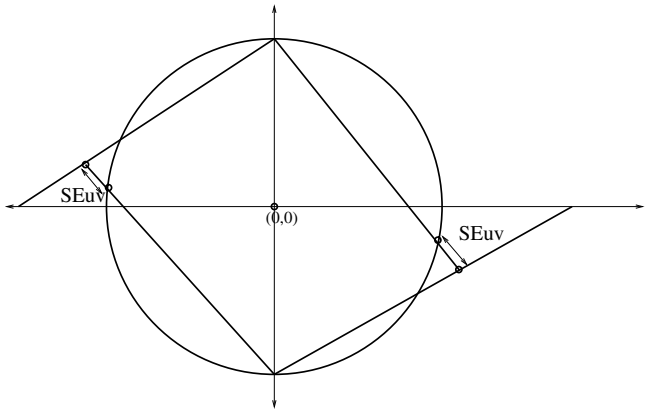


FIG. 7.4. *Artificial error in the* (U, V) *coordinates in the SP scheme.*

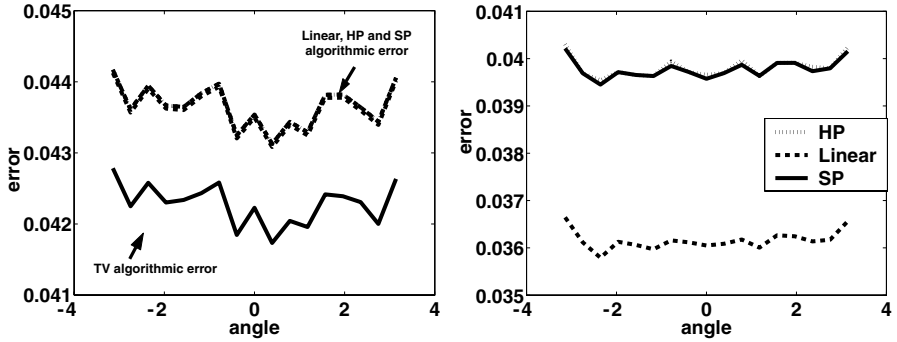


FIG. 7.5. *Algorithmic error for the linear, TV, HP, and SP schemes. Left: the four schemes together, using a small time step,* $dt = 0.00001$. *Right: the HP, SP, and linear schemes, using a larger time step,* $dt = 0.001$.

edge preservation is composed of two different directions. We apply each tested scheme to this image. We expect that the TV-based method will preserve edges better than the linear-based approach. As for the porcupine methods, we expect edge preservation quality to depend on the parameter β. The second test image is composed of two significantly different directions, where each direction is slowly varying. Using this test image, we may compare the edge preserving quality with the handling of the slowly varying data.

7.4. Comparison results and discussion. In this section we present the results of the numerical errors, algorithmic errors, and edge preserving performance.

In the test we go over S^1 from $-\pi$ to π using an equal step size. For each angle, random noise entries, chosen from a uniform distribution, are added to the vector field.

In Figure 7.5(left) we present the algorithmic error for the four schemes using a time step $dt = 0.00001$. All errors lie within the same range. However, the best performance is presented by the TV scheme, while the linear, HP, and SP approaches seem to have the same performance. In Figure 7.5(right) we used a larger time step, $dt = 0.001$, to observe the different behavior of the linear, HP, and SP schemes. The linear scheme has the smallest algorithmic error among the three schemes, and the HP and SP schemes seem to have the same algorithmic performance.

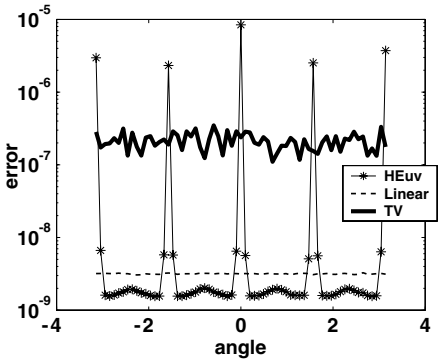


FIG. 7.6. *Numerical error for the TV, linear, and HP schemes. In this test we go over* S^1 *from* $-\pi$ *to* π *using an equal step size of* $\frac{\pi}{32}$.

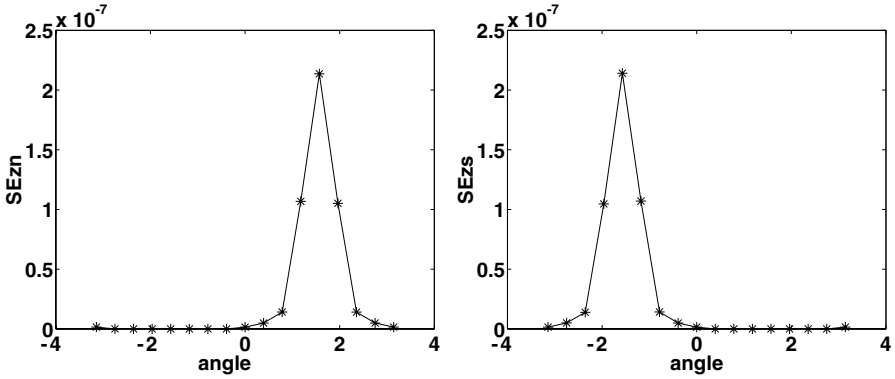


FIG. 7.7. *Left: numerical error for* Z_n. *Here we go over* S^1 *from* $-\pi$ *to* π *using an equal step size of* $\frac{\pi}{8}$. *As expected, it has a singular point at* $\frac{\pi}{2}$. *Right: numerical error for* Z_s. *As expected, it has a singular point at* $\frac{-\pi}{2}$.

Figure 7.6 compares the numerical errors of the HP, TV, and linear schemes. A logarithmic scale is used, as the error of the TV scheme is two orders of magnitude higher than the error of the linear and HP schemes! The HP error term has a periodic behavior, and it is very large at the singular points, $(-\pi, \frac{-\pi}{2}, 0, \frac{\pi}{2}, \pi)$. Away from the singular points, the HP error is slightly smaller than the linear scheme error, and the TV error is significantly higher than the HP error. However, as we approach the singularities, the HP error increases, and there the linear scheme's error is smaller.

In Figure 7.7 we show the numerical errors of Z_n and Z_s in the SP scheme. As expected, the errors have sharp maxima at $\frac{\pi}{2}$ and $\frac{-\pi}{2}$, respectively.

Another definition for the numerical error of the SP scheme was given, SE_{UV}, in which we refer to the (U, V) variables rather than the (Z_n, Z_s) variables. In Figure 7.8 this error is presented: the differences between the values of (U, V) when calculated using a coupled scheme for (Z_n, Z_s) and when calculated using an independent scheme for (Z_n, Z_s) are shown. It is interesting to note that this error has a periodic behavior, with maximum values at $(\frac{-\pi}{2}, \frac{\pi}{2})$, as can be expected.

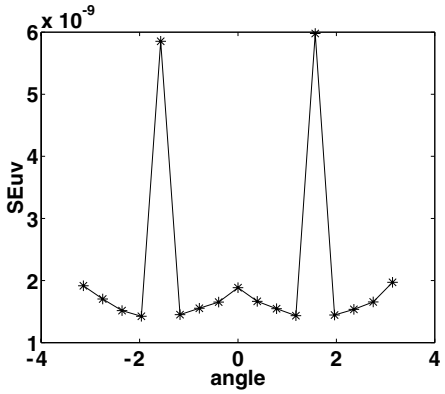


FIG. 7.8. *The differences between the values of* (u, v) *when calculated using the* (Z_n, Z_s) *coupled scheme and when independently calculated using* (Z_n, Z_s). *Here we go over* S^1 *from* $-\pi$ *to* π *using an equal step size of* $\frac{\pi}{8}$.

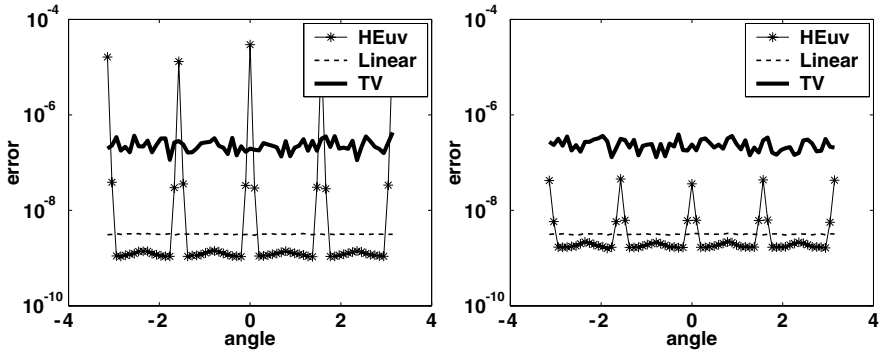


FIG. 7.9. *Numerical error for HP for* $\beta = 10$ *(left) and for* $\beta = 0$ *(right). Here we go over* S^1 *from* $-\pi$ *to* π *using an equal step size of* $\frac{\pi}{32}$.

The next step is to examine the dependence of the numerical error in the HP and SP schemes on the scaling parameter β. Figure 7.9 shows the numerical error for the HP scheme for a larger value of β (left) and for a smaller value of β (right). The scale used for presenting these results is again logarithmic. Away from the singular points, larger values of β produce smaller errors. In the vicinity of the singular points, the error increases when β increases.

The same goes for the SP scheme. In Figures 7.10 and 7.11 we present the results with respect to the three error measures we have defined for the SP scheme. The scale used for presenting the results is logarithmic. In Figure 7.10 the results for a larger value of β are presented, and in Figure 7.11 the results for a smaller value of β are presented.

When $\beta = 100$, the values of SE_{zn} in the range $[0, \pi]$ and away from the singularity at $\frac{\pi}{2}$ lie between the numerical errors of the linear and TV schemes. The error decreases as we move away from $\frac{\pi}{2}$ and is even smaller than the linear scheme error as we move closer to 0 and π. In the range $[-\pi, 0]$, SE_{zn} is equal to zero. SE_{zs} presents a mirror behavior. SE_{UV} is smaller than the numerical errors of the TV and linear schemes. It obtains maximum values at $\pm\frac{\pi}{2}$. When $\beta = 0$, the values of SE_{zn} in the

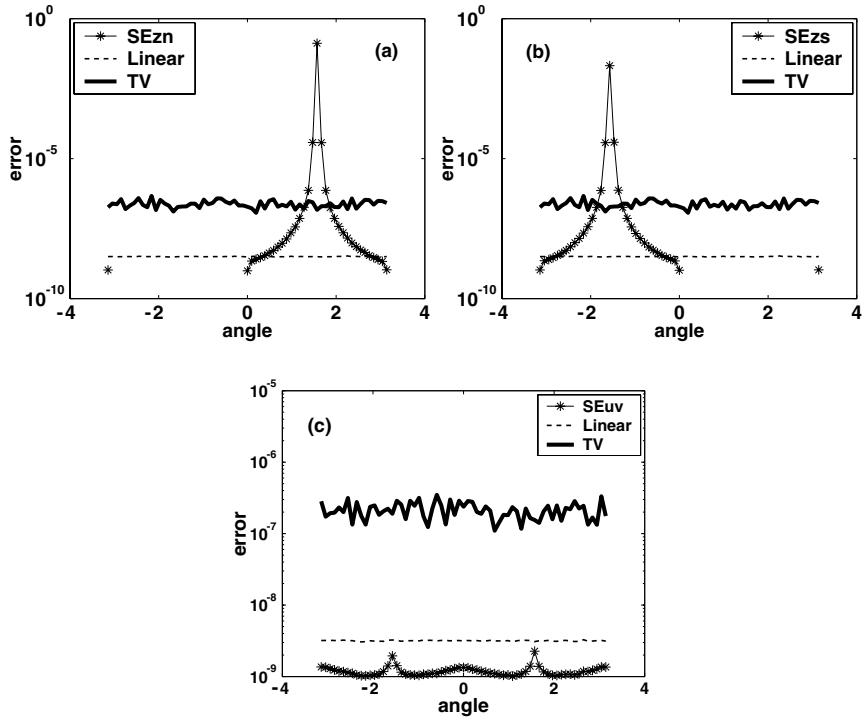


FIG. 7.10. *Numerical errors of the SP scheme for a large value of* $\beta = 100$. (a) *The numerical error of* Z_n. (b) *The numerical error of* Z_s. (c) *The numerical error of the* (U, V) *variables.*

range $[0, \pi]$ and away from the singularity at $\frac{\pi}{2}$ are a little bigger than those obtained for $\beta = 100$. Again, SE_{zs} presents a mirror behavior. In this case, SE_{UV}, away from the singular points $\pm\frac{\pi}{2}$, is higher than the one obtained for $\beta = 100$. Note that the error values in the vicinity of the singularities are much higher for the lower value of β.

Next, we examine the edge preserving quality of each direction diffusion scheme. The following synthetic data was generated so that there is a difference of $\frac{\pi}{2}$ radians between the left and right sides of the noise-free image. Random noise entries, chosen from a uniform distribution in the range $[\frac{-\pi}{9}, \frac{\pi}{9}]$, are added to the noise-free data, and each scheme is applied to the image. The noise-free and noisy initial images are shown in Figure 7.12. The diffusion results are presented for all schemes, while for the HP and SP approaches we show the results for both smaller and higher values of the parameter β. In Figure 7.13 the results for the linear and the TV schemes are presented. It is interesting to note that the linear scheme is less edge preserving than the TV scheme, as can be expected. In Figures 7.14 and 7.15 the results for the HP and SP schemes are also presented. Here, we note the dependence of the results on the value of the parameter β. We can go from linear to TV behavior simply by adjusting the value of β. If we examine the relationship between the numerical errors of the TV and linear schemes (see Figure 7.6), and their edge preserving quality, we note that while the linear scheme offers a low numerical error, it is less edge preserving, and while the TV scheme better preserves edges, it has a significantly higher numerical error. For the HP and SP schemes, both the numerical errors (see Figures 7.9, 7.10, 7.11) and the edge preserving quality depend on the parameter β. We may find a

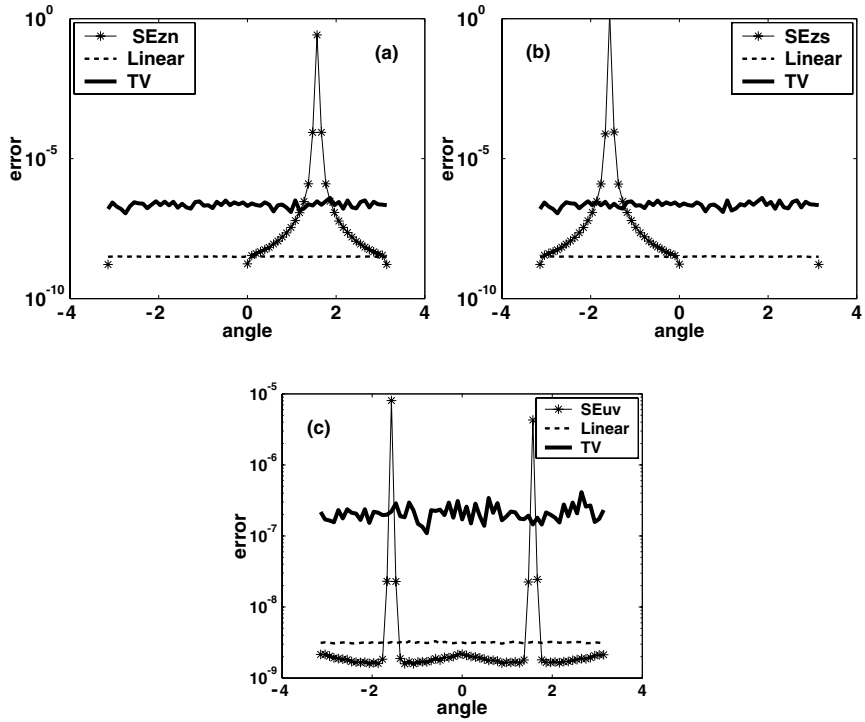


FIG. 7.11. *Numerical errors for the SP scheme for a small value of* $\beta = 0$. (a) *The numerical error of* Z_n. (b) *The numerical error of* Z_s. (c) *The numerical error of the* (U, V) *variables.*

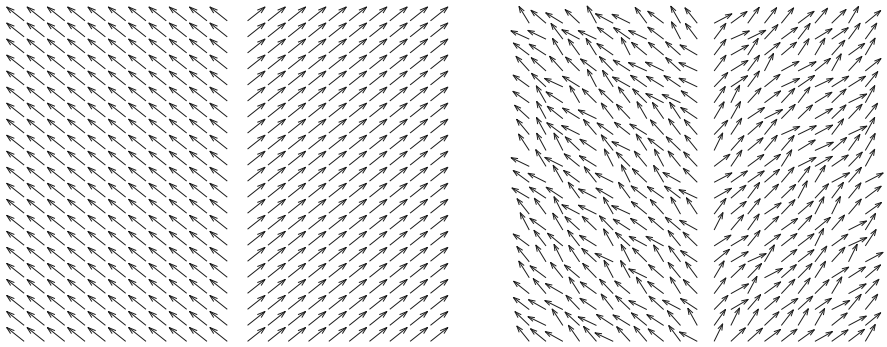


FIG. 7.12. *The original noise-free image (left) and the image after random noise was added (right).*

value of β in the HP and SP schemes so that we obtain a numerical error which is in the order of the linear scheme's error and an edge preserving quality which is comparable to that of the TV scheme.

Another example for exploring the edge preserving quality of each scheme is the direction fan example. The test image (Figure 7.15) is composed of a major gradient in directions in the image's center and a slowly varying angle as we move away from the center. The direction information is presented both by arrows (Figure 7.16 (left)) and by a color image, representing the angles (Figure 7.16 (right)). Random noise

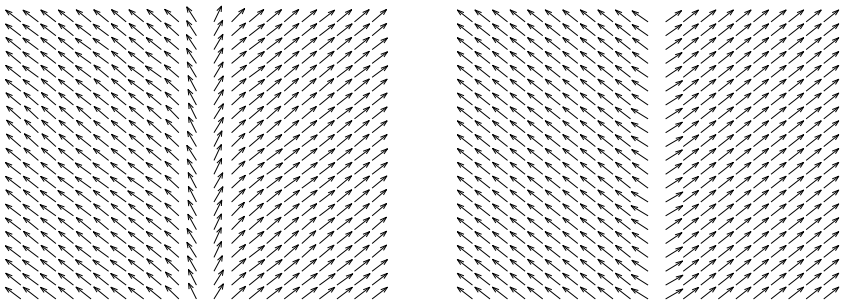


FIG. 7.13. *Left: the result of linear diffusion, with* 10,000 *iterations and a time step equal to* 0.0001. *Right: the result of TV diffusion, with* 100,000 *iterations and a time step equal to* 0.00001.

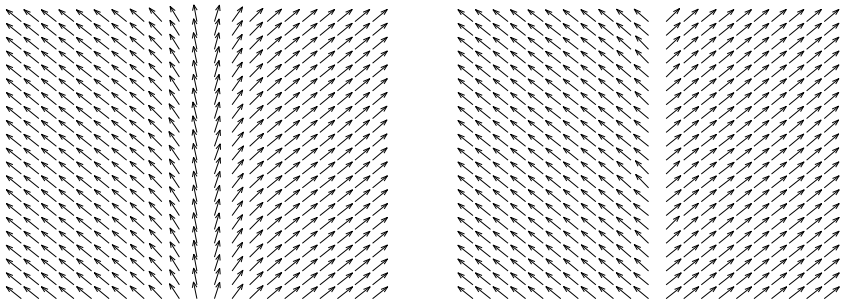


FIG. 7.14. *Left: the result of HP diffusion for* $\beta = 0$. *Right: the result of HP diffusion for* $\beta = 10$. *These results were obtained following* 10,000 *iterations with time step equal to* 0.001.

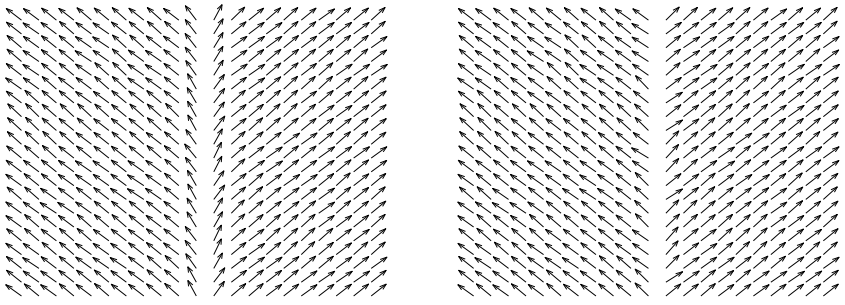


FIG. 7.15. *Left: the result of SP diffusion for* $\beta = 0$. *Right: the result of SP diffusion for* $\beta = 100$. *These results were obtained following* 10,000 *iterations with time step equal to* 0.0001.

entries, chosen from a uniform distribution in the range $[\frac{-\pi}{9}, \frac{\pi}{9}]$, are added to the noise-free data, and a noisy direction image is obtained (Figure 7.17). Next, each scheme is applied to the image with the time step, number of iterations, and value of β (for the HP and SP schemes) that produce the best results. When applying the linear scheme, the edge is blurred while the amount of noise is still significant (Figure 7.18). The TV approach results in a sharper boundary relative to the linear

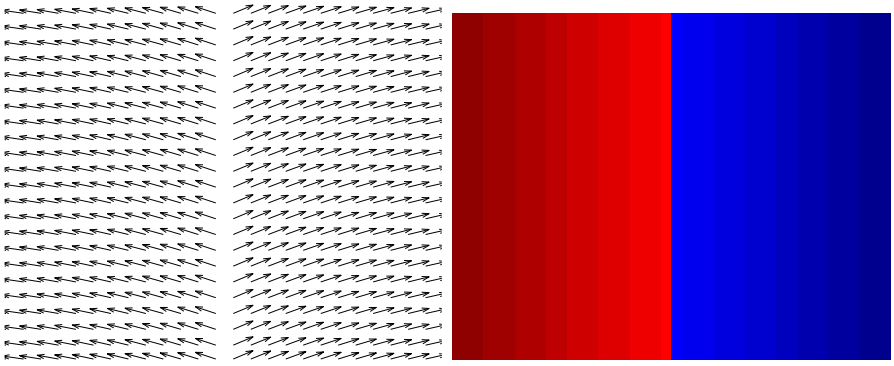


FIG. 7.16. *The noise-free direction fan image, represented by arrows (left) and as a color image (right).*

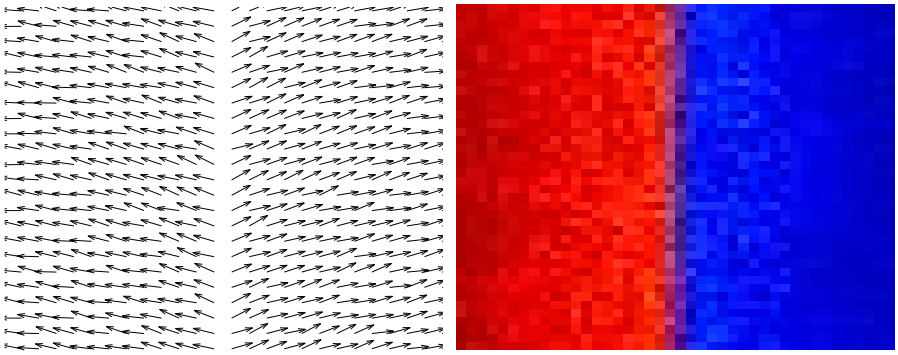


FIG. 7.17. *The noisy direction fan image, represented by arrows (left) and as a color image (right).*

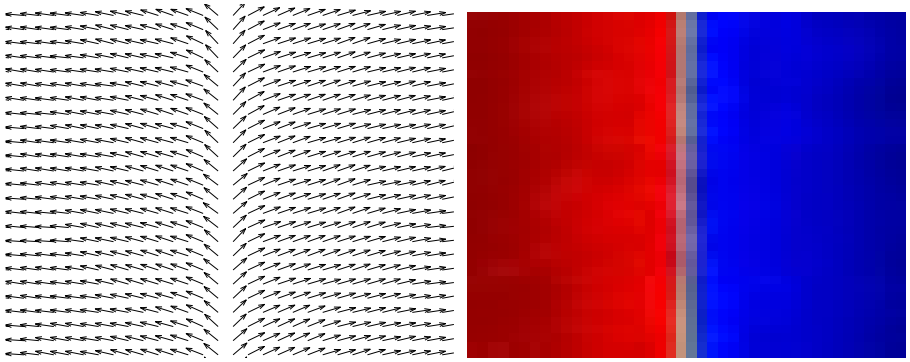


FIG. 7.18. *The result of linear diffusion following* 10,000 *iterations with time step* 0.0001*, represented by arrows (left) and as a color image (right).*

scheme, but if we examine the smoothed direction, we note a stair-casing effect; thus the smaller changes in direction are ignored (Figure 7.19).

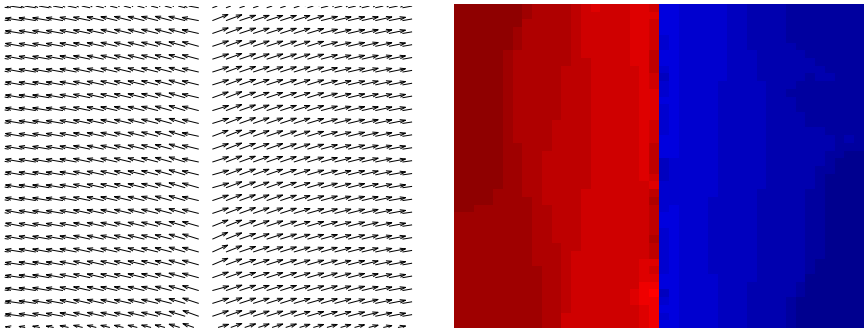


FIG. 7.19. *The result of TV diffusion following* 100,000 *iterations with time step* 0.00001*, represented by arrows (left) and as a color image (right).*

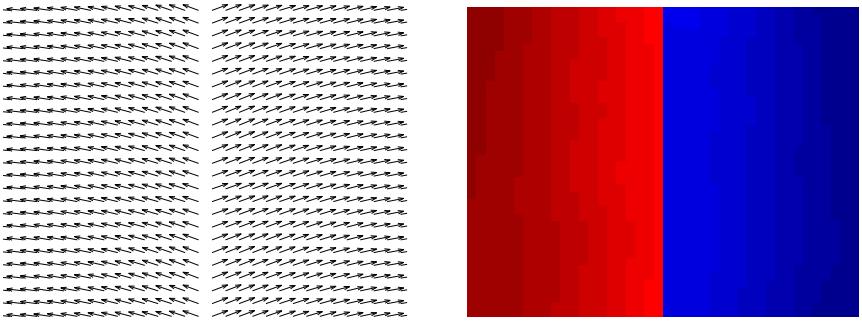


FIG. 7.20. *The result of HP diffusion following* 1,000 *iterations with time step* 0.01*. The value of* β *is* 1.5*. Representation by arrows (left) and as a color image (right).*

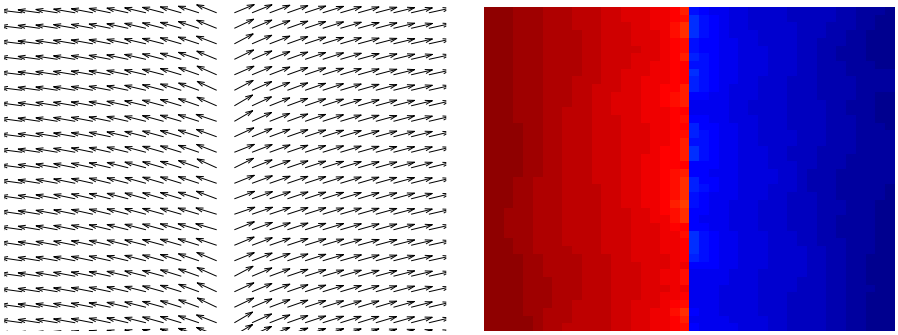


FIG. 7.21. *The result of SP diffusion following* 1,000 *iterations with time step* 0.01*. The value of* β *is* 10*. Representation by arrows (left) and as a color image (right).*

The HP scheme produces good results, as it keeps a sharp boundary and restores the original slowly changing behavior of the original direction data (Figure 7.20). The SP scheme produces similar results to those for the HP scheme, but as can be seen, some noise is still present (Figure 7.21).

8. Concluding remarks. There are some important issues in the process of denoising a constrained feature field. The first is to make the process compatible with the constraint in such a way that the latter is never violated along the flow. The second is the type of regularization which is applied in order to preserve significant discontinuities of the feature field while removing noise. The third is the numeric and algorithmic accuracy of the algorithms.

These issues are treated in this paper via the Beltrami framework. First a Riemannian structure, i.e., a metric, is introduced on the feature manifold, and several local coordinate systems are chosen to *intrinsically* describe the constrained feature manifold. The diffusion process acts on these coordinates, and the compatibility with the constraint is achieved through the intrinsic nature of the coordinate system. The difficulty in working on a non-Euclidean space transforms itself to the need to locally choose the best coordinate system to work with.

The preservation of significant discontinuities is dealt with by using the induced metric and the corresponding Laplace–Beltrami operator acting on feature coordinates only. This operation is in fact a projection of the mean curvature, in the normal(s) direction(s) to the surface, to the feature direction(s). This projection slows the diffusion process along significant (supported) discontinuities while letting the process proceed in the homogeneous regions at a normal speed.

The result of this algorithm is an adaptive smoothing process for a constrained feature space in every dimension and codimension. As an example we have shown how our geometric model coupled with a proper choice of charts handles the direction diffusion problem. This is a new application of the Beltrami framework, proposed in [18]. We tested the new model on vector fields restricted to the unit circle S^1, and hybrid spaces like the HSV color space. The integration of the spatial coordinates with the color coordinates yields a selective smoothing filter for images in which some of the coordinates are restricted to a circle.

Moreover, it is shown that even when algorithms are analytically equivalent, they may differ in their accuracy (numerical and algorithmic). It is shown that the hemispheric and stereographic coordinate systems present an advantage in the sense that a parameter β can be found, i.e., $\beta = 10$, or 100, respectively, such that the edge preserving quality is as good as that for the TV algorithm, while the numerical error is two orders of magnitude smaller!

Appendix A. The Levi–Civita method for S^2. Using (3.9) and the general formula

$$\Gamma^i_{jk} = \frac{1}{2} h^{il} \left(\partial_j h_{lk} + \partial_k h_{jl} - \partial_l h_{jk} \right), \tag{A.1}$$

we get, for example,

$$\begin{aligned}
\Gamma^3_{33} &= \frac{1}{2} h^{3l} \left(2\partial_3 h_{l3} - \partial_l h_{33} \right) = \frac{1}{2} \left(h^{33} \partial_3 h_{33} + 2h^{34} \partial_3 h_{34} - h^{34} \partial_4 h_{33} \right) \\
&= \frac{1}{2} \Bigg[(1-U^2) \frac{\partial}{\partial U} \left(\frac{1-V^2}{1-U^2-V^2} \right) - 2UV \frac{\partial}{\partial U} \left(\frac{UV}{1-U^2-V^2} \right) \\
&\qquad + UV \frac{\partial}{\partial V} \left(\frac{1-V^2}{1-U^2-V^2} \right) \Bigg],
\end{aligned}$$

and a straightforward calculation gives

$$\Gamma^3_{33} = \frac{U(1-V^2)}{1-U^2-V^2} = Uh_{33}. \tag{A.2}$$

Appendix B. The S^n diffusion flow. The hypersphere S^n is presented as an n-dimensional manifold embedded in $\mathbb{R}^{n+1}$ as the constrained hypersurface

$$\sum_{i=1}^{n+1}(U^i)^2 = 1.$$

We work in the chart where $\{U^i\}_{i=1}^n$ are local coordinates. On this chart, $U^{n+1} = \sqrt{1-\sum_{i=1}^n (U^i)^2}$.

THEOREM B.1. *The local S^n metric elements are*

$$\tilde{h}_{ij} = \delta^{ij} + \frac{U^i U^j}{1-\sum_{s=1}^n (U^s)^2}.$$

Proof. The hypersphere is embedded isometrically in $\mathbb{R}^{n+1}$. We use the induced metric technique as follows:

$$ds^2 = \sum_{i=1}^n (dU^i)^2 + (dU^{n+1})^2. \tag{B.1}$$

The U^{n+1} coordinate is a function of all the others, and as such we can apply the chain rule to get

$$dU^{n+1} = \sum_{i=1}^n \frac{\partial U^{n+1}}{\partial U^i} dU^i = -\sum_{i=1}^n \frac{U^i}{\sqrt{1-\sum_{s=1}^n (U^s)^2}} dU^i.$$

Using this expression in (B.1), we get

$$\begin{aligned} ds^2 &= \sum_{i,j=1}^n \tilde{h}_{ij} dU^i dU^j \\ &= \sum_{i=1}^n (dU^i)^2 + \left(-\sum_{i=1}^n \frac{U^i}{\sqrt{1-\sum_{s=1}^n (U^s)^2}} dU^i\right)\left(-\sum_{j=1}^n \frac{U^j}{\sqrt{1-\sum_{s=1}^n (U^s)^2}} dU^j\right) \\ &= \sum_{i,j=1}^n \delta_{ij} dU^i dU^j + \sum_{i,j=1}^n \frac{U^i U^j}{1-\sum_{s=1}^n (U^s)^2} dU^i dU^j \\ &= \sum_{i,j=1}^n \left(\delta_{ij} + \frac{U^i U^j}{1-\sum_{s=1}^n (U^s)^2}\right) dU^i dU^j, \end{aligned} \tag{B.2}$$

from which the assertion follows. ∎

THEOREM B.2. *The local S^n inverse metric elements are*

$$\tilde{h}^{-1}_{ij} = \delta^{ij} - U^i U^j.$$

Proof. By direct calculation,

$$\begin{aligned}\sum_{j=1}^{n}\tilde{h}_{ij}\tilde{h}^{-1}_{jk} &= \sum_{j=1}^{n}\left(\delta^{ij}+\frac{U^iU^j}{1-\sum_{s=1}^{n}(U^s)^2}\right)\left(\delta^{jk}-U^jU^k\right)\\ &= \delta^{ik}-U^iU^k+\frac{U^iU^k}{1-\sum_{s=1}^{n}(U^s)^2}-\sum_{j=1}^{n}\frac{U^i(U^j)^2U^k}{1-\sum_{s=1}^{n}(U^s)^2}\\ &= \delta^{ik}. \qquad \text{(B.3)}\end{aligned}$$

One can check similarly that

$$\sum_{j=1}^{n}\tilde{h}^{-1}_{ij}\tilde{h}_{jk}=\delta^{ik}. \qquad \square$$

THEOREM B.3. *The induced metric, and its inverse, are accordingly*

$$\begin{aligned}g_{\mu\nu} &= \delta_{\mu\nu}+\sum_{i,j=1}^{n}\tilde{h}_{ij}U^i_\mu U^j_\nu,\\ g^{\mu\nu} &= \frac{1}{g}\left(\delta^{\mu\nu}+\epsilon^{\mu\sigma}\epsilon^{\nu\rho}\sum_{i,j=1}^{n}\tilde{h}_{ij}U^i_\sigma U^j_\rho\right),\\ g &= \det(g_{\mu\nu})\\ &= 1+\sum_{i,j=1}^{n}\tilde{h}_{ij}(U^i_xU^j_x+U^i_yU^j_y)+\frac{1}{2}\epsilon^{\mu\sigma}\epsilon^{\nu\rho}\sum_{i,j,k,l=1}^{n}\tilde{h}_{ij}\tilde{h}_{kl}U^i_\mu U^j_\nu U^k_\rho U^l_\sigma, \qquad \text{(B.4)}\end{aligned}$$

where $(g^{\mu\nu})$ is the inverse of $(g_{\mu\nu})$, g is the determinant, and $\epsilon^{\mu\nu}$ is the 2D antisymmetric tensor

$$(\epsilon^{\mu\nu})=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}.$$

An implicit summation on all repeated Greek indices is assumed.

Proof. The calculation of the metric element is done directly by the induced metric identity

$$\begin{aligned}ds^2 &= g_{\mu\nu}dx^\mu dx^\nu = dx^2+dy^2+\sum_{i,j=1}^{n}\tilde{h}_{ij}dU^idU^j\\ &= \delta_{\mu\nu}dx^\mu dx^\nu+\sum_{ij}\tilde{h}_{ij}U^i_\mu U^j_\nu dx^\mu dx^\nu, \qquad \text{(B.5)}\end{aligned}$$

from which we extract the metric coefficients. The metric is a 2×2 matrix whose determinant is $g=g_{11}g_{22}-g_{12}^2=\epsilon^{\mu\nu}g_{1\mu}g_{2\nu}$. Using the explicit form of the metric, we get

$$
\begin{aligned}
g &= \left(1+\sum_{ij}\tilde{h}_{ij}U_x^iU_x^j\right)\left(1+\sum_{kl}\tilde{h}_{kl}U_y^kU_y^l\right)-\left(\sum_{ij}\tilde{h}_{ij}U_x^iU_y^j\right)\left(\sum_{kl}\tilde{h}_{kl}U_x^kU_y^l\right)\\
&= 1+\sum_{ij}\tilde{h}_{ij}(U_x^iU_x^j+U_y^iU_y^j)+\sum_{ijkl}\tilde{h}_{ij}\tilde{h}_{kl}U_x^i(U_x^jU_y^k-U_y^jU_x^k)U_y^l\\
&= 1+\sum_{ij}\tilde{h}_{ij}(U_x^iU_x^j+U_y^iU_y^j)+\epsilon^{\mu\nu}\sum_{ijkl}\tilde{h}_{ij}\tilde{h}_{kl}U_x^iU_\mu^jU_\nu^kU_y^l\\
&= 1+\sum_{ij}\tilde{h}_{ij}(U_x^iU_x^j+U_y^iU_y^j)\\
&\quad+\frac{1}{2}\epsilon^{\mu\nu}\sum_{ijkl}\tilde{h}_{ij}\tilde{h}_{kl}U_x^iU_\mu^jU_\nu^kU_y^l-\frac{1}{2}\epsilon^{\mu\nu}\sum_{ijkl}\tilde{h}_{ij}\tilde{h}_{kl}U_y^iU_\mu^jU_\nu^kU_x^l \qquad \text{(B.6)}\\
&= 1+\sum_{ij}\tilde{h}_{ij}(U_x^iU_x^j+U_y^iU_y^j)+\frac{1}{2}\epsilon^{\mu\nu}\epsilon^{\sigma\rho}\sum_{ijkl}\tilde{h}_{ij}\tilde{h}_{kl}U_\sigma^iU_\mu^jU_\nu^kU_\rho^l.
\end{aligned}
$$

Finally, we prove the formula for the inverse metric

$$
\begin{aligned}
g^{\mu\nu}g_{\nu\lambda} &= \frac{1}{g}\left(\delta^{\mu\nu}+\epsilon^{\mu\sigma}\epsilon^{\nu\rho}\sum_{i,j=1}^{n}\tilde{h}_{ij}U_\sigma^iU_\rho^j\right)\left(\delta_{\nu\lambda}+\sum_{i,j=1}^{n}\tilde{h}_{ij}U_\nu^iU_\lambda^j\right)\\
&= \frac{1}{g}\left(\delta_\lambda^\mu+\epsilon^{\mu\sigma}\epsilon^{\lambda\rho}\sum_{i,j=1}^{n}\tilde{h}_{ij}U_\sigma^iU_\rho^j+\sum_{k,l=1}^{n}\tilde{h}_{kl}U_\mu^kU_\lambda^l\right.\\
&\qquad\left.+\epsilon^{\mu\sigma}\epsilon^{\nu\rho}\sum_{i,j,k,l=1}^{n}\tilde{h}_{ij}\tilde{h}_{kl}U_\sigma^iU_\rho^jU_\nu^kU_\lambda^l\right) \qquad \text{(B.7)}\\
&= \frac{1}{g}\left(\delta_\lambda^\mu+\sum_{i,j=1}^{n}\tilde{h}_{ij}(U_x^iU_x^j+U_y^iU_y^j)\delta_\lambda^\mu+\epsilon^{\mu\sigma}\epsilon^{\nu\rho}\sum_{i,j,k,l=1}^{n}\tilde{h}_{ij}\tilde{h}_{kl}U_\sigma^iU_\rho^jU_\nu^kU_\lambda^l\right),
\end{aligned}
$$

where the last equality comes from a case-by-case analysis. Remember that $\lambda,\nu\in\{1,2\}$, and take, for example, $\mu=\lambda-1=1$. In this case we get

$$
\begin{aligned}
\sum_{\sigma,\rho=1}^{2}\epsilon^{1\sigma}\epsilon^{2\rho}\sum_{i,j=1}^{n}\tilde{h}_{ij}U_\sigma^iU_\rho^j+\sum_{k,l=1}^{n}\tilde{h}_{kl}U_x^kU_y^l &= \epsilon^{12}\epsilon^{21}\sum_{i,j=1}^{n}\tilde{h}_{ij}U_y^iU_x^j+\sum_{i,j=1}^{n}\tilde{h}_{ij}U_x^iU_y^a\\
&= -\sum_{i,j=1}^{n}\tilde{h}_{ij}U_y^iU_x^j+\sum_{i,j=1}^{n}\tilde{h}_{ij}U_x^jU_y^i=0, \qquad \text{(B.8)}
\end{aligned}
$$

where we have used the fact that the metric is a symmetric tensor. Other cases are analyzed in a similar manner. The third term is also analyzed on a case-by-case basis, and the result, as the reader can verify, is

$$
\epsilon^{\mu\sigma}\epsilon^{\nu\rho}\sum_{i,j,k,l=1}^{n}\tilde{h}_{ij}\tilde{h}_{kl}U_\sigma^iU_\rho^jU_\nu^kU_\lambda^l=\frac{1}{2}\delta_\lambda^\mu\epsilon^{\alpha\beta}\epsilon^{\nu\rho}\sum_{i,j,k,l=1}^{n}\tilde{h}_{ij}\tilde{h}_{kl}U_\alpha^iU_\nu^jU_\rho^kU_\beta^l.
$$

The whole expression in the parentheses in B.7 is, therefore, $\delta_\lambda^\mu g$, which completes our proof. ∎

The last piece of information needed for our machinery is the explicit form of the Levi–Civita coefficients.

THEOREM B.4. *The Levi–Civita coefficients are*

$$\Gamma^i_{jk} = U^i \tilde{h}_{jk}. \tag{B.9}$$

Proof. From the formula (2.11) we get

$$\begin{aligned}\Gamma^i_{jk} &= \frac{1}{2}\sum_l h^{-1}_{il}\left(\partial_j h_{lk} + \partial_k h_{jl} - \partial_l h_{jk}\right)\\ &= \frac{1}{2}\sum_l (\delta^{il} - U^i U^l)\left(\partial_j\left(\frac{U^l U^k}{1-\sum_s (U^s)^2}\right) + \partial_k\left(\frac{U^j U^l}{1-\sum_s (U^s)^2}\right) - \partial_l\left(\frac{U^j U^k}{1-\sum_s (U^s)^2}\right)\right).\end{aligned}$$

Let us compute the first term, for example,

$$\partial_j\left(\frac{U^l U^k}{1-\sum_s (U^s)^2}\right) = \frac{\delta^{jl}U^k}{1-\sum_s (U^s)^2} + \frac{\delta^{jk}U^l}{1-\sum_s (U^s)^2} + \frac{2U^j U^l U^k}{(1-\sum_s (U^s)^2)^2}. \tag{B.10}$$

Summing up the three terms, we get

$$\begin{aligned}\Gamma^i_{jk} = \frac{1}{2}\sum_l (\delta^{il} - U^i U^l)\Bigg(&\frac{\delta^{jl}U^k}{1-\sum_s (U^s)^2} + \frac{\delta^{kj}U^l}{1-\sum_s (U^s)^2} + \frac{2U^j U^l U^k}{(1-\sum_s (U^s)^2)^2}\\ &+ \frac{\delta^{kj}U^l}{1-\sum_s (U^s)^2} + \frac{\delta^{lk}U^j}{1-\sum_s (U^s)^2} + \frac{2U^j U^l U^k}{(1-\sum_s (U^s)^2)^2}\\ &- \frac{\delta^{lk}U^j}{1-\sum_s (U^s)^2} - \frac{\delta^{jl}U^k}{1-\sum_s (U^s)^2} - \frac{2U^j U^l U^k}{(1-\sum_s (U^s)^2)^2}\Bigg).\end{aligned}$$

Now simple algebra gives

$$\begin{aligned}\Gamma^i_{jk} &= \frac{1}{1-\sum_s (U^s)^2}\sum_l (\delta^{il} - U^i U^l)\left(\delta^{kj}U^l + \frac{U^j U^l U^k}{1-\sum_s (U^s)^2}\right)\\ &= \frac{1}{1-\sum_s (U^s)^2}\left(U^i - U^i\sum_l (U^l U^l)\right)\left(\delta^{kj} + \frac{U^j U^k}{1-\sum_s (U^s)^2}\right) = U^i \tilde{h}_{jk}. \qquad \square\end{aligned}$$

Acknowledgments. We thank Alfred Bruckstein from the Technion Israel for stimulating discussions on diffusion and averaging, and on color analysis. We also thank Guillermo Sapiro from the University of Minnesota for sharing with us his ideas and results on direction diffusion.

REFERENCES

[1] M. BERTALMÍO, L. T. CHENG, S. OSHER, AND G. SAPIRO, *Variational problems and partial differential equations on implicit surfaces*, J. Comput. Phys., 174 (2001), pp. 759–780.

[2] T. CHAN AND J. SHEN, *Variational restoration of nonflat image features: Models and algorithms*, SIAM J. Appl. Math., 61 (2000), pp. 1338–1361.

[3] R. COHEN, R. M. HARDT, D. KINDERLEHRER, S. Y. LIN, AND M. LUSKIN, *Minimum energy configurations for liquid crystals: Computational results*, in Theory and Applications of Liquid Crystals, J. L. Ericksen and D. Kinderlehrer, eds., IMA Vol. Math. Appl. 5, Springer, New York, 1987, pp. 99–121.

[4] L. Dascal and N. Sochen, *A maximum principle for the Beltrami color flow*, manuscript.
[5] L. Dascal, S. Kamin, and N. Sochen, *Existence and Uniqueness of the Weak Solutions to the Beltrami Flow*, in preparation.
[6] J. Eells and L. Lamarie, *A report on harmonic maps*, Bull. London Math. Soc., 10 (1978), pp. 1–68.
[7] J. Eells and L. Lamarie, *Another report on harmonic maps*, Bull. London Math. Soc., 20 (1988), pp. 385–524.
[8] R. Kimmel and N. Sochen, *Orientation diffusion or How to comb a porcupine*, J. Visual Communication and Image Representation, 13 (2002), pp. 238–248.
[9] R. Kimmel, R. Malladi, and N. Sochen, *Images as embedded maps and minimal surfaces: Movies, color, texture, and volumetric medical images*, Int. J. Computer Vision, 39 (2000), pp. 111–129.
[10] E. Kreyszig, *Differential Geometry*, Dover Publications, New York, 1991.
[11] J. Nash, *The imbedding problem for Riemannian manifolds*, Ann. Math., 63 (1965), pp. 20–63.
[12] A. M. Polyakov, *Quantum geometry of bosonic strings*, Phys. Lett. B, 103 (1981), pp. 207–210.
[13] P. Perona, *Orientation diffusion*, IEEE Trans. Image Process., 7 (1998), pp. 457–467.
[14] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation-based noise removal algorithms*, Phys. D, 60 (1991), pp. 259–268.
[15] A. R. Smith, *Color gamut transform pairs*, in SIGGRAPH: ACM Special Interest Group on Computer Graphics and Interactive Techniques, ACM Press, NY, 1979, pp. 12–19.
[16] N. Sochen, R. Deriche, and L. Lopez-Perez, *The Beltrami Flow over Manifolds*, Technical report TR-4897, INRIA Sophia-Antipolis, Sophia Antipolis, France, 2003.
[17] N. Sochen, R. M. Haralick, and Y. Y. Zeevi, *A Geometric Functional for Derivatives Approximation*, in Proceedings of the 2nd International Conference on Scale-Space Theories in Computer Vision, Lecture Notes in Comput. Sci. 1682, Springer-Verlag, Berlin, 1999, pp. 507–512.
[18] N. Sochen, R. Kimmel, and R. Malladi, *From High Energy Physics to Low Level Vision*, Technical report, LBNL 39243, University of California at Berkeley, Berkeley, CA, 1996.
[19] N. Sochen, R. Kimmel, and R. Malladi, *A general framework for low level vision*, IEEE Trans. Image Process., 7 (1998), pp. 310–318.
[20] N. Sochen and Y. Y. Zeevi, *Representation of colored images by manifolds embedded in higher dimensional non-Euclidean space*, in Proceedings of the International Conference on Image Processing, Chicago, 1998, IEEE, Los Alamitos, CA, pp. 166–170.
[21] B. Tang, G. Sapiro, and V. Caselles, *Direction diffusion*, in Proceedings of the International Conference on Computer Vision, Corfu, Greece, 1999, Vol. 2, pp. 1245–1252.
[22] B. Tang, G. Sapiro, and V. Caselles, *Diffusion of general data on non-flat manifolds via harmonic maps theory: The direction diffusion case*, Int. J. Computer Vision, 36 (2000), pp. 149–161.
[23] D. Tchumperlé and R. Deriche, *Regularization of orthonormal vector sets using coupled PDE's*, in Proceedings of the 2001 IEEE Workshop on Variational and Level Set Methods in Computer Vision (VLSM'01), Vancouver, 2001, IEEE, Los Alamitos, CA, pp. 3–10.
[24] J. Weickert, *Coherence-enhancing diffusion of colour images*, Image and Vision Computing, 17 (1999), pp. 201–212.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1509–1525

BURGERS–POISSON: A NONLINEAR DISPERSIVE MODEL EQUATION*

KLEMENS FELLNER† AND CHRISTIAN SCHMEISER†‡

Abstract. A dispersive model equation is considered, which has been proposed by Whitham [*Linear and Nonlinear Waves*, John Wiley & Sons, New York, 1974] as a shallow water model, and which can also be seen as a caricature of two-species Euler–Poisson problems. A number of formal properties as well as similarities to other dispersive equations are derived. A travelling wave analysis and some numerical tests are carried out. The equation features wave breaking in finite time. A local existence result for smooth solutions and a global existence result for weak entropy solutions are proved. Finally, a small dispersion limit is carried out for situations where the solution of the limiting equation is smooth.

Key words. dispersive models, entropy solutions, small dispersion limit

AMS subject classifications. 35L65, 35Q20

DOI. 10.1137/S0036139902410345

1. Introduction and motivation.

In this paper, a nonlinear dispersive model problem is proposed, the Burgers–Poisson (BP)-system:

$$u_t + uu_x = \varphi_x, \tag{1.1}$$

$$\varphi_{xx} = \varphi + u, \tag{1.2}$$

where u and φ depend on $(t,x) \in (0,\infty)\times\mathbb{R}$, and subscripts denote partial derivatives. The Burgers equation (1.1) is driven by the right-hand side φ_x, which is determined by solving the Poisson equation (1.2).

Using the Green's function $G(x) = -\frac{1}{2}e^{-|x|}$ (of $\partial_x^2 - 1$) to define the convolution operator

$$\varphi[u](x) = (G * u)(x) = \int_{\mathbb{R}} G(x-y)u(y)\,dy, \tag{1.3}$$

the BP-system reduces to the single BP-equation

$$u_t + uu_x = \varphi_x[u], \tag{1.4}$$

with the obvious notation $\varphi_x[u] := (\varphi[u])_x$.

A rescaled version of (1.4) was considered by Whitham in [30, section 13.14] as a shallow water equation modelling unidirectional water waves subject to weaker

*Received by the editors June 26, 2002; accepted for publication (in revised form) December 2, 2003; published electronically June 22, 2004. This work was supported by Austrian Science Foundation grants W008, P14876 and P16174-N05; the *"Wittgenstein Award"* of Peter A. Markowich; and the RTN network "HYKE—Hyperbolic and Kinetic Equations," contract number HPRN-CT-2002-00282 financed by the European Union.

http://www.siam.org/journals/siap/64-5/41034.html

†Institute for Analysis and Scientific Computing, Vienna University of Technology, Wiedner Hauptstraße 8-10/101, A-1040 Wien, Austria (k.fellner@tuwien.ac.at, christian.schmeiser@tuwien.ac.at).

‡Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria.

dispersive effects than the Korteweg–de Vries (KdV)-equation. More precisely, the rescaled kernel $\widetilde{G}(x) = \frac{\pi}{4} e^{-\frac{\pi}{2}|x|}$ was used as an approximation for the kernel

$$G_g(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \Big(\frac{g}{\kappa} \tanh(\kappa h_0) \Big)^{\frac{1}{2}} e^{i\kappa x} \, d\kappa \tag{1.5}$$

when g and h_0 are normalized to 1. Equation (1.5) is the Fourier transform of the dispersion relation of linearized (small amplitude, velocity potential) gravity (g is the gravitational acceleration) water waves in an inviscid, incompressible, and irrotational fluid of constant depth h_0 [30]. In contrast to the KdV-equation, where the dispersion kernel is singular, the above smooth kernels feature weaker dispersive effects and allow, in particular, peaking and breaking of waves.

The present study of (1.1), (1.2) has been motivated by earlier work [10], [11] on two-species-Euler–Poisson (2SEP)-systems modelling the dynamics of two oppositely charged species of particles subject to Coulomb interaction. A simple version in dimensionless form is given by

$$\rho_t + (\rho u)_x = 0, \tag{1.6}$$

$$u_t + u u_x + \frac{\rho_x}{\rho} = \varphi_x, \tag{1.7}$$

$$\varepsilon \varphi_{xx} = \rho - e^{-\varphi}. \tag{1.8}$$

Here, the unknowns ρ, u, and φ depend on position $x \in \mathbb{R}$ and time $t > 0$. The system (1.6), (1.7) comprises the isothermal Euler equations for the first species of particle with density ρ and velocity u. In the Poisson equation (1.8) for the electrostatic potential φ, the density of the second species is modelled by the equilibrium approximation $e^{-\varphi}$, resulting from an equation like (1.7) with the first two terms neglected and the opposite sign on the right-hand side. The dimensionless parameter ε denotes the scaled Debye length.

In view of the formal similarities between the BP- and the 2SEP-system, we shall use the terms position, time, velocity, and potential for the variables x, t, u, and φ, respectively. Note that the velocity (instead of the density) appears on the right-hand side of the Poisson equation. Nevertheless one can consider the BP-system as a caricature of the 2SEP-system with the Burgers equation replacing the Euler equations, and the potential terms in (1.2) coming from a linearization of the two-species Poisson equation (1.8).

The BP-system has a number of interesting formal properties, collected in section 2. In particular, we mention its relation to the Camassa–Holm equation [4], [5] and the Benjamin–Ono equation [2], [26], its Galilean invariance, and its Hamiltonian structure, as well as the existence of an entropy.

In section 3, a general travelling wave analysis of the BP-system is performed, recovering the result of Fornberg and Whitham [17] in the particular case of solitary waves. It turns out that the travelling wave structure of the BP-system and several versions of the 2SEP-system (see [10], [11]) are qualitatively equivalent. The section is completed with some numerical experiments.

In section 4, existence and uniqueness of smooth solutions locally in time for smooth initial data are proved. Recently, for three related problems—the Euler–Poisson system [13], the Camassa–Holm equation [9], [27], and the Rosenau regularization of viscous conservation laws [23], [28]—global existence of smooth solutions has been shown under certain conditions on the initial data. The methods employed there

do not apply to the BP-system. Also our numerical experiments with the BP-system indicate that comparable results are not true.

A global existence and uniqueness result for weak entropy solutions with initial data in $BV(\mathbb{R})$ is also derived. Comparable results have been shown for Rosenau regularized conservation laws [28], [21]. The Rosenau regularization is formally closely related to the right-hand side of (1.4). It is obtained from the BP-system, replacing u by $-u_x$ in the right-hand side of (1.2). As opposed to the dispersive nature of the BP-system, it is dissipative.

The Rosenau regularization has several favorable properties (similar to hyperbolic or viscous conservation laws, but different from the BP-system): It is L^1-contracting, TVD, and has a comparison principle [28]. For the BP-system we do not have uniform (in time) L^∞-bounds, and the total variation may grow with time.

Finally, the rescaling $x \to x/\varepsilon$, $t \to t/\varepsilon$, $0 < \varepsilon \ll 1$, is introduced in (1.1), (1.2), leading to

$$u_t^\varepsilon + u^\varepsilon u_x^\varepsilon = \varphi_x^\varepsilon, \tag{1.9}$$

$$\varepsilon^2 \varphi_{xx}^\varepsilon = \varphi^\varepsilon + u^\varepsilon. \tag{1.10}$$

A Chapman–Enskog expansion of (1.9), (1.10) recovers the Burgers equation with flux $(u^\varepsilon + 1)^2/2$ and the leading order perturbation $\varepsilon^2 u_{xxx}^\varepsilon$. For a rescaled system, the KdV-equation is formally obtained in the limit $\varepsilon \to 0$.

The travelling wave analysis and numerical simulations suggest that the quasi-neutral limit $\varepsilon \to 0$ in general is a weak limit, both for the 2SEP- and the BP-systems. Here, a result is shown for the BP-system which has been proved for a 2SEP-system in [12] and for the Rosenau regularization in [21]: convergence to smooth solutions of the formal limit problem. In general this is only a local-in-time result since the limiting inviscid Burgers equation can develop singularities in finite time. The situation is the same for a 2SEP-system, but not for the Rosenau regularization, where the limiting equation is the viscous Burgers equation with global smooth solutions.

2. Formal properties. First, we rewrite the BP-system as a single differential equation for u. By applying $1 - \partial_x^2$ to (1.1) and using (1.2) on the resulting right-hand side, we calculate

$$u_t - u_{xxt} + u_x + uu_x = 3u_x u_{xx} + uu_{xxx}. \tag{2.1}$$

The terms in (2.1) correspond to those in the Camassa–Holm equation [4]:

$$u_t - u_{xxt} + 2\kappa u_x + 3uu_x = 2u_x u_{xx} + uu_{xxx}, \tag{2.2}$$

where the constant $\kappa \geq 0$ is related to the critical shallow water wave speed. Conversely, the Camassa–Holm equation (2.2) can be written in "BP form":

$$u_t + uu_x = \varphi_x\,, \tag{2.3}$$

$$\varphi_{xx} = \varphi + 2\kappa u + u^2 + \frac{1}{2}u_x^2\,. \tag{2.4}$$

Note that for $\kappa = 1/2$ the BP-system is recovered by neglecting the two quadratic terms in (2.4).

The Camassa–Holm equation was introduced by Fokas and Fuchssteiner [15] as a formally integrable bi-Hamiltonian generalization of the KdV-equation. Camassa and Holm [4] rediscovered it as a shallow water equation by approximating the Hamiltonian

for the vertically averaged incompressible Euler equations. By the bi-Hamiltonian property, they derived an infinite sequence of conservation laws and showed that the associated flows of this hierarchy are isospectral and, thus, completely integrable.

Most commonly (cf. [5], [9], [8], [27]), the Camassa–Holm equation (2.2) is considered with $\kappa = 0$. Then the Camassa-Holm equation possesses peaked soliton solutions (called peakons), attractive travelling waves of the form $u(x,t) = c\exp(-|x-ct|)$, and other breaking phenomena, which is desirable for a shallow water equation and in contrast to the KdV-behavior.

For some initial data (e.g., with sufficiently large negative slope [8], [27]) the solution develops verticality within finite time. On the other hand, global well-posedness was proved (see [9], [27]) for initial data $u_0 \in H^s(\mathbb{R})$ with $s > 3/2$, provided that $\int |u_0|\,dx < \infty$ and $(1-\partial_x^2)u_0$ does not change sign. The Camassa–Holm equation is remarkable since it combines complete integrability with the formation of singularities.

In the existence analysis of section 4, (1.4) will be considered, subject to the initial conditions

$$u(x,0) = u_0(x), \qquad x \in \mathbb{R}. \tag{2.5}$$

Note that (1.4) contains additional information compared to (1.1), (1.2), since for bounded u, (1.3) is the unique bounded solution of the Poisson equation (1.2). The properties of the solution operator of the Poisson equation in a L^2-setting will be used, in particular the smoothing

$$\|\varphi[u]\|_{H^{k+2}(\mathbb{R})} \le \|u\|_{H^k(\mathbb{R})}, \qquad u \in H^k(\mathbb{R}),\ k \ge 0, \tag{2.6}$$

and the symmetry

$$\int_{\mathbb{R}} \varphi[u]v\,dx = \int_{\mathbb{R}} \varphi[v]u\,dx, \qquad u, v \in L^2(\mathbb{R}), \tag{2.7}$$

following from the evenness of the Green's function G.

The BP-equation (1.4) becomes the Benjamin–Ono equation when $\varphi[u]$ is replaced by $-2H[u_x]$, where H is the Hilbert transform

$$H[u] = \text{p.v.}\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{u(y)}{y-x}\,dy,$$

with p.v. denoting the Cauchy principle value. The Benjamin–Ono equation arises in the study of long internal gravity waves in stratified fluids of great depth [2], [26]. It is a completely integrable Hamiltonian system [19] possessing multisoliton solutions [3], [6]. There exists also the analogue of the inverse scattering method [1] and Bäcklund transformations [24]. Although the dispersive regularization by the Hilbert transform is weak compared to KdV (cf. [20]), the dispersion is strong enough that the Benjamin–Ono equation has globally smooth solutions for initial data $u_0 \in H^k(\mathbb{R})$, $k \ge 3/2$ (see [18], [25]), and even for sublinearly growing initial data [16].

The BP-system (1.1), (1.2) is *Galilean invariant*, i.e., invariant under changes of the reference frame of the form

$$x \to x + x_0 + u_0 t, \qquad u \to u + u_0, \qquad \varphi \to \varphi - u_0.$$

Note that the potential transforms like a velocity. The Galilean invariance will simplify the travelling wave analysis in section 3.

The *dispersion relation* of the BP-equation (1.4) linearized at $u = c$ is given by

$$\frac{\omega}{k} = c + \frac{1}{1+k^2}, \tag{2.8}$$

with the frequency ω and the wave number k. The group velocities lie between c and $c+1$, the limits for large and small wave numbers, respectively. The existence of a finite limit for large wave numbers indicates that (1.4) does not have smoothing properties.

Finally, we look for *conservation laws*. Obviously, (1.4) can be written in conservation form:

$$u_t + \left(\frac{u^2}{2} - \varphi[u]\right)_x = 0. \tag{2.9}$$

As a consequence, $\int_{\mathbb{R}} u\,dx$ is conserved for weak solutions with sufficiently strong decay for $x \to \pm\infty$. Moreover, for smooth convex functions Φ, multiplication of (1.4) with Φ' yields for smooth solutions

$$\Phi(u)_t + \Psi(u)_x = \varphi_x[u]\Phi'(u), \tag{2.10}$$

where Ψ is such that $\Psi'(u) = u\Phi'(u)$. In section 4, we prove that weak solutions constructed via a vanishing viscosity approach satisfy, instead of (2.10), entropy inequalities with the equality sign replaced by $\leq$.

We remark that only the choice $\Phi'(u) = u = \varphi_{xx} - \varphi$ allows us to write the right-hand side of (2.10) in conservation form. It leads, for smooth solutions, to the second conservation law

$$(u^2)_t + \left(\frac{2}{3}u^3 + \varphi^2 - \varphi_x^2\right)_x = 0, \tag{2.11}$$

while for weak solutions, the entropy $\int_{\mathbb{R}} u^2\,dx$ is nonincreasing.

The *jump conditions* for entropic shocks with velocity s are those of the Burgers equation:

$$s = \frac{1}{2}(u_l + u_r), \qquad u_l > u_r, \tag{2.12}$$

where u_l and u_r denote the left-sided and, respectively, right-sided limit of u at the shock.

The BP-equation has a *Hamiltonian structure* similar to the Benjamin–Ono equation. The bi-Hamiltonian structure of the Camassa–Holm equation is completely destroyed by dropping the quadratic terms in (2.4). With the Hamiltonian

$$H(u) = \frac{1}{2}\int_{\mathbb{R}} \left(-\varphi[u]u + \frac{u^3}{3}\right) dx,$$

equation (1.4) can be written as

$$u_t + \left(\frac{\delta H}{\delta u}\right)_x = 0,$$

where $\frac{\delta H}{\delta u} = -\varphi[u] + \frac{u^2}{2}$ is the L^2-representation of the Frechet-derivative of H. Note that this relies on the symmetry property (2.7) of $\varphi[\cdot]$. Conservation of the quantity $\int_{\mathbb{R}} H(u)\,dx$ corresponds to the third local conservation law

$$\left(-\varphi u + \frac{u^3}{3}\right)_t + \left(\varphi_x\varphi_t - \varphi\varphi_{xt} - \left(-\varphi + \frac{u^2}{2}\right)^2\right)_x = 0,$$

which (as (2.11)) can be expected to hold only for smooth solutions.

3. Travelling wave analysis. The results of this section should be compared to those of [10], [11] for different versions of the Euler–Poisson system. The qualitative similarities of the results have been one of the main motivations for this work.

By the Galilean invariance it suffices to consider only travelling waves with velocity 0, i.e., steady states. Travelling waves with velocity c are then found by adding the constant c to the velocity u (and $-c$ to the potential φ). After integration of the steady state version of (1.1) and using the result in (1.2), the steady state equations can be written as

$$uu_x = E, \tag{3.1}$$

$$E_x = \frac{u^2}{2} + u - d, \tag{3.2}$$

where we have used the notation $\varphi_x = E$ and d is the constant of integration. The system (3.1), (3.2) will be studied in the (u, E)-phase-plane. We shall also allow shocks (of course with velocity $s = 0$) satisfying the jump conditions (2.12):

$$-u_r = u_l > 0.$$

By (3.2), E is continuous across shocks.

Also worth mentioning is the line of singularities $u = 0$. In general, trajectories end (or begin) there with square root behavior. By (3.1), smooth trajectories can cross $u = 0$ only through the origin of the (u, E)-plane. Our analysis will be restricted to $d > -1/2$, whence there are two stationary points

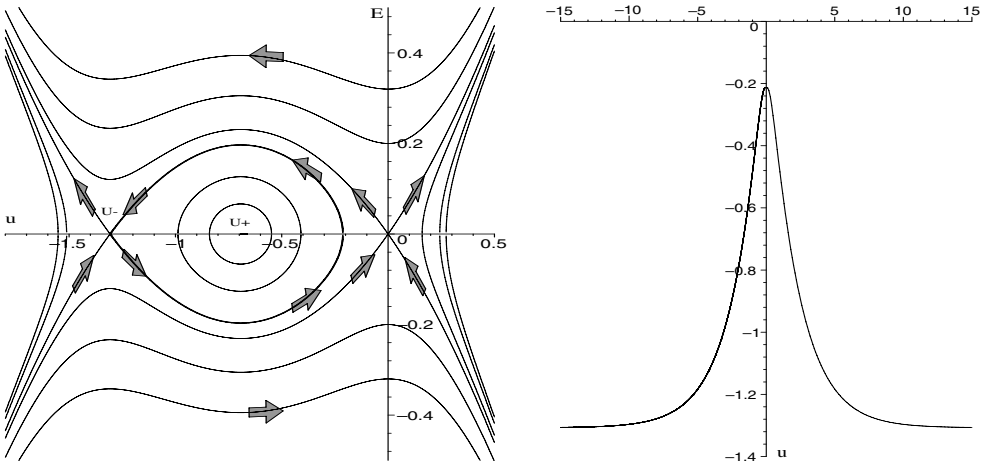
$$P_\pm = \begin{pmatrix} E_\pm \\ u_\pm \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \pm \sqrt{1+2d} \end{pmatrix}.$$

It is easily seen that u_- is always negative and a saddle. The second stationary point u_+ is negative and a center for $-1/2 < d < 0$. It becomes positive and a saddle for $d > 0$. By separation of variables, a first integral of (3.1), (3.2) can be found:

$$A = \frac{E^2}{2} - \frac{u^4}{8} - \frac{u^3}{3} + \frac{du^2}{2}. \tag{3.3}$$

Besides the stationary points, this family of curves (parametrized by A) has the origin as a critical point. Only away from the line $u = 0$ can these curves be seen as trajectories (with opposite orientation on opposite sides of $u = 0$).

Depending on the value of d, three generic cases of phase portraits occur as follows.



(a) Phase portrait for $-\frac{1}{2} < d \leq -\frac{4}{9}$.

(b) Solitary wave.

FIG. 1. *Solitary waves for* $-\frac{1}{2} < d \leq -\frac{4}{9}$.

Solitary waves for $-1/2 < d \leq -4/9$. The phase portraits for $-1/2 < d < -4/9$ are characterized by a homoclinic orbit (pulse, solitary wave) connecting P_- to itself (see Figure 1(a), (b)). Its interior is filled with periodic solutions around P_+. These features are reminiscent of the KdV-equation. By the singularity, the origin is a point of nonuniqueness for the initial value problem. Taylor expansion shows a pair of smooth trajectories passing through the origin. An implicit formula for the solitary waves has already been calculated in [17], together with a numerical simulation of the soliton-like interaction of two solitary waves.

In the critical case $d = -4/9$, the trajectories through the origin coincide with the stable and unstable manifolds of P_-. As a consequence of the nonuniqueness, we can switch from the unstable to the stable manifold at the origin, producing a nonsmooth solitary wave, reminiscent of the peakon solutions of the Camassa–Holm equation. It can be computed explicitly:

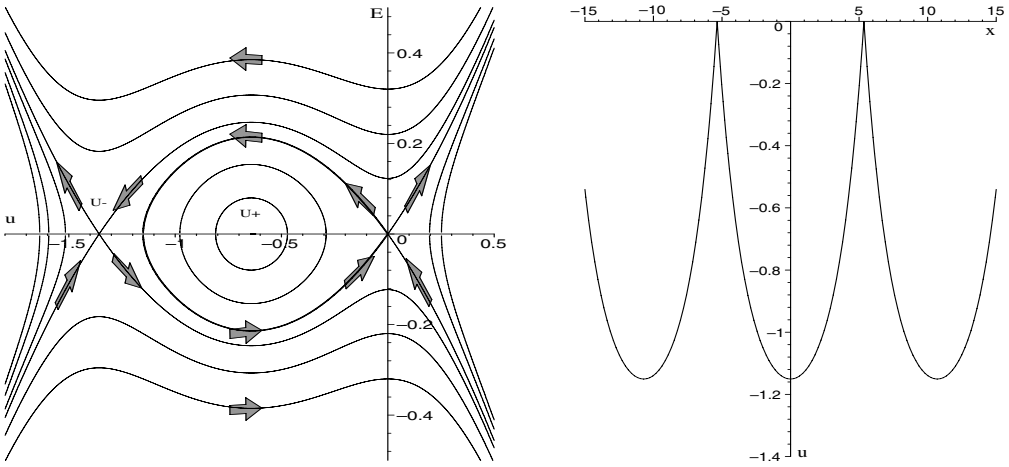
$$u(x) = \frac{4}{3}\left(e^{-|x|/2} - 1\right).$$

Peaked periodic solutions for $-4/9 < d < 0$. In this case the solitary wave disappears, and the trajectories passing through the origin connect to themselves (see Figure 2(a)). This closed curve in the left half plane corresponds to a peaked periodic solution (see Figure 2(b)). Again, these solutions can be computed explicitly. Taking the derivative of (3.1) and using (3.3) for the evaluation of u_x^2, we obtain (with $A = 0$)

$$u_{xx} = \frac{u}{4} + \frac{1}{3}.$$

The peaked periodic solution is given by

$$u(x) = \frac{4}{3}\left(\frac{\cosh(x/2)}{\cosh(p/2)} - 1\right)$$



(a) Phase portrait for $-\frac{4}{9} < d < 0$.

(b) Peaked periodic solution.

FIG. 2. *Periodic solutions for* $-\frac{4}{9} < d < 0$.

for $-p < x < p$ and by periodic continuation with period $2p$. The length of the period is connected to the parameter d by

$$\cosh\left(\frac{p}{2}\right) = \left(1 + \frac{9}{4}d\right)^{-1/2}.$$

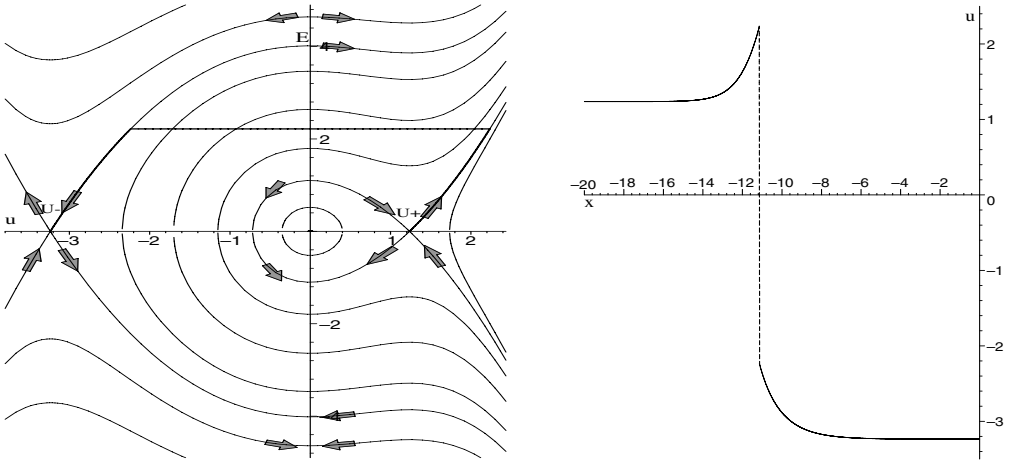
As d tends to zero, p tends to zero and the peaked periodic orbit shrinks to the origin in the phase portrait.

Heteroclinic connections for $d \geq 0$. For $d > 0$, the stationary points are saddles and lie on opposite sides of the line $u = 0$ (see Figure 3(a)). A heteroclinic connection (front wave) between them can be constructed using an entropic shock. There is a unique pair of points $P_l = (u_l, E_l) = (\sqrt{1+2d}, \sqrt{11/12+2d})$, $P_r = (-u_l, E_l)$, satisfying the jump conditions, with P_l lying on the unstable manifold of P_+, and P_r on the stable manifold of P_-. The u-component of the heteroclinic solution is depicted in Figure 3(b). Note that a solution with this qualitative behavior also exists for $d = 0$. In this case, however, the convergence towards P_+ is not exponential, since for $d = 0$, P_+ is a degenerate stationary point.

Remark 3.1. The question arises whether two arbitrary constant states $u_{-\infty}$, u_∞ can be connected by a front wave. The answer is negative. The set of admissible pairs $(u_{-\infty}, u_\infty)$ is constructed by shifting pairs $(u_+(d), u_-(d))$, $d \geq 0$ (exploiting the Galilean invariance). This leads to the requirement $u_\infty - u_{-\infty} \geq 2$.

Transient behavior, numerical experiments. We have studied the transient behavior of solutions of the BP-equation (1.4) numerically. A MATLAB program was written employing a straightforward explicit discretization: In a first step, the Poisson equation is solved for given u at the old time step. A centered difference scheme is used on a bounded interval with boundary conditions $\varphi + u = 0$ at both ends.

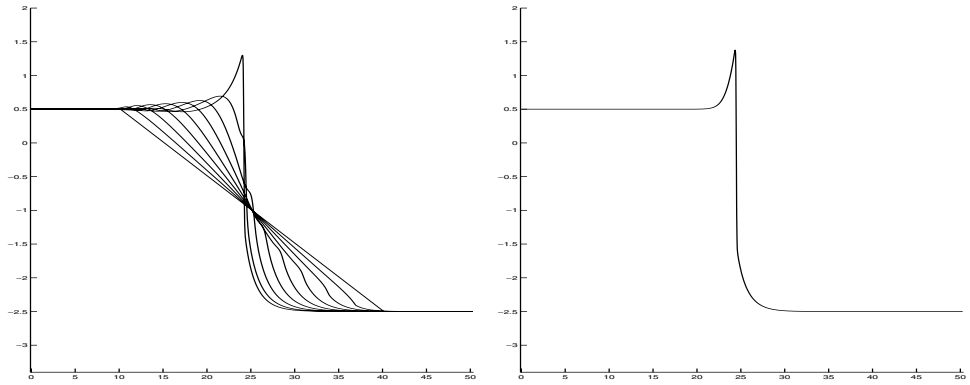
The result is used for the evaluation of the right-hand side of (1.4). Alternatively, we used an implicit spectral method (cf. [7, Part II, Chapter 8]) and obtained very



(a) Phase portrait for $d > 0$.

(b) Shock solution.

FIG. 3. *Shock solutions for* $d \geq 0$.



(a) Shock formation.

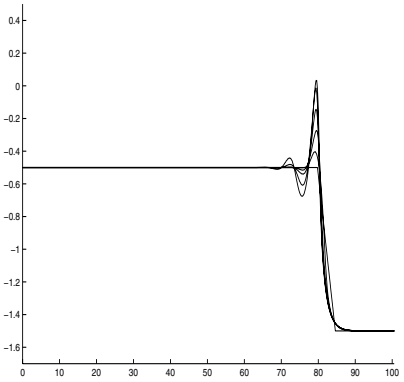
(b) Numerical stationary solution.

FIG. 4. *Initial ramp of height* $3 : 0.5 \searrow -2.5$.

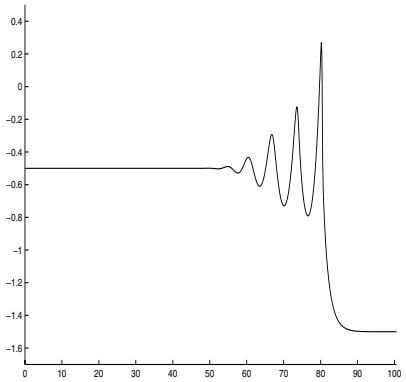
similar results. This spectral method is between three and four times faster due to the use of FFT, but it applies only for spatially periodic situations.

The Burgers flux term is discretized by the Lax–Friedrichs method. Time steps are chosen according to the CFL-condition. As initial data, linear ramps connecting two constant states are prescribed. Recalling Remark 3.1, we are interested in the behavior depending on the difference between the asymptotic states at $x = \pm\infty$.

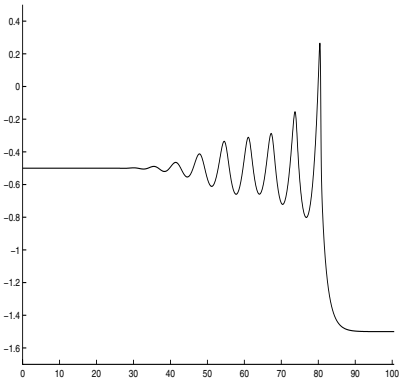
Our results for downward ramps of height larger than 2 suggest the conjecture that the heteroclinic waves constructed above are attractive. For a typical example, see Figure 4(a) and (b). For a ramp with height 3 and the constant states lying symmetric with respect to $u = -1$, we observe numerical convergence to the stationary solution



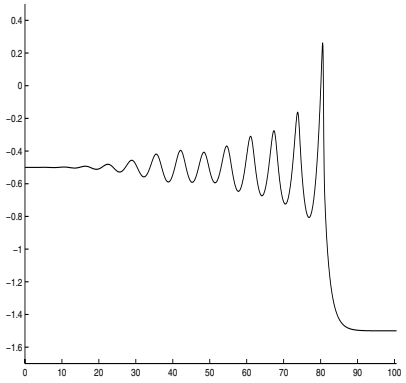
(a) Formation of first oscillations ($t = 0, 2, 4, 6, 10, 15$).



(c) Numerical solution at $t = 40$.



(b) Numerical solution at $t = 80$.



(d) Numerical solution at $t = 120$.

FIG. 5. *Initial ramp of height* $1 : -0.5 \searrow -1.5$.

of the type of Figure 3(b). The development of shocks seems not to depend on the steepness of the ramp.

A completely different behavior is observed for initial ramps with a height less than 2: In this case, there exists no stationary solution connecting the asymptotic states. For a typical example, Figure 5, the observed behavior is reminiscent of the KdV-equation and shows typical dispersive effects with oscillations at the left side of a smoothed ramp. This is in accordance with the dispersion relation (2.8) showing that high frequency components travel with lower velocities.

4. Existence. In this section, the initial value problem

$$u_t + uu_x = \varphi_x[u], \qquad u(x,0) = u_0(x) \tag{4.1}$$

is considered, where the operator $\varphi[\cdot]$ is defined in (1.3).

THEOREM 4.1 (local strong solution). *Assume $u_0 \in H^k(\mathbb{R})$ with $k > \frac{3}{2}$. Then there exists a time $T(\|u_0\|_{H^k(\mathbb{R})}) > 0$ and a unique solution*

$$u \in L^\infty\big((0,T); H^k(\mathbb{R})\big) \cap C\big([0,T]; H^{k-1}(\mathbb{R})\big)$$

of (4.1) *such that $T(\|u_0\|_{H^k(\mathbb{R})})$ depends only on $\|u_0\|_{H^k(\mathbb{R})}$.*

Proof. The proof is based on a contraction argument similar to that of [29, section 16.1]. We define the iteration map S_T as follows: For any function $v \in B_T$ with

$$B_T := \Big\{ w \in L^\infty((0,T), H^k(\mathbb{R})) \cap C([0,T]; H^{k-1}(\mathbb{R})) : \sup_{t\in[0,T]} \|w(\cdot,t)\|_{H^k(\mathbb{R})} \le 2\|u_0\|_{H^k(\mathbb{R})} \Big\}$$

the image $S_T(v)$ is the unique solution u of

$$u_t + uu_x = \varphi_x[v], \qquad u(t=0) = u_0. \tag{4.2}$$

First, we show that S_T maps B_T into itself for T small enough. We apply ∂_x^α to (4.2) for $\alpha \le k$ and take the L^2-scalar product with $\partial_x^\alpha u$:

$$\frac{1}{2}\frac{d}{dt}\big\|\partial_x^\alpha u\big\|^2_{L^2(\mathbb{R})} \quad + \quad \underbrace{\int_{\mathbb{R}} \partial_x^\alpha (uu_x)\, \partial_x^\alpha u \, dx}_{A} \quad = \quad \underbrace{\int_{\mathbb{R}} \varphi_x[\partial_x^\alpha v]\, \partial_x^\alpha u \, dx}_{B}. \tag{4.3}$$

The first factor in the integrand of A is differentiated by the product rule:

$$\partial_x^\alpha (uu_x) = u\, \partial_x^{\alpha+1} u + \sum_{l=1}^{\alpha} \binom{\alpha}{l} \partial_x^l u\, \partial_x^{\alpha+1-l} u.$$

Accordingly, A is split into two parts which are estimated separately:

$$\left| \int_{\mathbb{R}} u\, (\partial_x^{\alpha+1} u)(\partial_x^\alpha u)\, dx \right| = \frac{1}{2} \left| \int_{\mathbb{R}} u\, \partial_x (\partial_x^\alpha u)^2 \, dx \right| \le \frac{1}{2} \|u_x\|_{L^\infty(\mathbb{R})} \|u\|^2_{H^k(\mathbb{R})}. \tag{4.4}$$

For the second part of A, we use the Cauchy–Schwarz inequality and the interpolation estimate

$$\big\| \big(\partial_x^{l-1} f_x\big)\, \big(\partial_x^{\alpha-l} g_x\big) \big\|_{L^2(\mathbb{R})} \le c\, \big(\|f_x\|_{L^\infty(\mathbb{R})} \|g\|_{H^\alpha(\mathbb{R})} + \|f\|_{H^\alpha(\mathbb{R})} \|g_x\|_{L^\infty(\mathbb{R})} \big)$$

(see [29, Chapter 13, Proposition 3.6]) to obtain

$$\left| \int_{\mathbb{R}} \partial_x^\alpha u \sum_{l=1}^{\alpha} \binom{\alpha}{l} \partial_x^{l-1} u_x\, \partial_x^{\alpha-l} u_x \, dx \right| \le c \|u\|^2_{H^k(\mathbb{R})} \|u_x\|_{L^\infty(\mathbb{R})}\,,$$

with some constant c depending only on k. By the Sobolev imbedding $W^{1,\infty}(\mathbb{R}) \hookrightarrow H^k(\mathbb{R})$ for $k > 3/2$, we calculate

$$|A| \le c \|u\|^3_{H^k(\mathbb{R})}. \tag{4.5}$$

For B, we apply the Cauchy–Schwarz inequality and (2.6):

$$|B| \le \|\varphi_x[\partial_x^\alpha v]\|_{L^2(\mathbb{R})} \|\partial_x^\alpha u\|_{L^2(\mathbb{R})} \le c \|v\|_{H^k(\mathbb{R})} \|u\|_{H^k(\mathbb{R})}. \tag{4.6}$$

Using (4.5) and (4.6) in (4.3) gives

$$\frac{d}{dt}\|u\|_{H^k(\mathbb{R})} \le c\left(\|u\|^2_{H^k(\mathbb{R})} + \|v\|_{H^k(\mathbb{R})}\right).$$

For T small enough, a comparison principle shows $\|u(\cdot,t)\|_{H^k(\mathbb{R})} \le 2\|u_0\|_{H^k(\mathbb{R})}$ for $0 \le t \le T$. Since $u \in C([0,T];H^{k-1}(\mathbb{R}))$ is an obvious consequence of (4.2), we conclude that $S_T : B_T \to B_T$. In a second step, we prove S_T to be a strict contraction. Therefore, we consider two functions v_1 and v_2 in B_T and set $u_1 = S_T(v_1)$, $u_2 = S_T(v_2)$, and $u = u_1 - u_2$, $v = v_1 - v_2$. The difference of the equations for u_1, u_2 reads as

$$u_t + u(u_1)_x + u_2 u_x = \varphi_x[v], \qquad u(t=0) = 0.$$

We proceed similarly to (4.2), using B from (4.3):

$$\frac{1}{2}\frac{d}{dt}\|\partial_x^\alpha u\|^2_{L^2(\mathbb{R})} + \underbrace{\int_{\mathbb{R}} \partial_x^\alpha(u\partial_x u_1)\,\partial_x^\alpha u\,dx}_{A_1} + \underbrace{\int_{\mathbb{R}} \partial_x^\alpha(u_2 u_x)\,\partial_x^\alpha u\,dx}_{A_2} = B.$$

In contrast to (4.4), the highest order term of A_1 is not bounded in terms of the $H^\alpha(\mathbb{R})$-norm. Therefore, we are obliged to reduce the order of differentiation to $\alpha \le k-1$ and estimate as above for the second part of A:

$$\begin{aligned}|A_1| &\le c\|u\|_{H^{k-1}(\mathbb{R})}\left(\|\partial_x u_1\|_{L^\infty(\mathbb{R})}\|u\|_{H^{k-1}(\mathbb{R})} + \|u\|_{L^\infty(\mathbb{R})}\|u_1\|_{H^k(\mathbb{R})}\right)\\ &\le c\|u\|^2_{H^{k-1}(\mathbb{R})}\|u_1\|_{H^k(\mathbb{R})}.\end{aligned}$$

For A_2, we proceed as in (4.4)–(4.5):

$$\begin{aligned}|A_2| &= \left|\int_{\mathbb{R}}\left(\frac{u_2\,\partial_x(\partial_x^\alpha u)^2}{2} + \partial_x^\alpha u\sum_{l=1}^{\alpha}\binom{\alpha}{l}\partial_x^l u_2\,\partial_x^{\alpha-l}u_x\right)dx\right|\\ &\le c\|u\|^2_{H^{k-1}(\mathbb{R})}\|u_2\|_{H^k(\mathbb{R})}.\end{aligned}$$

Since $\|u_{1,2}\|_{H^k(\mathbb{R})} \le 2\|u_0\|_{H^k(\mathbb{R})}$, this leads to

$$\frac{d}{dt}\|u\|_{H^{k-1}(\mathbb{R})} \le c\left(\|u\|_{H^{k-1}(\mathbb{R})} + \|v\|_{H^{k-1}(\mathbb{R})}\right),$$

and the Gronwall lemma implies that for T small enough, S_T is a strict contraction with respect to the topology in $C([0,T];H^{k-1}(\mathbb{R}))$. ∎

THEOREM 4.2 (global weak solution). *Assume $u_0 \in BV(\mathbb{R})$. Then there exists a unique global weak solution*

$$u \in L^\infty_{loc}([0,\infty);BV(\mathbb{R})) \tag{4.7}$$

of (4.1)*, satisfying (in the distributional sense) the entropy inequalities*

$$\Phi(u)_t + \Psi(u)_x \le \varphi_x[u]\Phi'(u) \tag{4.8}$$

for convex entropy densities $\Phi \in C^1(\mathbb{R})$ and for Ψ such that $\Psi'(u) = u\Phi'(u)$.

Remark 4.1. For convex functions $\Phi \in W^{1,\infty}(\mathbb{R})$, it is possible to establish generalizations of (4.8) in the sense that the right-hand side of (4.8) has to be interpreted

as the limit of an approximating procedure. More precisely, considering a smoothing $\Phi_\delta \to \Phi$ in $W^{1,\infty}(\mathbb{R})$, the dominated convergence theorem and the upper bound $\varphi_x[u]\|\Phi'\|_{L^\infty}$ imply that the limit $\delta \to 0$ of the right-hand side of (4.8) exists, although it depends on the pointwise limit $\lim_{\delta\to 0} \Phi'_\delta$ since the level sets $\{u = \alpha\}$ may have nonzero measure.

Such problems do not occur in the theory of scalar conservation laws with local flux terms [29] where Kruzhkov's entropy densities $\Phi = |u - \alpha| \in W^{1,\infty}(\mathbb{R})$ for all $\alpha \in \mathbb{R}$ are well defined and allow us, for instance, to show L^1-contractivity (and therefore uniqueness) of the entropy solutions. For the BP-equation (4.1), only L^1-stability can be shown using the above generalizations of (4.8), which is nevertheless sufficient to show the uniqueness.

Proof. The proof is based on a viscosity method similar to that of [29, Chapter 16.6]. Instead of (4.1), we consider the regularized equation

$$u_t + uu_x = \varphi_x[u] + \nu u_{xx} \tag{4.9}$$

with $\nu > 0$. Local existence of a unique smooth solution of the initial value problem for (4.9) with $u(t = 0) = u_0 \in BV(\mathbb{R})$ can be shown by standard arguments. The next step is an L^1-stability result for (4.9). Let u_1, u_2 denote solutions of (4.9) with initial data $f_1, f_2 \in L^1(\mathbb{R})$. Then the difference $v = u_1 - u_2$ satisfies

$$v_t + (wv)_x = \varphi_x[v] + \nu v_{xx}, \qquad v(t = 0) = f_1 - f_2, \tag{4.10}$$

with $w = (u_1 + u_2)/2$. Let $\mathrm{abs}_\delta(\cdot)$ be a convex regularization of the modulus, uniformly converging to $|\cdot|$ as $\delta \to 0$, and satisfying $|\mathrm{abs}'_\delta(v)| \le 1$. Multiplication of (4.10) with $\mathrm{abs}'_\delta(v)$ and integration with respect to x leads to

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}} \mathrm{abs}_\delta(v)\,dx = &\int_{\mathbb{R}} wv\,\mathrm{abs}''_\delta(v)v_x\,dx - \int_{\mathbb{R}} \mathrm{abs}'_\delta(v)\varphi_x[v]\,dx \\ &- \nu\int_{\mathbb{R}} \mathrm{abs}''_\delta(v)\,(v_x)^2\,dx.\end{aligned} \tag{4.11}$$

Since the function $\int_0^v s\,\mathrm{abs}''_\delta(s)\,ds$ converges to zero uniformly as $\delta \to 0$, we have for the first term on the right-hand side of (4.11)

$$\int_{\mathbb{R}} wv\,\mathrm{abs}''_\delta(v)v_x\,dx = -\int_{\mathbb{R}} w_x \int_0^v s\,\mathrm{abs}''_\delta(s)\,ds\,dx \to 0 \qquad \text{as } \delta \to 0.$$

Boundedness of the operator $\varphi_x[\cdot] : L^1(\mathbb{R}) \to L^1(\mathbb{R})$ can be shown easily. With the properties of abs_δ, we obtain from (4.11) in the limit $\delta \to 0$

$$\frac{d}{dt}\|v\|_{L^1(\mathbb{R})} \le c\|v\|_{L^1(\mathbb{R})}. \tag{4.12}$$

Analogously to [29, Lemma 6.1], it can be shown that

$$\frac{d}{dt}\|v\|_{BV(\mathbb{R})} \le c\|v\|_{BV(\mathbb{R})}$$

holds for the solution of (4.9) as a consequence of (4.12). This is sufficient for proving that the solution of (4.9) is global and bounded in $L^\infty_{loc}([0,\infty); BV(\mathbb{R}))$ uniformly in ν. This again is sufficient to show that the solution converges pointwise a.e. on $(0,\infty)\times\mathbb{R}$ as $\nu \to 0$. For the details of the vanishing viscosity limit we refer to [29].

To show that the constructed solutions satisfy the entropy inequalities (4.8) we first consider smooth convex entropy densities Φ and multiply (4.9) with $\Phi'(u(t,x))$. In the weak formulation with nonnegative testfunctions $\theta \in C_0^\infty((0,\infty)\times\mathbb{R})$, integration by parts yields

$$-\int\limits_{(0,\infty)\times\mathbb{R}}\int (\Phi\,\theta_t + \Psi\,\theta_x)\,dtdx = \int\limits_{(0,\infty)\times\mathbb{R}}\int \left(\varphi_x[u]\Phi'(u) - \nu\Phi'' u_x^2 + \nu\Phi_{xx}\right)\theta\,dtdx,$$

where Ψ is as smooth as Φ and satisfies $\Psi' = u\Phi'$. Now, in the vanishing viscosity limit like above, $\Phi(u), \Psi(u)$, and $\varphi_x[u]\Phi'(u)$ converge pointwise a.e. on $(0,\infty)\times\mathbb{R}$ as $\nu \to 0$ by the smoothness of Φ and Ψ. Moreover (see [29]), $\nu\Phi_{xx} \to 0$ weakly, while $-\nu\Phi'' u_x^2 \le 0$ by the convexity of Φ, which implies the inequality sign (4.8). By an approximation argument, (4.8) holds clearly for $\Phi \in C^1(\mathbb{R})$.

For convex $\Phi \in W^{1,\infty}(\mathbb{R})$, Φ' is defined a.e. and monotonically increasing and has therefore at most a countable number of jumps. Hence, after defining Φ' pointwise everywhere in some way such that $|\Phi'|(x) \le \|\Phi'\|_{L^\infty}$, there exist smooth approximations $\Phi_\delta \to \Phi$ in $W^{1,\infty}(\mathbb{R})$ such that $|\Phi_\delta'|(x) \le |\Phi'|(x)$ and $\Phi_\delta' \to \Phi'$ pointwise everywhere. Therefore, considering a particular smoothing Φ_δ, the limit $\delta \to 0$ in the right-hand side of (4.8) is well defined by the dominated convergence theorem using the common upper bound $\varphi_x[u]\|\Phi'\|_{L_\infty}$.

To shown uniqueness of the entropy solution, we choose a particular smoothing Φ_δ of Kruzhkov's entropy densities $\Phi(u) = |u-\alpha|$ for $\alpha \in \mathbb{R}$ such that the limit $\lim_{\delta\to 0}\Phi_\delta'(0) = \operatorname{sign}(0) = 0$. Moreover, we consider two solutions $u(t,x)$ and $\widetilde{u}(s,y)$ and sum, respectively, two generalizations of (4.8), where we set $\alpha = \widetilde{u}(s,y)$ in the first and $\widetilde{\alpha} = u(t,x)$ in the latter. Hence, after estimating $|\operatorname{sign}(u-\alpha)| \le 1$ on the right-hand side, we obtain

$$\begin{aligned}-\iiiint |u-\widetilde{u}|\,(\theta_t+\theta_s)\,dtdxdsdy - \iiiint \Psi(u,\widetilde{u})(\theta_x+\theta_y)\,dtdxdsdy \\ \le \iiiint |\varphi_x[u]-\varphi_y[\widetilde{u}]|\,\theta\,dtdxdsdy,\end{aligned}$$

where $\Psi(u,\widetilde{u}) = \frac{1}{2}(u+\widetilde{u})|u-\widetilde{u}|$ and $\theta = \theta(t,x,s,y)$, a nonnegative testfunction, which we choose especially as

$$\theta(t,x,s,y) = f(t)\,\delta_\varepsilon\left(\frac{t-s}{2}\right)\,g(x)\,\delta_\varepsilon\left(\frac{x-y}{2}\right),$$

where $f, \delta_\varepsilon, g_\varepsilon$ are nonnegative in $C_0^\infty(\mathbb{R})$ and $\delta_\varepsilon(\cdot)$ approximates the delta distribution as $\varepsilon \to 0$, while g_ε is a plateau-function with $0 \le g_\varepsilon \le 1$ and $g_\varepsilon \equiv 1$ on $(-\varepsilon^{-1},\varepsilon^{-1})$. Since

$$\theta_t+\theta_s = f'\,\delta_\varepsilon\left(\frac{t-s}{2}\right)\,g\,\delta_\varepsilon\left(\frac{x-y}{2}\right), \qquad \theta_x+\theta_y = f\,\delta_\varepsilon\left(\frac{t-s}{2}\right)\,g'\,\delta_\varepsilon\left(\frac{x-y}{2}\right),$$

we obtain in the limit $\varepsilon \to 0$

$$-\int\limits_{(0,\infty)\times\mathbb{R}}\int |u(t,x)-\widetilde{u}(t,x)|f'\,dtdx \le \int\limits_{(0,\infty)\times\mathbb{R}}\int |\varphi_x[u]-\varphi_x[\widetilde{u}]|\,f\,dtdx.$$

Since obviously $\int_{\mathbb{R}} |\varphi_x[u-\widetilde{u}]|\,dx \le \|G_x\|_\infty \int_{\mathbb{R}} |u-\widetilde{u}|\,dx$ with $\|G_x\|_\infty = \frac{1}{2}$, we conclude (after integration by parts and by the arbitrariness of f) that

$$\frac{d}{dt}\int_{\mathbb{R}} |u-\widetilde{u}|\,dx \le \frac{1}{2}\int_{\mathbb{R}} |u-\widetilde{u}|\,dx,$$

and the Gronwall lemma implies the uniqueness. □

5. Asymptotics and the quasineutral limit. In this section, we investigate the rescaled $(x \to x/\varepsilon, t \to t/\varepsilon)$ BP-system

$$u_t^\varepsilon + u^\varepsilon u_x^\varepsilon = \varphi_x^\varepsilon, \tag{5.1}$$

$$\varepsilon^2 \varphi_{xx}^\varepsilon = \varphi^\varepsilon + u^\varepsilon, \tag{5.2}$$

with $\varepsilon \ll 1$. In accordance with the terminology taken from the Euler–Poisson system, the limit $\varepsilon \to 0$ will be called the quasi-neutral limit. With the rescaled potential operator

$$\varphi^\varepsilon[u](x) = -\frac{1}{2\varepsilon} \int_{\mathbb{R}} \exp\left(-\frac{|x-y|}{2\varepsilon}\right) u(y)dy,$$

the initial value problem

$$u_t^\varepsilon + u^\varepsilon u_x^\varepsilon = \varphi_x^\varepsilon[u^\varepsilon], \qquad u^\varepsilon(t=0) = u_0, \tag{5.3}$$

will be compared to its formal limit

$$u_t^0 + (u^0+1)u_x^0 = 0, \qquad u^0(t=0) = u_0. \tag{5.4}$$

The limit is the inviscid Burgers equation for the unknown $u^0 + 1$. Even for smooth initial data its solution can develop shocks in finite time. The travelling wave analysis of section 3 can be seen as an attempt to approximate solution profiles in the neighborhood of shocks of (5.4). The heteroclinic connections computed in section 3 are such profiles connecting two states u_l and u_r satisfying the jump conditions

$$s = \frac{1}{2}(u_l + u_r + 2), \qquad u_l > u_r.$$

However, these connections exist only for $u_l - u_r > 2$. For a better understanding of the situation, we expand the potential operator

$$\varphi^\varepsilon[u] = -u - \varepsilon^2 u_{xx} + O(\varepsilon^4).$$

Thus, an $O(\varepsilon^4)$-approximation of (5.3) is given by the KdV equation (for the unknown $u+1$)

$$u_t + (u+1)u_x + \varepsilon^2 u_{xxx} = 0. \tag{5.5}$$

Actually, the KdV equation can be obtained as a formal limit of (5.3). If (5.3) is rescaled by

$$t \to \frac{t}{\varepsilon^2}, \qquad u^\varepsilon \to -1 + \varepsilon^2 U,$$

then the formal limit of the resulting equation for U is

$$U_t + UU_x + U_{xxx} = 0.$$

In analogy to the well known results concerning the limit as $\varepsilon \to 0$ of (5.5) (e.g., [22], [14]), we expect that in general the limits of solutions of (5.3) are weak limits, which do not satisfy the formal limiting equation (5.4). In our last result, however,

we prove that—as long as the solution of the limiting equation remains smooth—it is the strong limit of the solution of (5.3). For $u_0 \in C^{0,1}(\mathbb{R})$, the space of all Lipschitz continuous function on $\mathbb{R}$, there exists a $T > 0$ such that the Burgers equation (5.4) has a solution $u^0 \in C([0,T];\, C^{0,1}(\mathbb{R}))$. Also, if $u_0 \in H^s(\mathbb{R})$, then $u^0 \in L^\infty((0,T);\, H^s(\mathbb{R}))$.

THEOREM 5.1. *Assume $u_0 \in C^{0,1}(\mathbb{R}) \cap H^s(\mathbb{R})$ with $s > 1$. Then, for a time interval $(0,T)$ as mentioned above, the solutions of* (5.3) *and* (5.4) *satisfy*

$$\|u^\varepsilon - u^0\|_{L^\infty((0,T);L^2(\mathbb{R}))} = O\left(\varepsilon^r\right), \qquad r = \min\{2, s-1\}.$$

Proof. Let $v = u^\varepsilon - u^0$ with initial data $v(t=0) = 0$. We subtract (5.4) from (5.3) to obtain an equation for v:

$$v_t + \left(\frac{v^2}{2} + u^0 v\right)_x = u^0_x + \varphi^\varepsilon[u^0_x] = \varepsilon^2 \varphi^\varepsilon_{xx}[u^0_x].$$

By taking the L^2-scalar product with v and by integration by parts, we calculate

$$\frac{1}{2}\frac{d}{dt}\|v\|^2_{L^2(\mathbb{R})} + \int_{\mathbb{R}} \frac{v^2}{2}\, u^0_x\, dx = \varepsilon^2 \int_{\mathbb{R}} v\, \varphi^\varepsilon_{xx}[u^0_x]\, dx,$$

which implies

$$\frac{d}{dt}\|v\|_{L^2(\mathbb{R})} \le \frac{1}{2}\|u^0_x\|_{L^\infty(\mathbb{R})}\|v\|_{L^2(\mathbb{R})} + \|\varepsilon^2 \varphi^\varepsilon_{xx}[u^0_x]\|_{L^2(\mathbb{R})}. \tag{5.6}$$

With the Fourier transform $\widehat{u^0}(k,t)$ of $u^0(x,t)$, the last term can be estimated by

$$\|\varepsilon^2 \varphi^\varepsilon_{xx}[u^0_x]\|_{L^2(\mathbb{R})} = \left\| \frac{\varepsilon^2 |k|^3 |\widehat{u^0}|}{1+\varepsilon^2 k^2} \right\|_{L^2(\mathbb{R})} \le \sup_{k\in\mathbb{R}} \frac{\varepsilon^2 |k|^3}{(1+\varepsilon^2 k^2)(1+k^2)^{s/2}} \|u^0\|_{H^s(\mathbb{R})}.$$

The factor on the right-hand side is obviously $O(\varepsilon^2)$ for $s \ge 3$. For $s < 3$, it can be estimated by

$$\frac{\varepsilon^2 |k|^3}{(1+\varepsilon^2 k^2)|k|^s} = \varepsilon^{s-1} \frac{|\varepsilon k|^{3-s}}{1+|\varepsilon k|^2} = O(\varepsilon^{s-1}),$$

and thus

$$\|\varepsilon^2 \varphi^\varepsilon_{xx}[u^0_x]\|_{L^2(\mathbb{R})} \le c\varepsilon^r \|u^0\|_{H^s(\mathbb{R})}.$$

The statement of the theorem is now a direct consequence of the Gronwall lemma applied to (5.6). ∎

REFERENCES

[1] M. J. ABLOWITZ AND A. S. FOKAS, *The inverse scattering transform for the Benjamin–Ono equation—A pivot to multidimensional problems*, Stud. Appl. Math., 68 (1983), pp. 1–10.

[2] T. B. BENJAMIN, *Internal waves of permanent form in fluids of great depth*, J. Fluid Mech., 29 (1967), pp. 559–592.

[3] T. L. BOCK AND M. D. KRUSKAL, *A two-parameter Miura transformation of the Benjamin–Ono equation*, Phys. Lett. A, 74 (1979), pp. 173–176.

[4] R. CAMASSA AND D. D. HOLM, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.

[5] R. Camassa, D. D. Holm, and J. M. Hyman, *A new integrable shallow water equation*, Adv. Appl. Mech., 31 (1994), pp. 1–33.
[6] K. M. Case, *Benjamin–Ono related equations and their solutions*, Proc. Natl. Acad. Sci. USA, 76 (1979), pp. 1–3.
[7] C. Cercignani and D. H. Sattinger, *Scaling Limits and Models in Physical Processes*, DMV Seminar Band 28, Birkhäuser Verlag, Basel, 1998.
[8] A. Constantin, *On the Cauchy problem for the periodic Camass–Holm equation*, J. Differential Equations, 141 (1997), pp. 218–235.
[9] A. Constantin and J. Escher, *Well-posedness, global existence, and blowup phenomena for a periodic quasi-linear hyperbolic equation*, Comm. Pure Appl. Math., 51 (1998), pp. 475–504.
[10] S. Cordier, P. Degond, P. Markowich, and C. Schmeiser, *Travelling wave analysis and jump relations for the Euler–Possion model in the quasineutral limit*, Asymptot. Anal., 11 (1995), pp. 209–240.
[11] S. Cordier, P. Degond, P. Markowich, and C. Schmeiser, *Travelling wave analysis of an isothermal Euler–Poisson model*, Ann. Fac. Sci. Toulouse Math. (6), 5 (1996), pp. 599–643.
[12] S. Cordier and E. Grenier, *Quasineutral limit of an Euler–Poisson system arising from plasma physics*, Comm. Partial Differential Equations, 25 (2000), pp. 1099–1113.
[13] S. Engelberg, H. Liu, and E. Tadmor, *Critical thresholds in Euler–Poisson equations*, Indiana Univ. Math. J., 50 (2001), pp. 109–157.
[14] H. Flaschka, M. Forest, and D. McLaughlin, *Multiphase averaging and the inverse spectral solution of the Korteweg–de Vries equation*, Comm. Pure Appl. Math., 33 (1980), pp. 739–784.
[15] A. Fokas and B. Fuchssteiner, *Sympletic structures, their Bäcklund transformations and hereditary symmetries.*, Phys. D, 4 (1981), pp. 47–66.
[16] G. Fonseca and F. Linares, *Benjamin–Ono equation with unbounded data*, J. Math. Anal. Appl., 247 (2000), pp. 426–447.
[17] B. Fornberg and G. B. Whitham, *A numerical and theoretical study of certain nonlinear wave phenomena*, Phil. Trans. Roy. Soc. London A, 227 (1978), pp. 373–404.
[18] R. J. Iório, *On the Cauchy problem for the Benjamin–Ono equation*, Comm. Partial Differential Equations, 11 (1986), pp. 1031–1081.
[19] D. J. Kaup, T. I. Lakoba, and Y. Matsuno, *Complete integrability of the Benjamin–Ono equation by means of action-angle variables*, Phys. Lett. A, 238 (1998), pp. 123–133.
[20] C. E. Kenig, G. Ponce, and L. Vega, *On the generalized Benjamin–Ono equation*, Trans. Amer. Math. Soc., 342 (1994), pp. 155–172.
[21] C. Lattanzio and P. Marcati, *Global well-posedness and relaxation limits of a model for radiating gas*, J. Differential Equations, 190 (2003), pp. 439–465.
[22] P. D. Lax and D. Levermore, *The zero dispersion limit for the Korteweg-de Vries KdV equation*, Proc. Natl. Acad. Sci. USA, 76 (1979), pp. 3602–3606.
[23] H. Liu and E. Tadmor, *Critical thresholds in a convolution model for nonlinear conservation laws*, SIAM J. Math. Anal., 33 (2001), pp. 930–945.
[24] A. Nakamura, *Bäcklund transformations and conservation laws of the Benjamin–Ono equation*, J. Phys. Soc. Japan, 47 (1979), pp. 1335–1340.
[25] G. Ponce, *On the global well-posedness of the Benjamin–Ono equation*, Differential Equations, 4 (1991), pp. 527–542.
[26] H. Ono, *Algebraic solitary waves in stratified fluids*, J. Phys. Soc. Japan, 39 (1975), pp. 1082–1091.
[27] G. Rodríguez-Blanco, *On the Cauchy problem for the Camassa–Holm equation*, Nonlinear Anal., 46 (2001), pp. 309–327.
[28] S. Schochet and E. Tadmor, *Regularized Chapman–Enskog expansion for scalar conservation laws*, Arch. Ration. Mech. Anal., 119 (1992), pp. 95–107.
[29] M. E. Taylor, *Partial Differential Equation* III*, Nonlinear Equations*, Springer Series Appl. Math. Sci. 117, Springer-Verlag, New York, 1996.
[30] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley & Sons, New York, 1974.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1526–1549

LOW-FIELD LIMIT FOR A NONLINEAR DISCRETE DRIFT-DIFFUSION MODEL ARISING IN SEMICONDUCTOR SUPERLATTICES THEORY*

THIERRY GOUDON†, OSCAR SÁNCHEZ‡, JUAN SOLER‡, AND LUIS L. BONILLA§

Abstract. Charge transport in semiconductor superlattices can be described through a discrete drift-diffusion model. In this model, we identify some small parameter $h > 0$, related to the ratio between the length of a superlattice period and the observation length scale. Specifically, we investigate a regime where the length of the superlattice period is small while the doping profile is low. In the limit $h \to 0$, we are led to a nonlinear drift-diffusion model, coupled to the Poisson equation.

Key words. semiconductor supperlattices, drift-diffusion models

AMS subject classifications. 35Q99, 35K55, 82C70

DOI. 10.1137/S003613990241730X

1. Introduction.

A semiconductor superlattice (SL) is a periodic array of layers of two different semiconductors whose lateral dimension is much larger than the length of one period. These devices exhibit nonlinear charge transport phenomena due to the existence of electric field domains. Depending on the charge density (produced by doping or irradiating the SL) and on the applied voltage, different qualitative responses of the current can be obtained, as shown, e.g., by Bonilla [6]. In experiments at intermediate values of the charge density, stationary responses and self-sustained oscillations are observed depending on the values of the voltage. There exist solutions corresponding to low voltages which are stationary and typically develop low electric fields. It is very important to understand the time evolution of the solutions toward these stationary profiles (see [5] where relocation experiments are studied).

Electronic transport in such semiconductor devices can be described by a discrete drift-diffusion model. Details on the modeling will be given in section 2, following the works by Aguado, Platero, Moscoso, and Bonilla [1] and Bonilla, Platero, and Sánchez [4] and review papers by Bonilla [6] and Wacker [15]. The model consists of the Poisson equation coupled to charge continuity equations for the electron density n and average electric field F at each SL period. Tunneling currents across barriers are approximated by a discrete drift-diffusion (DDD) law, whose coefficients are themselves field dependent. We aim at investigating asymptotics regimes for this model. To this end, we shall write the equations in dimensionless form. Hence, we are able to identify some small parameter—denoted $h > 0$ in what follows—by means of physically relevant dimensionless parameters of the DDD system. Having set up this DDD system, we prove that the solutions converge, in an appropriate weak setting, to solutions of a continuous drift-diffusion-Poisson problem with field-dependent mobilities,

*Received by the editors November 5, 2002; accepted for publication (in revised form) November 3, 2003; published electronically June 22, 2004. This research was partially supported by the EU financed network IHP-HPRN-CT-2002-00282 and by MCYT (Spain), Proyecto BFM2002–00831.
http://www.siam.org/journals/siap/64-5/41730.html

†CNRS-UMR8524, Université des Sciences et Technologies Lille 1, Cité scientifique, 59655 Villeneuve d'Ascq cedex, France (thierry.goudon@univ-lille1.fr).

‡Departamento de Matematica Aplicada, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain (ossanche@ugr.es, jsoler@ugr.es).

§Departamento de Matemáticas, Escuela Politécnica Superior, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Spain, and Unidad Asociada al Instituto de Ciencia de Materiales (CSIC), 28049 Cantoblanco, Spain (bonilla@ing.uc3m.es).

as the parameter h tends to 0. The limit equation reads

$$\text{(1.1)}\qquad \begin{cases} \partial_t n + \partial_x J(F,n) = 0 & \text{in } (0,T)\times[-X,X], \\ J(F,n) = v(F)n - D(F)\partial_x n\,, & \\ \partial_x F = n - N_D & \text{in } (0,T)\times[-X,X]. \end{cases}$$

These equations are completed by bias, boundary, and initial conditions

$$\begin{cases} \displaystyle\int_{-X}^{X} F\,dx = V & \text{on } (0,T), \\ J(F,n)(X) = W^{(f)}(F)n(X) & \text{on } (0,T), \\ J(F,n)(-X) = (j^{(e)}(F) - W^{(b)}(F)n)(-X) & \text{on } (0,T), \\ n(t=0,x) = n^0(x) & \text{on } [-X,X]. \end{cases}$$

The techniques employed to prove the convergence are based on a priori estimates and compactness properties. At this point, we have to remark that the solutions to the DDD model are stepwise constant functions, which forces us to consider the solutions in the framework of the bounded variation (BV) spaces.

Our main result admits a reversal lecture. Since the dimensionless system of the DDD model coincides with a finite difference discretization of (1.1), the analysis proposed can be read as a convergence analysis for numerical approximations of that drift-diffusion continuous equation. In this direction our problem is related with other works on approximation of field-dependent mobilities. System (1.1) is a monopolar one-dimensional version of the drift-diffusion system analyzed by Gajewski and Gröger [8] and Jerome [10]. Our DDD system is simpler than the general version studied in those works, where different boundary conditions are considered. Here, we deal with the time-dependent problem, while, to the best of our knowledge, the previous analyses are devoted to approximate steady state solutions.

This paper is organized as follows. In section 2 we present in detail the DDD model, recalling some aspects of its derivation. We also justify its well-posedness. In section 3 we perform the dimension analysis of the system. Then, we derive the dimensionless equations for which the analysis is actually performed and we state precisely our main convergence result for the DDD model. Actually, our analysis applies either when considering a Dirichlet boundary condition for the electric field or with the bias condition. Section 4 is devoted to the crucial a priori estimates satisfied by the solution. For the sake of simplicity, we start with the Dirichlet boundary condition. Then, in section 5 we use these estimates to show the convergence to the continuous model, through compactness arguments. Finally, section 6 sketches the slight adaptation of the proof to treat the DDD system endowed with the bias condition. The paper ends with two appendices: in the first one we deal with a technical auxiliary result, and the latter investigates uniqueness of the limit system.

2. Discrete drift-diffusion model. Since the two semiconductors constituting the SL have different energy gaps, the conduction band of an SL can be viewed as a periodic array of potential wells and barriers, of widths w and d, respectively, with $\ell = d + w$ the length of one period. We assume that scattering times are shorter than escape times from quantum wells, the latter being shorter than dielectric relaxation times. In such a weakly coupled semiconductor SL, the dominant mechanism of charge transport is sequential resonant tunneling. In the simplest situation, the center of each quantum well is n-doped and the thermal energy is large compared to the energy of the lowest miniband. Then, a description of charge transport in such devices has been

proposed through a DDD model; see [1, 4, 6, 15]. This model has been extended by taking into account stochastic effects by Bonilla, Sánchez, and Soler [5], in comparison with the experimental results of Rogozia et al. [12].

In such a modeling, we consider an array of $2N+1$ consecutive cells, which are well-barrier pairs, labeled by the index $i \in \{-N, \dots, +N\}$. The barrier separating the injecting contact from the first well of the SL is considered as the $(-N-1)$th barrier, while the barrier of the Nth SL period separates the Nth well from the collector. The model assumes that the electrons are singularly concentrated on a two-dimensional region located in the center of the quantum well. The unknowns are the two-dimensional electron density n_i (number of electrons per unit area of the SL cross section at the center of the ith well) and the average electric field F_i in each cell. These quantities are related through the following discrete Poisson equation:

$$F_i - F_{i-1} = \frac{e}{\overline{\varepsilon}}(n_i - N_D^w), \quad i \in \{-N, \dots, N\}. \tag{2.1}$$

In (2.1), N_D^w stands for the two-dimensional doping in the wells, assumed to be constant, while $\overline{\varepsilon}$ is the average permittivity in the SL and $(-e)$ stands for the electron charge. Notice that the set of relations (2.1) involves as an additional unknown the electric field F_{-N-1} at the injecting contact. On the other hand, denoting by $eJ_{i\to i+1}$ the tunneling current density through the barrier separating the cells $\#i$ and $\#(i+1)$, the density in the ith cell satisfies the following charge continuity equation:

$$\frac{dn_i}{dt} = J_{i-1\to i} - J_{i\to i+1}, \quad i \in \{-N, \dots, N\}. \tag{2.2}$$

Consequently, differentiating (2.1) and using (2.2), we notice that the quantity

$$\frac{\overline{\varepsilon}}{e}\frac{dF_i}{dt} + J_{i\to i+1} = J(t), \quad i \in \{-N-1, \dots, N\} \tag{2.3}$$

does not depend on the considered cell. This is the so-called Ampère's law, where $eJ(t)$ stands for the total current density through the SL which does not depend on the index i.

Then, the model is completed by a constitutive law which defines the current density $eJ_{i\to i+1}$ by means of the (n_k, F_k)'s. The tunneling current density depends on the electrochemical potentials at cells $\#i$ and $\#(i+1)$ and on the average electric field F_i; see [6, 15]. The electrochemical potentials that "drive" the tunneling current (a nonzero current is a consequence of unequal electrochemical potentials at cells $\#i$ and $\#(i+1)$) are functions of the electron densities and therefore, according to [6, 15], we may consider that the tunneling current $eJ_{i\to i+1}$ depends on n_i, n_{i+1} and F_i. First-principles calculations of $eJ_{i\to i+1}$ are at best sketchy. In the literature, formulas have been derived from quantum kinetic equations for Green's functions (see [15], assuming constant electric field across the SL, simplified hopping Hamiltonians, and scattering) and from the transfer Hamiltonian formalism as in [1, 4, 6] (a many-body version of the WKB method originally proposed by Bardeen [3]). At high (room) temperature, all these formulas imply that the tunneling current is given by the difference of a drift term and a diffusion term as follows:

$$J_{i\to i+1} = \frac{n_i v(F_i)}{\ell} - \frac{D(F_i)(n_{i+1} - n_i)}{\ell^2}, \quad i \in \{-N, \dots, N-1\}. \tag{2.4}$$

The drift velocity and the diffusion coefficient are defined through functions v and D of the electric field, which depend on the physical properties of the material used in the SL; see [6] for more details. The special nature of the three-dimensional emitter and collector layers (different from the essentially two-dimensional quantum wells that form the SL) is considered in the calculation of the boundary tunneling current. By using the transfer Hamiltonian formalism, the following approximate formulas can be derived [4]:

$$J_{-N-1\to -N} = j^{(e)}(F_{-N-1}) - \frac{n_{-N} W^{(b)}(F_{-N-1})}{\ell}, \tag{2.5}$$

$$J_{N\to N+1} = \frac{n_N W^{(f)}(F_N)}{\ell}. \tag{2.6}$$

These equations involve the emitter current density $ej^{(e)}$, the emitter backward velocity $W^{(b)}$, and the collector forward velocity $W^{(f)}$, which are given functions of the electric field. All the coefficients $v, D, W^{(b)}, W^{(f)}, j^{(e)}$ are supposed to be nonnegative and satisfy some regularity properties. Typical graphs for these functions can be found in [5].

We remark that one equation is still missing since we have one more unknown than we have equations. There are several ways to close the system. The simplest way is to assume that the electric field at the emitter is prescribed as

$$F_{-N-1}(t) = F_-(t), \tag{2.7}$$

the right-hand side being a given function $F_- : \mathbb{R}^+ \to \mathbb{R}$. This Dirichlet boundary condition has been proposed when the number of periods considered in the SL is high enough (infinite supperlattice). Therefore, this condition is well adapted to our work since we shall deal with an asympotic problem where the number of cells goes to infinity.

However, from a physical viewpoint, it is certainly more realistic to complete the system by using the so-called voltage bias condition: the total voltage across the SL,

$$\ell \sum_{i=-N}^{N} F_i = V, \tag{2.8}$$

remains equal to a given quantity V. In what follows we essentially deal with the Dirichlet-like boundary condition (2.7) for the electric field. We will come back to the voltage bias condition (2.8) at the end of the paper.

Relations (2.1), (2.2), and (2.7) form a closed system of equations for n_i and F_i with $i \in \{-N, \dots, N\}$, referred to in what follows as the DDD model. We remark that the electric field in the cell $\#i$ can be expressed as a function of the incoming field F_- and the density in the previous cells as follows:

$$F_i(t) = F_-(t) + \frac{e}{\bar{\varepsilon}} \sum_{j=-N}^{i} (n_j(t) - N_D^w), \quad i \in \{-N, \dots, N\} \quad \forall t \in [0, T]. \tag{2.9}$$

Consequently, we can rewrite the initial value problem associated to the DDD model in terms of the densities

$$\frac{d\vec{n}}{dt} = g(t, \vec{n}(t)), \qquad \vec{n}(0) = \vec{n}^0, \tag{2.10}$$

where $\vec{n}(t) = (n_{-N}, \dots, n_N)^T \in \mathbb{R}^{2N+1}$, $g : \mathbb{R}^{2N+1} \to \mathbb{R}^{2N+1}$ is a smooth function, and $\vec{n}^0 \in \mathbb{R}^{2N+1}$ is the initial condition.

THEOREM 2.1. *Let $n_i^0 \geq 0$ for $i \in \{-N, \dots, N\}$ be the initial data for the DDD system. Let F_- be a C^1 function of time. Also let $v, D, W^{(b,f)}, j^{(e)}$ be C^1 nonnegative functions. Then, there exists a unique global solution associated with the initial value problem* (2.10). *The solution verifies $n_i(t) \geq 0$ for all $i \in \{-N, \dots, N\}$, $t \geq 0$.*

Proof. Local existence and uniqueness follow by a direct application of the Cauchy–Lipschitz theorem for ODE, since the function g inherits the regularity properties of the coefficients. The estimates proved in section 4, especially in Lemma 4.1, also provide a uniform bound on the solution which prevents a finite time blowup. Consequently, the solution is globally defined. There remains only to justify the nonnegativeness of the solution. To this end, it is convenient to rewrite (2.2) as a difference between a gain term and a loss term as follows:

$$\frac{dn_i}{dt} = \begin{cases} \frac{v(F_{i-1})}{\ell} n_{i-1} + \frac{D(F_i)}{\ell^2} n_{i+1} + \frac{D(F_{i-1})}{\ell^2} n_{i-1} - \Big(\frac{v(F_i)}{\ell} + \frac{D(F_i)}{\ell^2} + \frac{D(F_{i-1})}{\ell^2} \Big) n_i \\ \qquad\qquad \text{for } i \in \{-N+1, \dots, N-1\}, \\ \frac{D(F_{-N})}{\ell^2} n_{-N+1} + j^{(e)}(F_{-N-1}) - \Big(\frac{v(F_{-N})}{\ell} + \frac{D(F_{-N})}{\ell^2} + \frac{W^{(b)}(F_{-N-1})}{\ell} \Big) n_{-N} \\ \qquad\qquad \text{for } i = -N, \\ \frac{v(F_{N-1})}{\ell} n_{N-1} + \frac{D(F_{N-1})}{\ell^2} n_{N-1} - \Big(\frac{D(F_{N-1})}{\ell^2} + \frac{W^{(f)}(F_N)}{\ell} \Big) n_N \\ \qquad\qquad \text{for } i = N. \end{cases}$$

Let $t \geq 0$ such that $n_i(t) \geq 0$ for any $i \in \{-N, \dots, N\}$. Suppose $n_j(t) = 0$ for some $j \in \{-N, \dots, N\}$. Thus, we notice that its time derivative $\frac{dn_j}{dt}(t)$ is nonnegative and, hence, we deduce the nonnegative character of the solution along the time evolution. ∎

3. Dimensionless equations. The aim of this section is to write the system in dimensionless form. Hence, we will identify some dimensionless physical parameters. Next, we appropriately order these parameters in terms of a quantity $h > 0$ intended to tend to 0. Studying the limit $h \to 0$ we obtain a nonlinear continuous drift-diffusion model, as described in the introduction. This approach relating discrete to continuous models is reminiscent of hydrodynamic limits in kinetic theory (see [9]). Actually, it has been used for models of phase transition, for example, in [7].

Let us introduce time and length units, respectively, denoted by $\mathcal{T}$ and $\mathcal{L}$. They correspond to observation scales. We also need characteristic values for the electron density and for the electric field, respectively, denoted by $\mathcal{N}$ and $\mathcal{F}$. For instance, it is quite natural to define $\mathcal{N}$ from the doping profile N_D^w and $\mathcal{F}$ from the emitter field F_-. Then, using the convention that overlined quantities are dimensionless, we set

$$\begin{cases} \mathcal{N}\, \overline{n_i}(\bar{t}) = n_i(\mathcal{T}\bar{t}), & \mathcal{N}\, \overline{N_D} = N_D^w, \\ \mathcal{F}\, \overline{F_i}(\bar{t}) = F_i(\mathcal{T}\bar{t}), & \mathcal{F}\, \overline{F_-}(\bar{t}) = F_-(\mathcal{T}\bar{t}), \\ \dfrac{\mathcal{L}}{\mathcal{T}}\, \overline{v}(\overline{F}) = v(\mathcal{F}\overline{F}), & \dfrac{\mathcal{L}}{\mathcal{T}}\, \overline{W^{(b,f)}}(\overline{F}) = W^{(b,f)}(\mathcal{F}\overline{F}), \\ \dfrac{\mathcal{L}^2}{\mathcal{T}}\, \overline{D}(\overline{F}) = D(\mathcal{F}\overline{F}), & \dfrac{\overline{\varepsilon}\mathcal{F}}{e}\, \dfrac{1}{\mathcal{T}}\, \overline{j^{(e)}}(\overline{F}) = j^{(e)}(\mathcal{F}\overline{F}). \end{cases}$$

Note that the emitter current density has been scaled with respect to the density $\frac{\overline{\varepsilon}}{e}\mathcal{F}$ instead of with respect to $\mathcal{N}$ (the other choice is also possible; the proof adapts immediately and the emitter current density disappears as $h \to 0$ in that case).

Therefore, we are led to the continuity equations in the following dimensionless form:

$$\frac{d\overline{n}_i}{d\bar{t}} = \frac{\mathcal{L}}{\ell}\Big(\overline{v}(\overline{F}_{i-1})\overline{n}_{i-1} - \overline{v}(\overline{F}_i)\overline{n}_i - \frac{\mathcal{L}}{\ell}\overline{D}(\overline{F}_{i-1})(\overline{n}_{i-1} - \overline{n}_i) + \frac{\mathcal{L}}{\ell}\overline{D}(\overline{F}_i)(\overline{n}_i - \overline{n}_{i+1})\Big)$$

for $i \in \{-N+1, \dots, N-1\}$ and

$$\begin{aligned}
\frac{d\overline{n}_{-N}}{d\bar{t}} &= \frac{\mathcal{L}}{\ell}\Big(\frac{\ell}{\mathcal{L}}\frac{\overline{\varepsilon}\mathcal{F}}{\epsilon\mathcal{N}}\bar{j}^{(e)}(\overline{F}_{-N-1}) - \overline{n}_{-N}\overline{W}^{(b)}(\overline{F}_{-N-1}) \\
&\qquad -\overline{v}(\overline{F}_{-N})\overline{n}_{-N} - \frac{\mathcal{L}}{\ell}\overline{D}(\overline{F}_{-N})(\overline{n}_{-N} - \overline{n}_{-N+1})\Big), \\
\frac{d\overline{n}_N}{d\bar{t}} &= \frac{\mathcal{L}}{\ell}\Big(\overline{v}(\overline{F}_{N-1})\overline{n}_{N-1} + \frac{\mathcal{L}}{\ell}\overline{D}(\overline{F}_{N-1})(\overline{n}_{N-1} - \overline{n}_N) - \overline{n}_{-N}\overline{W}^{(f)}(\overline{F}_N)\Big).
\end{aligned}$$

On the other hand, the Poisson equation reads

$$\frac{\overline{\varepsilon}\mathcal{F}}{e\mathcal{N}}\,(\overline{F}_i - \overline{F}_{i-1}) = (\overline{n}_i - \overline{N}_D)$$

for $i \in \{-N, \dots, N\}$.

In these expressions, we identify two dimensionless parameters

$$\alpha = \frac{\overline{\varepsilon}\mathcal{F}}{e\mathcal{N}}, \quad \beta = \frac{\mathcal{L}}{\ell}.$$

Roughly speaking, we go from the discrete equations to a continuous description by interpreting the difference between consecutive cells as differential quotients. It means that we shall consider the situation

$$\alpha = \beta = \frac{1}{h} \gg 1,$$

where h is a positive quantity intended to tend to 0. (Actually, we might suppose, with some obvious adaptations in the proofs, that $\alpha = \frac{1}{h} \gg 1$, and $\frac{\alpha}{\beta}$ has a finite positive limit.) Coming back to the physical meaning, the ordering for β means that the size of the cells is small compared to the observation length scale $\ell \ll \mathcal{L}$, while the ordering for α is an assumption about the data; the doping profile N_D^w is small compared to the density $\frac{\overline{\varepsilon}}{e}F_-$ associated with the electric field at the injecting contact ($\mathcal{N} \ll \frac{\overline{\varepsilon}}{e}\mathcal{F}$). Furthermore, we shall assume that the total length of the SL is given and is equal to $2X$, so that the number of cells in the SL also should be appropriately rescaled. Namely, the number of cells is defined in terms of the parameter $h > 0$ by

$$N^h = X/h.$$

The limit performed in this paper is motivated by the comparison between the profiles of the drift velocity and of the diffusion coefficient. Figure 3.1 shows these profiles for a 9nm/4nm GaAs/AlAs SL at $5K$, while the inset picture enlarges these coefficients in the low-field range. In this region the diffusion coefficient is larger than the drift velocity, which is close to zero; i.e., $v(F) \ll D(F)/\ell$ holds. This implies that the diffusion coefficient is large (order h^{-2}) in comparison to the drift velocity (order h^{-1}). Accordingly, we call this asymptotic approach low-field limit. A continuum limit also can be performed in a regime in which $v(F) \approx D(F)/\ell$; this high-field regime will be investigated in a forthcoming work. A complementary interpretation of our asymptotic analysis can be given in terms of a parameter (the so-called Lorentzian half-width) defining the Lorentzian functions involved in the expression of the coefficients of the DDD model; see [4]. The smaller the Lorentzian half-width, the lower the field.

A stationary solution for the DDD model can be obtained in the low-field range as shown by the dotted line in Figure 3.1 (right). In this experiment, we have applied

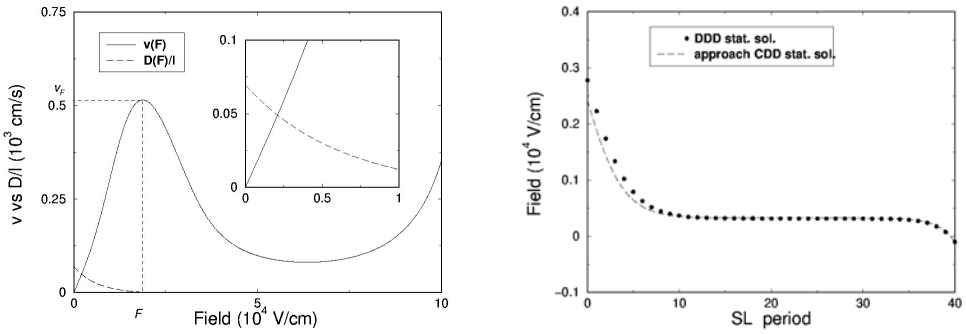


FIG. 3.1. *Left: Drift velocity versus diffusion coefficient for a 9nm/4nm GaAs/AlAs SL. Inset: Detailed low-field range. Right: Electric field distribution for a stationary solution of 40 periods 9nm/4nm GaAs/AlAs SL (dots) and numerical approach to the continuum drift-diffusion model (dashed line) with bias constraint $V = 0.52$ V.*

voltage 0.52 V, $N_D^w = 0.05 \times 10^{11}$ cm^{-2} and contact doping $N_D = 0.2 \times 10^{18}$ cm^{-3}. The Lorentzian half-width involved in the computation of the coefficients of the DDD model is 1 eV and the other parameters are equal to those used in [5]. Thus, $\mathcal{N} = N_D^w$, $\mathcal{F} \approx 0.3\,10^4$ V/cm, $v \approx 0.001\,10^3$ cm/s, and $D/\ell \approx 0.076\,10^3$ cm/s. This leads to the values $\alpha = 3.29$ and $\beta = 7.67$; a sequence of values of the same order can be obtained (low-field limit) by modifying V.

Let us summarize the low-field problem we are interested in as follows. We drop the overlines and emphasize the dependence of the solution (n, F) with respect to the parameter h by a superscript. Hence, we consider the system

$$\frac{dn_i^h}{dt} = \frac{1}{h}(J_{i-1\to i}^h - J_{i\to i+1}^h), \quad i \in \{-N^h, \dots, N^h\}, \tag{3.1}$$

coupled to

$$F_i^h - F_{i-1}^h = h(n_i^h - N_D), \quad i \in \{-N^h, \dots, N^h\}, \tag{3.2}$$

with $F_{-N^h-1}^h = F_-$ given. Note that, coming back to (2.9), we also have

$$F_i^h(t) = F_-(t) + h \sum_{j=-N^h}^{i} (n_j^h(t) - N_D), \quad i \in \{-N^h, \dots, N^h\}. \tag{3.3}$$

Here, we used the following definition for the tunneling currents:

$$\begin{cases} J_{i\to i+1}^h = n_i^h v_i^h - \dfrac{1}{h} D(F_i^h)(n_{i+1}^h - n_i^h), \quad i \in \{-N^h, \dots, N^h - 1\}, \\ J_{-N^h-1\to -N^h}^h = j^{(e)}(F_-) - n_{-N^h}^h W^{(b)}(F_-), \\ J_{N^h\to N^h+1}^h = n_{N^h} W^{(f)}(F_{N^h}^h). \end{cases}$$

The idea is to investigate the limit as $h \to 0$.

To this end, we set $I = (-X, +X) = (-N^h h, N^h h)$ and we associate to the unknowns $(n_{-N^h}^h, \dots, n_{N^h-1}^h) \in \mathbb{R}^{2N^h}$ and $(F_{-N^h}^h, \dots, F_{N^h-1}^h) \in \mathbb{R}^{2N^h}$, the stepwise

constant functions $n^h(t,x)$ and $F^h(t,x)$ defined almost everywhere on $[0,\infty)\times I$ by saying

$$n^h(t,x)=n_i^h(t),\quad F^h(t,x)=F_i^h(t),\quad ih<x<(i+1)h,\quad i\in\{-N^h,\dots,N^h-1\}.$$

Note that it is not relevant to define these functions on the negligible set of points $\{ih,\, i\in\{-N^h,\dots,N^h\}\}$; note also that $F_-, n^h_{N^h}, F^h_{N_h}$ seem to play no role in these definitions. However, they will be used in the definition of traces in the limit $h\to 0$. As a consequence of these definitions, we shall use that sums of n_i^h or F_i^h can be considered as integrals: for example, for any function $\psi:\mathbb{R}\to\mathbb{R}$ we have

$$\int_{-X}^{+X}\psi(n^h)\,dx=h\sum_{i=-N^h}^{N^h-1}\psi(n_i^h)\,,$$

because n_i^h is constant on $ih<x<(i+1)h$. Then, passing to a continuous variable, it is tempting to interpret finite differences as differential quotients. Following this rough idea, we formally guess that the limiting problem corresponding to $h\to 0$ consists of the following nonlinear drift-diffusion equation:

$$\left\{\begin{array}{ll}
\partial_t n+\partial_x J(F,n)=0 & \text{in } (0,T)\times I,\\
J(F,n)=v(F)n-D(F)\partial_x n, & \\
\partial_x F=n-N_D & \text{in } (0,T)\times I,\\
F(-X)=F_- & \text{on } (0,T),\\
J(F,n)(X)=W^{(f)}(F)n(X) & \text{on } (0,T),\\
J(F,n)(-X)=(j^{(e)}(F)-W^{(b)}(F)n)(-X) & \text{on } (0,T),\\
n(t=0,x)=n^0(x) & \text{on } I.
\end{array}\right.\tag{3.4}$$

Thus, the main result of the paper is the following.

THEOREM 3.1. *Let $v,D,W^{(b,f)},j^{(e)}:\mathbb{R}\to\mathbb{R}$ be continuous and nonnegative functions. Suppose that $D(F)>0$ and $W^{(b,f)}(F)>0$ for any $F\in\mathbb{R}$. Let $F_-\in \mathrm{C}^1(\mathbb{R}^+)$. Let $n^{h,0}=(n^{h,0}_{-N^h},\dots,n^{h,0}_{N^h})\in\mathbb{R}^{2N^h+1}$ be the initial data for the rescaled problem. We suppose that $n_i^{h,0}\ge 0$ satisfy*

$$\sup_{h>0}\left(h\sum_{i=-N^h}^{N^h}|n_i^{h,0}|^2\right)\le C_0<\infty.\tag{3.5}$$

Let (n^h,F^h) be the associated solution of (3.1), (3.2). *Then, up to a subsequence, we have*

$$\left\{\begin{array}{ll} n^h\to n & \text{strongly in } \mathrm{L}^2((0,T)\times I) \text{ and in } \mathrm{C}^0([0,T];\mathrm{L}^2(I)-\text{weak}),\\ F^h\to F & \text{uniformly in } [0,T]\times\overline{I}.\end{array}\right.$$

Furthermore, the limits satisfy $n\in\mathrm{L}^2(0,T;\mathrm{H}^1(I))$, $F\in\mathrm{C}^0([0,T]\times\overline{I})$ and solve the nonlinear problem (3.4) *in the sense that*

$$\frac{d}{dt}\int_{-X}^{X}n\phi\,dx=\int_{-X}^{X}J(F,n)\phi'\,dx+W^{(f)}(F)n\phi(X)+(j^{(e)}(F)-W^{(b)}(F)n)\phi(-X)$$

holds in $\mathcal{D}'(0,T)$ for any test function $\phi\in\mathrm{C}^\infty(\overline{I})$, coupled to the Poisson equation

$$\partial_x F=n-N_D,\quad F(-X)=F_-$$

also considered in the sense of the distributions.

This kind of nonlinear parabolic equation, coupled to the Poisson equation, has been investigated by Liang [11]. Actually, in [11] the diffusion coefficient is constant and the boundary conditions are slightly different. In the convergence proof, we need only to assume the continuity of the coefficients; however, using locally Lipschitz properties of the coefficients, we can prove the uniqueness of solution for (3.4); see Appendix B. Consequently, assuming the convergence of the initial data, in Theorem 3.1 the entire sequence converges.

A stationary solution (continuous line) for the continuous drift-diffusion model (CDD) has been obtained in Figure 3.1 (right). The corresponding stationary solution for the DDD model, with the same voltage, can be seen now as a numerical approach to that of the CDD model with $h \in [\frac{1}{7.67}, \frac{1}{3.29}]$. We can observe that the agreement between the solutions to the discrete model and the continuous one is better at the inner periods, where the low-field hypothesis plays a determinant role. The difference between both profiles in the emitter region comes from the fact that the simulations have been done under bias constraint.

4. A priori estimates. This section is devoted to the derivation of the crucial estimates on the solutions (n^h, F^h) that will lead us to rigorously perform the limit $h \to 0$. We assume that the initial data $n_i^{h,0} \geq 0$ satisfies (3.5). This implies that the $\mathrm{L}^1[-X, X]$-norm is bounded as follows:

$$h \sum_{i=-N^h}^{N^h} n_i^{h,0} \leq \left(h \sum_{i=-N^h}^{N^h} |n_i^{h,0}|^2 \right)^{1/2} \sqrt{(2N^h + 1)h}$$

is bounded independently of $h \in (0, 1)$. We recall that

$$\left\{ \begin{array}{l} D, W^{(b,f)}, j^{(e)}, v \in \mathrm{C}^0(\mathbb{R}), \\ v(F) \geq 0, \quad j^{(e)} \geq 0, \\ W^{(b,f)}(F) > 0, \quad D(F) > 0. \end{array} \right. \tag{4.1}$$

We split our argument into several steps. We shall use the convention that C_T stands for a constant possibly depending on T and on the data $F_-, j^{(e)}, W^{(b,f)}$, or on the estimates (3.5), but not on h. Also, we denote as usual by $\mathcal{M}(I)$ the set of Radon measures on the open interval I. Elements of $\mathcal{M}(I)$ identify with distributions Φ on I satisfying $|\langle \Phi, \varphi \rangle| \leq C \|\varphi\|_{L^\infty(I)}$ for all $\varphi \in C_c^\infty(I)$ for some $C > 0$ being independent of the support of the test function (see, e.g., [13]). As usual we denote by $\mathrm{BV}(I)$ the set of bounded variation functions, i.e., functions which are in $\mathrm{L}^1(I)$ and such that their distributional derivative belongs to $\mathcal{M}(I)$.

LEMMA 4.1 (L^1 estimate on the density). *The sequence n^h is bounded in* $\mathrm{L}^\infty(0, T; \mathrm{L}^1(I))$.

Proof. Summing up the equations in (3.1), we obtain

$$\begin{aligned} h \frac{d}{dt} \sum_{i=-N^h}^{N^h} n_i^h &= \sum_{i=-N^h}^{N^h} (J_{i-1\to i}^h - J_{i\to i+1}^h) = J_{-N^h-1\to -N^h}^h - J_{N^h\to N^h+1}^h \\ &= j^{(e)}(F_-) - n_{-N^h}^h W^{(b)}(F_-) - n_{N^h}^h W^{(f)}(F_{N^h}^h). \end{aligned}$$

Therefore, integrating with respect to time and using $n_i^h \geq 0$ and $W^{(b,f)} \geq 0$, we find

$$
\begin{aligned}
& h \sum_{i=-N^h}^{N^h} n_i^h(t) + \int_0^t n_{-N^h}^h W^{(b)}(F_-(s))\,ds + \int_0^t n_{N^h}^h W^{(f)}(F_{N^h}^h)(s)\,ds \\
& = h \sum_{i=-N^h}^{N^h} n_i^{h,0} + \int_0^t j^{(e)}(F_-(s))\,ds \leq C_0 + \|j^{(e)}(F_-)\|_{L^1(0,T)} \leq C_T\,,
\end{aligned} \tag{4.2}
$$

which concludes the proof. □

LEMMA 4.2 (estimates on the electric field). *The sequence F^h is bounded in $\mathrm{L}^\infty((0,T)\times I)$ and in $\mathrm{L}^\infty(0,T;\mathrm{BV}(I))$, while $F_{N_h}^h$ is bounded in $\mathrm{L}^\infty(0,T)$.*

Proof. We combine the estimate in Lemma 4.1 with the identity (3.3) to yield

$$
\begin{aligned}
|F_i^h(t)| &= \left| F_-(t) + h \sum_{j=-N^h}^{i} (n_j^h(t) - N_D) \right| \\
&\leq |F_-(t)| + h \sum_{j=-N^h}^{i} n_j^h(t) + h(i + N^h + 1)N_D \\
&\leq |F_-(t)| + h \sum_{j=-N^h}^{N^h} n_j^h(t) + (2X + h)N_D \leq C_T\,,
\end{aligned}
$$

which proves that F^h is bounded in $L^\infty((0,T)\times I)$ and implies the estimate on $F_{N^h}^h$.

Next, let $\phi \in \mathrm{C}_0^\infty(I)$ be a test function. We have

$$
\begin{aligned}
\langle \partial_x F^h, \phi \rangle &= -\int_{-X}^{X} F^h(t,x)\phi'(x)\,dx = -\sum_{i=-N^h}^{N^h-1} F_i^h \int_{ih}^{(i+1)h} \phi'(x)\,dx \\
&= \sum_{i=-N^h}^{N^h-1} F_i^h(\phi(ih) - \phi((i+1)h)) \\
&= \sum_{i=-N^h}^{N^h} \Big((F_i^h - F_{i-1}^h)\ \phi(ih) \Big) + F_{-N^h-1}^h \phi(-N^h h) - F_{N^h}^h \phi(N^h h) \\
&= h \sum_{i=-N^h}^{N^h} \Big((n_i^h - N_D)\ \phi(ih) \Big) + F_-\phi(-X) - F_{N^h}^h \phi(X),
\end{aligned}
$$

where we have used (3.2). Hence, by using the above bounds we deduce that the following estimate,

$$
|\langle \partial_x F^h, \phi \rangle| \leq \|\phi\|_{\mathrm{L}^\infty(I)} \left(h \sum_{i=-N^h}^{N^h} n_i^h + (2X + h)N_D \right) \leq \|\phi\|_{\mathrm{L}^\infty(I)}\ C_T\,,
$$

holds. This proves that $\partial_x F^h$ is bounded in $\mathrm{L}^\infty(0,T;\mathcal{M}(I))$. □

Remark 4.3. Since the functions $W^{(b,f)}$ and D are continuous and positive in $\mathbb{R}$, the uniform bound on F_i^h guarantees that

$$\begin{cases} \inf\limits_{h>0, i\in\{-N^h,\dots,N^h\}, 0\le t\le T} D(F_i^h(t)) \ge \delta > 0, \\ \inf\limits_{h>0, 0\le t\le T} W^{(f)}(F_{N^h}^h(t)) \ge \delta > 0, \\ \inf\limits_{0\le t\le T} W^{(b)}(F_-(t)) \ge \delta > 0 \end{cases}$$

for some $\delta > 0$. Coming back to (4.2), we deduce that the boundary terms $n_{\pm N^h}$ are bounded in $\mathrm{L}^1(0,T)$. Similarly, there exists $0 < M < \infty$ such that

$$\begin{cases} \sup\limits_{h>0, i\in\{-N^h,\dots,N^h\}, 0\le t\le T} |D(F_i^h)| \le M, \\ \sup\limits_{h>0, i\in\{-N^h,\dots,N^h\}, 0\le t\le T} |v(F_i^h)| \le M, \\ \sup\limits_{h>0, 0\le t\le T} |W^{(f)}(F_{N_h}^h)| \le M, \\ \sup\limits_{h>0, 0\le t\le T} |W^{(b)}(F_-^h)| \le M, \\ \sup\limits_{0\le t\le T} |j^{(e)}(F_-)| \le M. \end{cases}$$

LEMMA 4.4 (L^2 estimate on the density). *The sequence n^h is bounded in $\mathrm{L}^\infty(0,T;\mathrm{L}^2(I))$. The "boundary terms" $n^h_{\pm N^h}$ are bounded in $\mathrm{L}^2(0,T)$. Moreover, we have*

$$\int_0^T \sum_{i=-N^h}^{N^h-1} \frac{|n_{i+1}^h - n_i^h|^2}{h}\, ds \le C_T\,.$$

Proof. Multiplying (3.1) by n_i^h and summing over i, we obtain

$$\begin{aligned} \frac{h}{2}\frac{d}{dt}\sum_{i=-N^h}^{N^h} |n_i^h|^2 &= \sum_{i=-N^h}^{N^h} (J^h_{i-1\to i} - J^h_{i\to i+1}) n_i^h \\ &= \sum_{i=-N^h}^{N^h-1} J^h_{i\to i+1}(n_{i+1}^h - n_i^h) + J^h_{-N^h-1\to -N^h} n^h_{-N^h} - J^h_{N^h\to N^h+1} n^h_{N^h} \\ &= \sum_{i=-N^h}^{N^h-1} \left(n_i^h v(F_i^h) - \frac{1}{h} D(F_i^h)(n_{i+1}^h - n_i^h) \right) (n_{i+1}^h - n_i^h) \\ &\qquad + j^{(e)}(F_-) n^h_{-N^h} - |n^h_{-N^h}|^2 W^{(b)}(F_-) - |n_{N^h}|^2 W^{(f)}(F^h_{N^h}). \end{aligned}$$

By using Remark 4.3, we deduce the inequality

$$\frac{h}{2}\sum_{i=-N^h}^{N^h}|n_i^h(t)|^2+\delta\int_0^t\left(\sum_{i=-N^h}^{N^h-1}\frac{|n_{i+1}^h-n_i^h|^2}{h}+|n_{-N^h}^h|^2+|n_{N^h}|^2\right)ds$$
$$\leq\frac{h}{2}\sum_{i=-N^h}^{N^h}|n_i^h(0)|^2+M\int_0^t\left(\sum_{i=-N^h}^{N^h-1}n_i^h|n_{i+1}^h-n_i^h|+n_{-N^h}^h\right)ds\,.$$

Now, by using the Young inequality we estimate

$$\int_0^t\sum_{i=-N^h}^{N^h-1}n_i^h|n_{i+1}^h-n_i^h|\,ds\leq\frac{2Mh}{\delta}\int_0^t\sum_{i=-N^h}^{N^h}|n_i^h|^2\,ds+\frac{\delta}{2M}\int_0^t\sum_{i=-N^h}^{N^h-1}\frac{|n_{i+1}^h-n_i^h|^2}{h}\,ds.$$

It follows that

$$\frac{h}{2}\sum_{i=-N^h}^{N^h}|n_i^h(t)|^2+\frac{\delta}{2}\int_0^t\sum_{i=-N^h}^{N^h-1}\frac{|n_{i+1}^h-n_i^h|^2}{h}\,ds+\delta\int_0^t(|n_{-N^h}^h|^2+|n_{N^h}|^2)\,ds$$
$$\leq\frac{h}{2}\sum_{i=-N^h}^{N^h}|n_i^h(0)|^2+\frac{2M^2}{\delta}\int_0^t\left(h\sum_{i=-N^h}^{N^h}|n_i^h|^2\right)ds+M\int_0^t n_{-N^h}^h\,ds.$$

We conclude the proof by applying the Gronwall inequality and by taking into account that $n_{-N^h}^h$ is bounded in $\mathrm{L}^1(0,T)$ (see Remark 4.3). ∎

In order to study the limit in boundary terms we consider the next statement.

LEMMA 4.5 (H^1 estimate of the electric field at the boundary). *The sequence $F_{N^h}^h$ is bounded in $\mathrm{H}^1(0,T)$.*

Proof. We have proved that $F_{N^h}^h$ is bounded in $\mathrm{L}^\infty(0,T)$. There remains to bound its time derivative in $\mathrm{L}^2(0,T)$. This is a consequence of (3.3) together with the estimates in Lemma 4.2 and 4.4. Indeed, we get (see the argument given in Lemma 4.1)

$$\begin{aligned}\left|\frac{d}{dt}F_{N^h}^h(t)\right| &= \left|\frac{d}{dt}F_-+\frac{d}{dt}\left(h\sum_{i=-N^h}^{N^h}(n_i^h-N_D)\right)\right|\\ &= \left|\frac{d}{dt}F_-+j^{(e)}(F_-)-n_{-N^h}^hW^{(b)}(F_-)-n_{N^h}^hW^{(f)}(F_{N^h}^h)\right|\\ &\leq \left\|\frac{d}{dt}F_-\right\|_{\mathrm{L}^\infty(0,T)}+M(1+n_{-N^h}^h+n_{N^h}^h).\end{aligned}$$

By Lemma 4.4 the right-hand side is bounded in $\mathrm{L}^2(0,T)$, which ends the proof. ∎

LEMMA 4.6 (BV estimate on the density). *The sequence n^h is bounded in $\mathrm{L}^2(0,T;\mathrm{BV}(I))$.*

Proof. Once the L^2 estimate on n^h is known, we derive some bounds for $\partial_x n^h$.

Consider $\phi \in \mathrm{C}_0^\infty(I)$. We have

$$
\begin{aligned}
|\langle \partial_x n^h, \phi\rangle| &= \left| -\int_{-X}^{X} n^h \phi'\, dx \right| = \left| -\sum_{i=-N^h}^{N^h-1} n_i^h \int_{ih}^{(i+1)h} \phi'\, dx \right| \\
&= \left| -\sum_{i=-N^h}^{N^h-1} n_i^h \ (\phi((i+1)h) - \phi(ih)) \right| \\
&= \left| \sum_{i=-N^h+1}^{N^h} (n_i^h - n_{i-1}^h)\ \phi(ih) + n_{-N^h}^h \phi(-N^h h) - n^h N^h \phi(N^h h) \right| \\
&\le \left(h \sum_{i=-N^h+1}^{N^h} |\phi(ih)|^2 \right)^{1/2} \left(\frac{1}{h} \sum_{i=-N^h+1}^{N^h} |n_i^h - n_{i-1}^h|^2 \right)^{1/2} \\
&\le \|\phi\|_{\mathrm{L}^\infty(I)}\ (2X)^{1/2} \left(\sum_{i=-N^h}^{N^h-1} \frac{|n_{i+1}^h - n_i^h|^2}{h} \right)^{1/2}. \qquad (4.3)
\end{aligned}
$$

Lemma 4.4 implies that the $\mathrm{L}^2(0,T)$-norm of the right-hand side of (4.3) is bounded uniformly with respect to h. Hence, we conclude that $\partial_x n^h$ is in $\mathrm{L}^2(0,T;\mathcal{M}(I))$. ∎

LEMMA 4.7 (estimate on the time derivative). *The sequences $\partial_t n^h$ and $\partial_t F^h$ are bounded in $\mathrm{L}^2(0,T;\mathcal{M}(I)) + \mathrm{L}^2(0,T;\mathrm{W}^{-1,1}(I))$ and in $\mathrm{L}^2(0,T;\mathcal{M}(I))$, respectively.*

Proof. Let $\phi \in \mathrm{C}_0^\infty(I)$ and denote

$$\Gamma_i^h = \int_{ih}^{(i+1)h} \phi(x)\, dx$$

for $i \in \{-N^h, \dots, N^h - 1\}$. Since the support of ϕ is included in I, we can extend Γ_i^h by 0 for $i \ge N^h$. We shall use the following basic estimates:

$$
\begin{cases}
|\Gamma_i^h| \le h\ \|\phi\|_{\mathrm{L}^\infty(I)}, \\
|\Gamma_{i+1}^h - \Gamma_i^h| \le h^2\ C\|\phi'\|_{\mathrm{L}^\infty(I)}.
\end{cases}
$$

Now we estimate the time derivative of the electric field by using the Ampère equations (2.3). We have

$$
\begin{aligned}
\langle \partial_t F^h, \phi\rangle &= \sum_{i=-N^h}^{N^h-1} \frac{d}{dt} F_i^h \int_{ih}^{(i+1)h} \phi(x)\, dx \\
&= J^h \sum_{i=-N^h}^{N^h-1} \Gamma_i^h - \sum_{i=-N^h}^{N^h-1} J_{i\to i+1}^h \Gamma_i^h = I_1 + I_2, \qquad (4.4)
\end{aligned}
$$

where $J^h(t)$ stands for the total current density, which is defined by the $(-N^h-1)$th Ampère equation,

$$J^h(t) = \frac{d}{dt} F_- + J_{-N^h-1\to -N^h}^h = \frac{d}{dt} F_- + j^{(e)}(F_-) - W^{(b)}(F_-) n_{-N^h}^h.$$

By Lemma 4.4, this quantity is bounded in $\mathrm{L}^2(0,T)$. Therefore, the first term of the right-hand side of (4.4) is bounded by

$$|I_1| \le \|\varphi\|_{\mathrm{L}^\infty(I)}\ 2hN^h\ |J^h| = \|\varphi\|_{\mathrm{L}^\infty(I)}\ 2X\ |J^h|,$$

which belongs to a bounded set in $\mathrm{L}^2(0,T)$. Next, I_2 is estimated as follows:

$$\begin{aligned}
|I_2| &\le \left| \sum_{i=-N^h}^{N^h-1} n_i^h v(F_i^h)\ \Gamma_i^h \right| + \left| \sum_{i=-N^h}^{N^h-1} \frac{1}{h} D(F_i^h)(n_i^h - n_{i+1}^h)\ \Gamma_i^h \right| \\
&\le M\ \|\phi\|_{\mathrm{L}^\infty(I)}\ h \left(\sum_{i=-N^h}^{N^h-1} n_i^h + \sum_{i=-N^h}^{N^h-1} \frac{|n_i^h - n_{i+1}^h|}{h} \right) \\
&\le M\ \|\phi\|_{\mathrm{L}^\infty(I)} \left(h \sum_{i=-N^h}^{N^h-1} n_i^h + \sqrt{2hN^h \sum_{i=-N^h}^{N^h-1} \frac{|n_i^h - n_{i+1}^h|^2}{h}} \right).
\end{aligned}$$

We conclude that $\partial_t F^h$ is bounded in $\mathrm{L}^2(0,T;\mathcal{M}^1(I))$.

Similarly, we deal with the time derivative of n^h. We have

$$\begin{aligned}
|\langle \partial_t n^h, \varphi \rangle| &= \left| \sum_{i=-N^h}^{N^h-1} \frac{dn_i^h}{dt} \int_{ih}^{(i+1)h} \phi(x)\,dx \right| = \left| \frac{1}{h} \sum_{i=-N^h}^{N^h-1} (J_{i-1\to i}^h - J_{i\to i+1}^h)\ \Gamma_i^h \right| \\
&= \frac{1}{h} \left| \sum_{i=-N^h}^{N^h-1} J_{i\to i+1}^h\ (\Gamma_{i+1}^h - \Gamma_i^h) + J_{-N^h-1\to -N^h}^h \Gamma_{-N^h}^h - J_{N^h-1\to N^h}^h \Gamma_{N^h}^h \right| \\
&\le \frac{1}{h} \left| \sum_{i=-N^h}^{N^h-1} v(F_i^h) n_i^h\ (\Gamma_{i+1}^h - \Gamma_i^h) \right| + \frac{1}{h^2} \left| \sum_{i=-N^h}^{N^h-1} D(F_i^h)(n_i^h - n_{i+1}^h)(\Gamma_{i+1}^h - \Gamma_i^h) \right| \\
&\quad + \frac{1}{h} |j^{(e)}(F_-) - n_{-N^h}^h W^{(b)}(F_-)|\ |\Gamma_{-N^h}^h| \\
&\le C\ h^2\ \|\phi'\|_{\mathrm{L}^\infty(I)} \left(\frac{M}{h} \sum_{i=-N^h}^{N^h-1} n_i^h + \frac{M}{h^2} \sum_{i=-N^h}^{N^h-1} |n_i^h - n_{i+1}^h| \right) \\
&\quad + h\ \|\phi\|_{\mathrm{L}^\infty(I)}\ \frac{M}{h}(1 + n_{-N^h}^h) \\
&\le C\ \|\phi'\|_{\mathrm{L}^\infty(I)} \left(h \sum_{i=-N^h}^{N^h-1} n_i^h + \sqrt{2hN^h \sum_{i=-N^h}^{N^h-1} \frac{|n_i^h - n_{i+1}^h|^2}{h}} \right) \\
&\quad + \|\phi\|_{\mathrm{L}^\infty(I)}\ M(1 + n_{-N^h}^h),
\end{aligned}$$

which proves the estimate on $\partial_t n^h$. ∎

5. Continuous model. Let us combine the estimates discussed in the previous section with the following classical compactness result (see, e.g., [2], [14]).

PROPOSITION 5.1. *Consider Banach spaces B, X, and Y. We suppose that $X \subset B \subset Y$, the first embedding being compact. Let $\mathcal{C}$ be a bounded set in $\mathrm{L}^p(0,T;X)$, $1 \le p \le \infty$. Assume that $\partial_t \mathcal{C} = \{\partial_t f,\ f \in \mathcal{C}\}$ is a bounded set in $\mathrm{L}^r(0,T;Y)$. Then, $\mathcal{C}$ is relatively compact in $\mathrm{L}^p(0,T;B)$ if $1 \le p < \infty$ and $r \ge 1$, or in $\mathrm{C}^0([0,T];B)$ if $p = \infty$ and $r > 1$.*

Hence, from the previous estimates we have, possibly at the cost of extracting subsequences, that

$$\left\{\begin{array}{ll} n^h \to n & \text{strongly in } \mathrm{L}^2((0,T)\times I) \text{ and in } \mathrm{C}^0([0,T];L^2(I)-\textit{weak}),\\ \partial_x n^h \rightharpoonup \partial_x n & \text{weakly}-* \text{ in } \mathrm{L}^2(0,T;\mathcal{M}(I)),\\ F^h \to F & \text{strongly in } \mathrm{C}^0([0,T];L^p(I)) \text{ for any } 1 \le p < \infty, \end{array}\right. \tag{5.1}$$

as h goes to 0. Notice in particular that the convergence of traces in time makes sense and

$$n^h(t,x)_{|t=0} = n^{h,0}(x) \rightharpoonup n^0(x) = n(t,x)_{|t=0} \text{ weakly in } \mathrm{L}^2(I)$$

holds, with $n^{h,0}(x) = n_i^h$ for $ih < x < (i+1)h$, $i \in \{-N^h, \dots, N^h - 1\}$. In other words, we recover the initial condition in the limit $h \to 0$. Finally, we can also guarantee from Lemmas 4.4 and 4.5 the following properties:

$$\left\{\begin{array}{ll} n^h_{\pm N^h} \rightharpoonup n_\pm & \text{weakly in } \mathrm{L}^2(0,T),\\ F^h_{N^h} \to F_+ & \text{uniformly in } \mathrm{C}^0([0,T]). \end{array}\right. \tag{5.2}$$

We first get the continuous Poisson equation.

PROPOSITION 5.2. *The electric field limit F and the density limit n satisfy the continuous Poisson equation*

$$\partial_x F = n - N_D, \quad F_{|x=-X} = F_-$$

in a weak sense.

Remark 5.3. The Poisson relation with $n \in \mathrm{L}^2((0,T)\times I)$ implies, by the Sobolev embedding, that F is in $\mathrm{L}^2(0,T;\mathrm{C}^0(\bar I))$ so that the traces of F are well defined.

Proof. Let $\phi \in C^\infty(\bar I)$ and $\phi_i^h = \phi(ih)$ for $i \in \{-N^h, \dots, N^h\}$. We denote by ϕ^h the associated stepwise constant function. For the sake of simplicity it will be convenient to also introduce the stepwise constant function $\nabla^h(\phi)(x) = \frac{\phi^h_{i+1} - \phi^h_i}{h}$ for $x \in (ih, (i+1)h)$. Multiplying (3.2) by ϕ_i^h, we get

$$\begin{aligned} h \sum_{i=-N^h}^{N^h} \frac{F_i^h - F_{i-1}^h}{h}\, \phi_i^h &= h \sum_{i=-N^h}^{N^h} (n_i^h - N_D)\, \phi_i^h \\ &= \int_{-X}^{X} (n^h - N_D)\, \phi^h\, dx + h\, (n^h_{N^h} - N_D)\, \phi(X) \\ &= h \sum_{i=-N^h}^{N^h-1} F_i^h\, \frac{\phi_i^h - \phi_{i+1}^h}{h} + F^h_{N^h}\phi(X) - F_-\phi(-X) \\ &= -\int_{-X}^{X} F^h\, \nabla^h(\phi)\, dx + F^h_{N^h}\phi(X) - F_-\phi(-X). \end{aligned}$$

Since $\nabla^h(\phi)$ converges uniformly to $\phi'(x)$ on $\overline{I}$, we have

$$\int_{-X}^{X} (n^h - N_D)\ \phi^h\, dx \to -\int_{-X}^{X} F\ \phi'(x)\, dx + F_+\phi(X) - F_-\phi(-X)$$

as $h \to 0$.

We conclude that $\partial_x F = n - N_D \in \mathrm{L}^2((0,T)\times I)$ and, by the Sobolev embedding, F lies in $\mathrm{L}^2(0,T;\mathrm{C}^0(\overline{I}))$ and the traces of F are well defined and are given by $F(t,\pm X) = F_\pm(t)$. ∎

Let us now show that the limit n is more regular than n^h is. In fact, we will prove that $n \in \mathrm{L}^2(0,T;\mathrm{H}^1(I))$, which guarantees that $n \in \mathrm{L}^2(0,T;\mathrm{C}^0(\overline{I}))$ due to the Sobolev embedding, so that the traces of the limit n with respect to the space variable are also well defined.

PROPOSITION 5.4. *The density limit n of n^h belongs to* $\mathrm{L}^2(0,T;\mathrm{H}^1(I))$.

Proof. Let $\phi \in \mathrm{C}_c^\infty(I)$. We have seen in the proof of Lemma 4.6 that the estimate

$$\|\langle \partial_x n^h, \phi\rangle\|_{\mathrm{L}^2(0,T)} \le C_T \left(h \sum_{i=-N^h+1}^{N^h} |\phi(ih)|^2 \right)^{1/2} = C_T\ \|\phi^h\|_{\mathrm{L}^2(I)}$$

holds. We also readily check that ϕ^h tends to ϕ in $\mathrm{L}^2(I)$. Hence, letting $h \to 0$ leads to

$$\|\langle \partial_x n, \phi\rangle\|_{\mathrm{L}^2(0,T)} \le C_T\ \|\phi\|_{\mathrm{L}^2(I)}.$$

By a density argument the estimate can be extended for any function $\phi \in \mathrm{L}^2(I)$. We conclude that $\partial_x n \in L^2((0,T)\times I)$. ∎

Convergence properties stronger than (5.1) will be necessary due to the nonlinear term. The idea is that the estimate in Lemma 4.4 is close to an $\mathrm{L}^2(0,T;\mathrm{H}^1(I))$ estimate on n^h. To this end we introduce the following $\mathbb{P}_1$ approximation: for $x \in (ih,(i+1)h)$, $i \in \{-N^h,\dots,N^h-1\}$, we set

$$\text{(5.3)} \qquad \begin{cases} m^h(t,x) = \dfrac{n^h_{i+1} - n^h_i}{h}\ (x - ih) + n^h_i, \\ G^h(t,x) = \dfrac{F^h_{i+1} - F^h_i}{h}\ (x - ih) + F^h_i. \end{cases}$$

Then, the sequences (m^h, G^h) are close to the original quantities (n^h, F^h) and enjoy better compactness properties as shown in the following lemma.

LEMMA 5.5. *The following estimates are verified:*

$$\begin{cases} \|n^h - m^h\|_{\mathrm{L}^2((0,T)\times I)} \le C_T\ h, \\ \|F^h - G^h\|_{\mathrm{L}^\infty((0,T)\times I)} \le C_T\ \sqrt{h}. \end{cases}$$

Furthermore, $\left(m^h\right)_{h>0}$ is relatively compact in $\mathrm{L}^2(0,T;\mathrm{C}^0(\overline{I}))$ and $\left(G^h\right)_{h>0}$ is relatively compact in $\mathrm{C}^0([0,T]\times\overline{I})$.

Proof. By taking into account the definition of the $\mathbb{P}_1$ approximations, we have

$$m^h(t,x) - n^h(t,x) = \frac{n^h_{i+1} - n^h_i}{h}\ (x - ih)$$

in the interval $(ih,(i+1)h)$, $i\in\{-N^h,\dots,N^h-1\}$. Hence, by using Lemma 4.4 we get

$$\begin{aligned}\|m^h-n^h\|^2_{\mathrm{L}^2((0,T)\times I)} &= \int_0^T\sum_{i=-N^h}^{N^h-1}\Big|\frac{n^h_{i+1}-n^h_i}{h}\Big|^2\int_{ih}^{(i+1)h}(x-ih)^2\,dx\,ds\\ &= \frac{h^2}{3}\int_0^T\sum_{i=-N^h}^{N^h-1}\frac{|n^h_{i+1}-n^h_i|^2}{h}\,ds\le C_T\ h^2\,.\end{aligned}$$

On the other hand, (3.2) yields

$$\begin{aligned}|G^h(t,x)-F^h(t,x)| &= \left|\frac{F^h_{i+1}-F^h_i}{h}\ (x-ih)\right|\\ &= |n^h_{i+1}-N_D|\ (x-ih)\le |n^h_{i+1}-N_D|\ h\end{aligned}$$

for $x\in(ih,(i+1)h)$, $i\in\{-N^h,\dots,N^h-1\}$. Therefore, Lemma 4.4 allows us to control this quantity as follows:

$$\begin{aligned}|G^h(t,x)-F^h(t,x)| &\le \sqrt{h}\ \sqrt{h}\ (n^h_{i+1}+N_D)\\ &\le \sqrt{h}\left(\Big(h|n^h_{i+1}|^2\Big)^{1/2}+\sqrt{h}N_D\right)\\ &\le \sqrt{h}\left(\left(h\sum_{j=-N^h}^{N^h}|n^h_j|^2\right)^{1/2}+\sqrt{h}N_D\right)\le C_T\ \sqrt{h}.\end{aligned}$$

This proves the first part of the result.

Note that m^h and G^h are bounded in $\mathrm{L}^2(0,T;\mathrm{H}^1(I))$ and $\mathrm{L}^\infty(0,T;\mathrm{H}^1(I))$, respectively. Indeed, we have $\partial_x m^h=(n^h_{i+1}-n^h_i)/h$ on $(ih,(i+1)h)$, and the bound for $\partial_x m^h$ in L^2 follows directly from Lemma 4.4. For the approximate electric field we have $\partial_x G^h=(F^h_{i+1}-F^h_i)/h=n^h_{i+1}-N_D$, so that

$$\begin{aligned}\|\partial_x G^h\|^2_{\mathrm{L}^2(I)} &= \sum_{i=-N^h}^{N^h}|n^h_{i+1}-N_D|^2\int_{ih}^{(i+1)h}dx\le 2\sum_{i=-N^h}^{N^h}\Big(|n^h_{i+1}|^2+N_D^2\Big)\ h\\ &\le 2\left(h\sum_{i=-N^h}^{N^h}|n^h_{i+1}|^2+(2X+h)N_D^2\right)\le C_T.\end{aligned}$$

Hence, to justify the compactness properties there remains to obtain some estimates on the time derivatives. We check that (see Appendix A)

$$\begin{aligned}&\partial_t(G^h-F^h)\quad\text{is bounded in } \mathrm{L}^2(0,T;\mathcal{M}(I)),\\ &\partial_t(m^h-n^h)\quad\text{is bounded in } \mathrm{L}^2(0,T;\mathcal{M}(I))+\mathrm{L}^2(0,T;\mathrm{W}^{-1,1}(I)).\end{aligned}\tag{5.4}$$

Then, combining this information with Lemma 4.7, we deduce the asserted compactness by application of Proposition 5.1. ∎

As a consequence of the compactness property, and by identifying limits, we can assure that

$$\begin{cases}G^h\to F & \text{uniformly on } [0,T]\times\overline{I},\\ m^h\to n & \text{strongly in } \mathrm{L}^2(0,T;C^0(\overline{I})),\\ \partial_x m^h\rightharpoonup\partial_x n & \text{weakly in } \mathrm{L}^2((0,T)\times I).\end{cases}\tag{5.5}$$

Since G^h is $\sqrt{h}$-close to F^h in the L^∞-norm, we can improve the convergence in (5.1). Actually, we have

$$F^h \to F \text{ uniformly on } [0,T]\times\overline{I}. \tag{5.6}$$

Notice also in (5.5) that the traces are well defined and the following convergences

$$\begin{cases} m^h(\pm X) = n^h_{\pm N^h} \to n(\pm X) = n_\pm & \text{strongly in } \mathrm{L}^2(0,T),\\ G^h(\pm X) = F^h_{\pm N^h} \to F(\pm X) = F_\pm & \text{strongly in } \mathrm{L}^2(0,T), \end{cases}$$

hold. In particular, the traces of n at $\pm X$ can be identified with the limits $n_\pm$, respectively, which were defined in (5.2).

In order to pass to the limit in the equation, we write a discrete weak formulation. Let $\phi \in \mathrm{C}^\infty(\overline{I})$. We denote $\phi^h_i = \phi(ih)$, and ϕ^h stands for the associated piecewise constant approximation. Then, we get

(5.7)
$$\begin{aligned} h\sum_{i=-N^h}^{N^h} \frac{d}{dt} n^h_i\ \phi^h_i &= \sum_{i=-N^h}^{N^h} \left(J^h_{i-1\to i} - J^h_{i\to i+1}\right)\ \phi^h_i \\ &= \sum_{i=-N^h}^{N^h-1} J^h_{i\to i+1}\ (\phi^h_{i+1}-\phi^h_i) - J^h_{N^h\to N^h+1}\phi^h_{N^h} + J^h_{-N^h-1\to -N^h}\phi^h_{-N^h} \\ &= \sum_{i=-N^h}^{N^h-1} v(F^h_i) n^h_i\ (\phi^h_{i+1}-\phi^h_i) - \sum_{i=-N^h}^{N^h-1} D(F^h_i)\frac{1}{h}\ (n^h_{i+1}-n^h_i)\ (\phi^h_{i+1}-\phi^h_i) \\ &\quad -W^{(f)}(F^h_{N^h}) n^h_{N^h}\phi^h_{N^h} + (j^{(e)}(F_-) - W^{(b)}(F_-) n^h_{-N^h})\phi^h_{-N^h}. \end{aligned}$$

Let us rewrite the discrete sums as integrals as follows:

$$\begin{aligned} &\frac{d}{dt}\int_{-X}^{X} n^h\ \phi^h\,dx + h\frac{d}{dt} n^h_{N^h}\ \phi(X) \\ &\quad = \int_{-X}^{X} v(F^h) n^h\ \nabla^h\phi\,dx - \int_{-X}^{X} D(F^h)\partial_x m^h\ \nabla^h\phi\,dx \\ &\qquad -W^{(f)}(F^h_{N^h}) n^h_{N^h}\phi(X) + (j^{(e)}(F_-) - W^{(b)}(F_-) n^h_{-N^h})\phi(-X), \end{aligned} \tag{5.8}$$

following the notation $\nabla^h\phi(x) = (\phi^h_{i+1}-\phi^h_i)/h$, for $x \in (ih,(i+1)h)$. We can now pass to the limit $h\to 0$.

We check that $\phi^h \to \phi$ and $\nabla^h\phi \to \phi'$ uniformly on $\overline{I}$. Let us pass to the limit in each term of (5.8). Taking into account that $n^h \to n$ in $\mathrm{C}^0([0,T];\mathrm{L}^2(I)-weak)$, we have $\int_{-X}^X n^h\phi^h\,dx \to \int_{-X}^X n\phi\,dx$ in $\mathrm{C}^0([0,T])$. Since $n^h_{N^h}$ is bounded in $\mathrm{L}^2(0,T)$, the second term in the left-hand side of (5.8) vanishes as $h\to 0$ in $\mathcal{D}'(0,T)$. Next, by using (5.6), $v(F^h)\nabla^h\phi \to v(F)\phi'$ and $D(F^h)\nabla^h\phi \to D(F)\phi'$ uniformly on $[0,T]\times\overline{I}$. To do that we combine the strong convergence $n^h \to n$ and the weak convergence $\partial_x m^h \to \partial_x n$ in $\mathrm{L}^2((0,T)\times I)$ so that the integrals in the right-hand side of (5.8) tend to

$$\int_X^X v(F) n\ \phi',dx - \int_X^X D(F)\partial_x n\ \phi'\,dx$$

as $h\to 0$ in $\mathcal{D}'(0,T)$. Finally, for the boundary terms we combine the convergence properties in (5.2) to find as the limit as $h\to 0$ the expression

$$-W^{(f)}(F) n\phi(X) + \left(j^{(e)}(F) - W^{(b)}(F) n\right)\phi(-X)\,.$$

Therefore, letting $h \to 0$ in (5.8), we have

$$\frac{d}{dt}\int_X^X n\ \phi, dx = \int_X^X v(F)n\ \phi', dx - \int_X^X D(F)\partial_x n\phi'\, dx + W^{(f)}(F)n\phi(X) + \left(j^{(e)}(F) - W^{(b)}(F)n\right)\phi(-X)$$

in $\mathcal{D}'(0,T)$. This ends the proof of Theorem 3.1.

Remark 5.6. The proof adapts readily if instead of assuming a constant doping density N_D, we deal with a sequence $\{N^h_{D,i},\ i \in \{-N^h, \dots, N^h\}\}$ verifying

$$h \sum_{i=-N^h}^{N^h} |N^h_{D,i}|^2 < \infty.$$

Accordingly, we obtain in the continuous limit a (possibly nonconstant) $\mathrm{L}^2(-X,X)$ doping density.

6. The bias constraint. In this section we reconsider the bias condition (2.8) as an alternative to the prescription of the emitter electric field (2.7). The arguments are exactly those of the previous section and we point out only the main differences in the proof. In rescaled form the condition is

$$h \sum_{i=-N^h}^{N^h} F^h_i = V, \tag{6.1}$$

which is added to the system (3.1), (3.2). This scaling means that the ratio $\frac{\mathcal{LF}}{\mathcal{V}}$ has order 1, $\mathcal{V}$ being a characteristic value for the total voltage. Of course, the L^1 estimate in Lemma 4.1 still holds, provided that $j^{(e)}$ is a bounded function. Then, the key point in the previous analysis is to establish a uniform estimate (with respect to h) on the electric field $F^h_{-N^h-1}$.

LEMMA 6.1. *The quantity $F^h_{-N^h-1}$ is bounded in $\mathrm{L}^\infty((0,T))$.*

Proof. Let us sum the relations (3.3). We get

$$\begin{aligned} h \sum_{i=-N^h}^{N^h} F^h_i &= V = h \sum_{i=-N^h}^{N^h} \left(F^h_{-N^h-1} + h \sum_{j=-N^h}^{i} (n^h_j - N_D) \right) \\ &= (2N^h+1)h\ F^h_{-N^h-1} + h^2 \sum_{j=-N^h}^{N^h} \left((n^h_j - N_D) \sum_{i=j}^{N^h} 1 \right) \\ &= (2N^h+1)h\ F^h_{-N^h-1} + h^2 \sum_{j=-N^h}^{N^h} (n^h_j - N_D)(N^h - j + 1). \end{aligned}$$

Consequently, the electric field at the emitter is given by

$$F^h_{-N^h-1} = \frac{V}{(2N^h+1)h} - \frac{h}{2N^h+1} \sum_{j=-N^h}^{N^h} (n^h_j - N_D)(N^h - j + 1). \tag{6.2}$$

It follows that

$$\begin{aligned}
|F^h_{-N^h-1}| &\leq \frac{|V|}{(2N^h+1)h} + \frac{h^2}{(2N^h+1)h}\sum_{j=-N^h}^{N^h} |n^h_j - N_D|\,|N^h - j + 1| \\
&\leq \frac{|V|}{2X} + \frac{h}{(2N^h+1)h}\left(h\sum_{j=-N^h}^{N^h} n^h_j + (2N^h+1)hN_D\right)(2N^h+1) \\
&\leq \frac{|V|}{2X} + h\sum_{j=-N^h}^{N^h} n^h_j + (2X+h)N_D.
\end{aligned}$$

This leads to the estimate of $F^h_{-N^h-1}$ in $\mathrm{L}^\infty((0,T))$. □

Once we have this estimate, we can justify the bounds in Lemmas 4.2 and 4.4. We also need some control on the time derivative of $F^h_{-N^h-1}$.

LEMMA 6.2. *The quantity $F^h_{-N^h-1}$ is bounded in $\mathrm{H}^1((0,T))$.*

Proof. Differentiating (6.2), we find

$$\begin{aligned}
\frac{d}{dt}F^h_{-N^h-1} &= \frac{h}{2N^h+1}\sum_{i=-N^h}^{N^h}\left(\sum_{j=-N^h}^{i}\frac{d}{dt}n^h_j\right) \\
&= \frac{1}{2N^h+1}\sum_{i=-N^h}^{N^h}\left(\sum_{j=-N^h}^{i}(J^h_{j-1\to j} - J^h_{j\to j+1})\right) \\
&= \frac{1}{2N^h+1}\sum_{i=-N^h}^{N^h}(J^h_{-N^h-1\to -N^h} - J^h_{i\to i+1}) \\
&= J^h_{-N^h-1\to -N^h} - \frac{1}{2N^h+1}\,J_{N^h\to N^h+1} \\
&\quad + \frac{1}{2N^h+1}\sum_{i=-N^h}^{N^h-1}\left(v(F^h_i)n^h_i - D(F^h_i)\frac{n^h_{i+1}-n^h_i}{h}\right).
\end{aligned}$$

Using the bounds of Lemmas 6.1 and 4.2, we can bound $v(F^h_i)$, $D(F^h_i)$, $j^{(e)}(F^h_{-N^h-1})$, $W^{(b)}(F^h_{-N^h-1})$, and $W^{(f)}(F^h_{N^h})$ by some constant $0 < M < \infty$. Hence, we deduce that

$$\begin{aligned}
\left|\frac{d}{dt}F^h_{-N^h-1}\right| &\leq M(1 + n^h_{-N^h} + n^h_{N^h}) \\
&\quad + \frac{M}{(2N^h+1)h}\left(h\sum_{i=-N^h}^{N^h} n^h_i + \sum_{i=-N^h}^{N^h}|n^h_{i+1} - n^h_i|\right) \\
&\leq M(1 + n^h_{-N^h} + n^h_{N^h}) + \frac{M}{2X}\,h\sum_{i=-N^h}^{N^h} n^h_i \\
&\quad + \frac{M}{\sqrt{2X}}\left(\sum_{i=-N^h}^{N^h}\frac{|n^h_{i+1} - n^h_i|^2}{h}\right)^{1/2}.
\end{aligned}$$

We conclude by applying the estimates of Lemma 4.4. □

By using these estimates, we can reproduce mutatis mutandis the arguments of the previous section. We conclude with the following result.

THEOREM 6.3. *Assume that $j^{(e)}$ is a bounded function. Then, the conclusions of Theorem* 3.1 *are still valid by replacing the condition* (2.7) *by* (6.1). *Accordingly, in the limit problem the electric field satisfies the Poisson equation $\partial_x F = n - N_D$ coupled to the constraint $\int_{-X}^{X} F\,dx = V$.*

Appendix A. Proof of (5.4). We write $m^h = \nu^h + n^h$, $G^h = \Phi^h + F^h$. Recall that ν^h, Φ^h are defined on $(0,T)\times(ih,(i+1)h)$, $i \in \{-N^h, \dots, N^h - 1\}$, by

$$\nu^h(t,x) = \frac{1}{h}\,(n^h_{i+1} - n^h_i), \qquad \Phi^h(t,x) = \frac{1}{h}\,(F^h_{i+1} - F^h_i) = n^h_{i+1} - N_D\,,$$

where we have used (3.2) in the second relation. As in the proof of Lemma 4.7, we consider a test function $\phi \in \mathrm{C}^\infty_0(I)$ and set $\Gamma^h_i = \int_{ih}^{(i+1)h}(x-ih)\phi(x)\,dx$, which verifies $|\Gamma^h_i| \le \|\phi\|_{\mathrm{L}^\infty(I)} h^2/2$. We have

$$\begin{aligned}
\langle \partial_t \Phi^h, \phi\rangle &= \sum_{i=-N^h}^{N^h-1} \frac{dn^h_{i+1}}{dt} \int_{ih} (i+1)h(x-ih)\phi(x)\,dx \\
&= \sum_{i=-N^h}^{N^h-1} \frac{1}{h}\,(J^h_{i\to i+1} - J^h_{i+1\to i+2})\,\Gamma^h_i \\
&= \sum_{i=-N^h}^{N^h-1} J^h_{i\to i+1}\,\frac{1}{h}\,(\Gamma^h_i - \Gamma^h_{i-1}) - \frac{1}{h} J^h_{N^h\to N^h+1}\Gamma^h_{N^h-1}\,.
\end{aligned}$$

We can bound this expression as follows:

$$\begin{aligned}
|\langle \partial_t \Phi^h, \phi\rangle| &\le \|\phi\|_{\mathrm{L}^\infty(I)} h \left(\sum_{i=-N^h}^{N^h-1} \left| v(F^h_i) n^h_i + \frac{1}{h} D(F^h_i)(n^h_{i+1} - n^h_i) \right| \right) \\
&\quad + \|\phi\|_{\mathrm{L}^\infty(I)} h\,|W^{(f)}(F^h_{N^h}) n^h_{N^h}| \\
&\le \|\phi\|_{\mathrm{L}^\infty(I)}\; M \left(h \sum_{i=-N^h}^{N^h-1} n^h_i + \left(\sum_{i=-N^h}^{N^h-1} \frac{|n^h_{i+1} - n^h_i|^2}{h} \right)^{1/2} \sqrt{2X} + n^h_{N^h} \right).
\end{aligned}$$

Thus, from Lemma 4.4 we deduce that $\partial_t \Phi^h$ is bounded in $\mathrm{L}^2(0,T;\mathcal{M}(I))$.

We proceed with ν^h in a similar way. Indeed, we can write

$$\begin{aligned}
\langle \partial_t \nu^h, \phi\rangle &= \frac{1}{h} \sum_{i=-N^h}^{N^h-1} \left(\frac{dn^h_{i+1}}{dt} - \frac{dn^h_i}{dt} \right) \int_{ih} (i+1)h(x-ih)\phi(x)\,dx \\
&= \frac{1}{h^2} \sum_{i=-N^h}^{N^h-1} (-J^h_{i+1\to i+2} + 2J^h_{i\to i+1} - J^h_{i-1\to i})\,\Gamma^h_i \\
&= \frac{1}{h^2} \sum_{i=-N^h}^{N^h-1} J^h_{i\to i+1}\,(-\Gamma^h_{i+1} + 2\Gamma^h_i - \Gamma^h_{i-1}) \\
&\quad - \frac{1}{h^2} J^h_{N^h\to N^h+1}\Gamma^h_{N^h-1} - \frac{1}{h^2} J^h_{-N^h-1\to -N^h}\Gamma^h_{-N^h}.
\end{aligned} \tag{A.1}$$

The boundary terms in (A.1) are bounded by

$$M(1+n^h_{-N^h}+n^h_{N^h})\|\phi\|_{\mathrm{L}^\infty(I)},$$

which belongs to a bounded set of $\mathrm{L}^2(0,T)$. Next, we have the bound

$$\frac{1}{h^2}|-\Gamma^h_{i+1}+2\Gamma^h_i-\Gamma^h_{i-1}|\le C\|\phi'\|_{\mathrm{L}^\infty(I)}\ h.$$

Therefore, the sum in the right-hand side of (A.1) can be estimated by

$$\begin{aligned}C\|\phi'\|_{\mathrm{L}^\infty(I)}\ h\sum_{i=-N^h}^{N^h-1}|J^h_{i\to i+1}| \ &\le\ CM\|\phi'\|_{\mathrm{L}^\infty(I)}\left(h\sum_{i=-N^h}^{N^h-1}n^h_i\right.\\ &\quad\left.+\left(\sum_{i=-N^h}^{N^h-1}\frac{|n^h_{i+1}-n^h_i|^2}{h}\right)^{1/2}\sqrt{2X}\right),\end{aligned}$$

as we did in the previous proof for Φ^h. We conclude that $\partial_t\nu^h$ is bounded in $\mathrm{L}^2(0,T;\mathcal{M}(I))+\mathrm{L}^2(0,T;\mathrm{W}^{-1,1}(I))$. This ends the proof of (5.4). ∎

Appendix B. Uniqueness for the limit problem. In this section, we show the uniqueness of the solution of (3.4). Let us consider two solutions (n_1,F_1) and (n_2,F_2) of (3.4) with $n_i\in\mathrm{C}^0([0,T];\mathrm{L}^2(I))\cap\mathrm{L}^2(0,T;\mathrm{H}^1(I))$. For the difference, we have

$$\partial_t(n_1-n_2)+\partial_x J(F_1,n_1-n_2)+\partial_x\Big((v(F_1)-v(F_2))n_2-(D(F_1)-D(F_2))\partial_x n_2\Big)=0,$$

where $J(F,n)=v(F)n-D(F)\partial_x n$. The boundary conditions read

$$\begin{cases}J(F_1,n_1-n_2)(X)=W^{(f)}(F_1)(n_1-n_2)+(W^{(f)}(F_1)-W^{(f)}(F_2))n_2,\\ J(F_1,n_1-n_2)(X)=j^{(e)}(F_1)-j^{(e)}(F_2)-W^{(b)}(F_1)(n_1-n_2)\\ \qquad\qquad\qquad\qquad -(W^{(b)}(F_1)-W^{(b)}(F_2))n_2.\end{cases}$$

Thus, we are left with only the task of evaluating

$$\begin{aligned}&\frac{d}{dt}\int_{-X}^X\frac{|n_1-n_2|^2}{2}\,dx+\int_{-X}^X D(F_1)|\partial_x(n_1-n_2)|^2\,dx\\ &=\int_{-X}^X v(F_1)(n_1-n_2)\partial_x(n_1-n_2)\,dx+\int_{-X}^X(v(F_1)-v(F_2))n_2\partial_x(n_1-n_2)\,dx\\ &\quad-\int_{-X}^X(D(F_1)-D(F_2))\partial_x n_2\partial_x(n_1-n_2)\,dx\\ &\quad+J(F_1,n_1-n_2)(n_1-n_2)(-X)-J(F_1,n_1-n_2)(n_1-n_2)(X).\end{aligned}\tag{B.1}$$

Denote by A,B,C,D, and E the five terms in the right-hand side of (B.1). Recall that F_i belongs to L^∞, so that the coefficients are lying in a bounded set. Also denote by Λ a Lipschitz constant for the functions $v,D,j^{(e)}$ and $W^{(b,f)}$ in the range of values of F_1 and F_2. Let $\nu>0$ be a parameter to be specified later on. By using the Cauchy–Schwarz and Young inequalities, we can estimate

$$|A|\le C_\nu\int_{-X}^X|n_1-n_2|^2\,dx+\nu\int_{-X}^X|\partial_x(n_1-n_2)|^2\,dx.$$

Next, we have

$$\begin{aligned} |B| &\leq \Lambda \|F_1 - F_2\|_{\mathrm{L}^\infty(I)} \int_{-X}^{X} |n_2| \, |\partial_x (n_1 - n_2)| \, dx \\ &\leq C_\nu \Lambda^2 \int_{-X}^{X} |n_2|^2 \, dx \, \|F_1 - F_2\|^2_{\mathrm{L}^\infty(I)} + \nu \int_{-X}^{X} |\partial_x (n_1 - n_2)|^2 \, dx. \end{aligned}$$

The Poisson equations yield to

$$(F_1 - F_2)(t,x) = F_{-,1} - F_{-,2} + \int_{-X}^{x} (n_1 - n_2)(t,y) \, dy,$$

which provides the bound

$$\|F_1 - F_2\|^2_{\mathrm{L}^\infty(I)} \leq 2|F_{-,1} - F_{-,2}|^2 + 4X \int_{-X}^{X} |n_1 - n_2|^2 \, dx.$$

Hence, we get (changing the value of C_ν)

$$|B| \leq C_\nu \int_{-X}^{X} |n_2|^2 \, dx \left(|F_{-,1} - F_{-,2}|^2 + \int_{-X}^{X} |n_1 - n_2|^2 \, dx \right) + \nu \int_{-X}^{X} |\partial_x (n_1 - n_2)|^2 \, dx.$$

A similar reasoning for C leads to

$$|C| \leq C_\nu \int_{-X}^{X} |\partial_x n_2|^2 dx \left(|F_{-,1} - F_{-,2}|^2 + \int_{-X}^{X} |n_1 - n_2|^2 dx \right) + \nu \int_{-X}^{X} |\partial_x (n_1 - n_2)|^2 dx.$$

For the boundary terms, we get rid of the terms $-W^{(b,f)}(F_1)|n_1 - n_2|^2$ which are nonnegative and get

$$D + E \leq \Lambda \Big((1 + n_2) \, |F_1 - F_2| \, |n_1 - n_2|(-X) + n_2 \, |F_1 - F_2| \, |n_1 - n_2|(+X) \Big).$$

Then, we use the Sobolev embedding to control the traces of $n_1 - n_2$ with the H^1-norm. Finally, we obtain

$$\begin{aligned} D + E &\leq C_\nu (1 + |n_2(-X)|^2 + |n_2(X)|^2) \left(|F_{-,1} - F_{-,2}|^2 + \int_{-X}^{X} |n_1 - n_2|^2 \right) \\ &\quad + \nu \left(\int_{-X}^{X} |n_1 - n_2|^2 \, dx + \int_{-X}^{X} |\partial_x (n_1 - n_2)|^2 \, dx \right). \end{aligned}$$

Having disposed of these preliminaries, we recall that $D(F_1)$ is bounded from below by some $\delta > 0$. Then, we put all the pieces together and choose $\nu = \nu(\delta)$ appropriately so that we finally find

$$\begin{aligned} &\frac{d}{dt} \int_{-X}^{X} |n_1 - n_2|^2 \, dx + \frac{\delta}{2} \int_{-X}^{X} |\partial_x (n_1 - n_2)|^2 \, dx \\ &\qquad \leq f(t) \int_{-X}^{X} |n_1 - n_2|^2 \, dx + g(t) |F_{-,1} - F_{-,2}|^2, \end{aligned}$$

where the nonnegative functions $f, g \in \mathrm{L}^1(0,T)$ depend on Λ, δ and $\int_{-X}^{X}(n_2^2 + |\partial_x n_2|^2)\,dx$. The Gronwall lemma provides the inequality

$$\int_{-X}^{X}|n_1 - n_2|^2(t,x)\,dx \leq e^{\int_0^t f(s)\,ds}\left(\int_{-X}^{X}|n_1 - n_2|^2(0,x)\,dx + \int_0^t g(s)|F_{-,1} - F_{-,2}|^2(s)\,ds\right).$$

This proves the continuity of the solution with respect to the data and, consequently, the uniqueness of the solution. We skip the adaptation of the proof to the bias condition.

REFERENCES

[1] R. Aguado, G. Platero, M. Moscoso, and L. L. Bonilla, *Microscopic model for sequential tunneling in semiconductor multiple quantum wells*, Phys. Rev. B, 55 (1997), R 16053–16056.

[2] J. P. Aubin, *Un théorème de compacité*, C. R. Acad. Sci. Paris Sér. I Math., 256 (1963), pp. 5042–5044.

[3] J. Bardeen, *Tunneling from a many-particle point of view*, Phys. Rev. Lett., 6 (1961), pp. 57–59.

[4] L. L. Bonilla, G. Platero, and D. Sánchez, *Microscopic derivation of transport coefficient and boundary conditions in discrete drift-diffusion models of weakly coupled superlattices*, Phys. Rev. B, 62 (2000), pp. 2786–2796.

[5] L. L. Bonilla, O. Sánchez, and J. Soler, *Nonlinear stochastic discrete drift-diffusion theory of charge fluctuations and domain relocation times in semiconductor superlattices*, Phys. Rev. B, 65 (2002), pp. 195308/1–8.

[6] L. L. Bonilla, *Theory of nonlinear charge transport, wave propagation and self-oscillations in semiconductor superlattices*, J. Phys. Condens. Matter, 14 (2002), R341–R381.

[7] J.-F. Collet, T. Goudon, F. Poupaud, and A. Vasseur, *The Beker–Döring system and its Lifshitz–Slyozov limit*, SIAM J. Appl. Math., 62 (2002), pp. 1488–1500.

[8] H. Gajewski and K. Gröger, *On the basic equations for carrier transport in semiconductors*, J. Math. Anal. Appl., 113 (1986), pp. 12-35.

[9] F. Golse, *From kinetic to macroscopic models in kinetic equations and asymptotic theory*, in Kinetic Equations and Asymptotic Theory, Ser. Appl. Math. 4, B. Perthame and L. Desvillettes, eds., Gauthiers-Villars, Paris, 2000, pp. 41–126.

[10] J. W. Jerome, *The approximation problem for drift-diffusion systems*, SIAM Rev., 37 (1995), pp. 552–572.

[11] J. Liang, *On a nonlinear integrodifferential drift-diffusion semiconductor model*, SIAM J. Math. Anal., 25 (1994), pp. 1375–1392.

[12] M. Rogozia, S. W. Teitsworth, H. T. Grahn, and K. H. Ploog, *Statistics of the domain-boundary relocation time in semiconductor superlattices*, Phys. Rev. B, 64 (2001), 041308(R).

[13] W. Rudin, *Real and Complex Analysis*, 3rd ed., McGraw–Hill, New York, 1987.

[14] J. Simon, *Compact sets in the space* $\mathrm{L}^p(0,T;B)$, Ann. Mat. Pura Appl. 4, 146 (1987), pp. 65–96.

[15] A. Wacker, *Semiconductor superlattices: A model system for nonlinear transport*, Phys. Rep. 357 (2002), pp. 1–111.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1550–1587

WELL-POSEDNESS OF TWO NONRIGID MULTIMODAL IMAGE REGISTRATION METHODS*

OLIVIER FAUGERAS† AND GERARDO HERMOSILLO†

Abstract. Image registration methods may be designed as solutions of minimization problems on a set of geometric deformations. In the nonrigid case, solving these problems often means computing the steady state of a system of evolution equations involving the "gradient" of the error criterion, defined from the Euler–Lagrange equations of the corresponding minimization problem. The well-posedness of the registration method requires showing the existence of a solution to the minimization problem, as well as that of a stable solution of the evolution equations derived from it. We provide such proofs in the case where the error criterion is derived from two different statistical similarity measures: global mutual information and local cross-covariance. We also describe our numerical implementation for solving the corresponding evolution equations and show examples of registrations of real 2D and 3D images achieved with these algorithms. The proofs are quite general and can be applied to most of the known nonrigid image registration methods.

Key words. multimodal image matching, variational methods, image registration, mutual information, cross-covariance, Euler–Lagrange equations, initial-value problems, analytical semigroups of linear operators

AMS subject classifications. 34G20, 35B65, 35D10, 35K55, 35K90, 35Q80, 47D03, 47D06, 47H06, 47H07, 47J35, 47N60

DOI. 10.1137/S0036139903424904

1. Introduction.

The problem of estimating the geometric deformation between two images has a long history. Most of the existing methods have been developed for images acquired with the same sensor or at least the same modality (e.g., in the visible spectrum). They often rely on the minimization of an error criterion which takes into account two sources of a priori knowledge: (a) the properties of the images' intensities that characterize their similarity and (b) the constraints on the possible geometric deformations.

For the first point, the basic idea is that the intensities at corresponding points should be "similar," i.e., equal as in the case of the optical flow problem [28]. This is, of course, too strict a requirement in many cases, and it has been relaxed to an average intensity similarity in the neighborhoods of the corresponding points, as in the case of local image differences [29] or more general block matching strategies [51, 41]. One more step and one can characterize similarity as a large score of a nonlocal similarity measure such as the cross-correlation [20, 21, 11, 39], the correlation ratio [49], or mutual information [55, 58, 56, 32], among several others [59, 26, 45, 31].

As for the second point, the possible geometric deformations, a first idea is to restrict the search to sets of low-dimensional parametric transformations (e.g., Euclidean, affine, quadratic, or spline-interpolation between a set of control points [36, 50]). Another example of a constrained deformation is the stereo case, in which knowledge of the fundamental matrix allows one to restrict the search for the matching point along the epipolar line [1, 60]. If the deformation is not defined parametrically,

*Received by the editors March 18, 2003; accepted for publication (in revised form) November 10, 2003; published electronically June 22, 2004. This work was partially supported by NSF grant DMS-9972228, EC grant Mapawamo QLG3-CT-2000-30161, INRIA ARCs IRMf and MC2, and the Mexican National Council for Science and Technology, Conacyt.

http://www.siam.org/journals/siap/64-5/42490.html

†INRIA Sophia Antipolis, 2004 route des Lucioles, BP 93 06902, Sophia-Antipolis Cedex, France (faugeras@sophia.inria.fr, ghermosi@sophia.inria.fr).

the constraint may consist of requiring some smoothness of the displacement field, possibly preserving discontinuities [53, 1, 47, 2, 34, 35, 4, 3, 18]. Concerning this regularization, we can distinguish the approaches based on explicit smoothing of the field, as in Thirion's deamons algorithm [53] (we refer to [44] for a variational interpretation of this algorithm), from those considering an additive term in the error criterion, yielding (possibly anisotropic) diffusion terms [5, 57]. For a comparison of these two approaches, we refer to the work of Cachier and Ayache [9, 10]. Approaches relying on a regularization functional induce a certain amount of competition between the similarity of the images and the smoothness of the deformation. Fluid methods do not impose this competition, although they still require a parameter that fixes the amount of desired smoothness or fluidness of the result [13, 54, 12].

Many of the previous methods are "differential" and are mostly valid for small displacements. Special techniques are required in order to recover large deformations. For instance, Alvarez, Weickert, and Sánchez [2] use a scale-space focusing strategy. Christensen, Miller, and Vannier [13] adopt a different approach. They look for a continuously invertible mapping which is obtained by the composition of small displacements. Each small displacement is calculated as the solution of an elliptic partial differential equation describing the nonlinear kinematics of fluid-elastic materials under deforming forces given by the matching term (in their case the image differences). Trouvé [54] has generalized this approach using Lie group ideas on sets of diffeomorphisms. Under a similar formalism, a very general framework which also allows for changes in the intensity values is proposed by Miller and Younes [37].

In the case of multimodal image registration (e.g., visible and infrared, anatomical and functional magnetic resonance images, etc.) the similarity in the intensities is "weak" and one has to rely on statistical definitions. Such definitions have been widely used in the case of low-dimensional parametric transformations. Mutual information was introduced by Viola and colleagues [55, 58, 56] and independently by Maes et al. [32]. The correlation ratio was first proposed as a similarity measure for image matching by Roche et al. [49]. Other statistical approaches rely on learning the joint distribution of intensities, as done, for instance, by Leventon and Grimson [31]. Extensions to more complex (nonrigid) transformations using statistical similarity measures include approaches relying on more complex parametric transformations [36, 50], block matching strategies [33, 24, 22], and parametric intensity corrections [48]. Some recent approaches rely on the computation of the gradient of the local cross correlation [11, 39].

Several papers have discussed the theoretical well-posedness of similar registration problems in the monomodal case as, for instance, [27, 14]. However, none of those treating the multimodal case has dealt significantly with the problem of the existence and uniqueness of a solution of the registration problem. Our work is an attempt to improve this situation. We consider the problem of multimodal image registration and use statistical considerations to define similarity. In order to be able to deal with strongly nonstationary variations in the intensity we extend the usual statistical framework to act locally instead of globally in the images. Since most of the geometric deformations that occur in practice are nonparametric, we model them as dense deformation fields that are constrained to belong to some reasonable functional spaces, e.g., Sobolev spaces. We then classically define the problem as a variational one and show that it is well posed.

In detail, we consider the problem of dense matching between two images using statistical dissimilarity criteria, which are well adapted to the case of multimodal im-

age data often encountered, e.g., in medical imaging. The minimization of the sum of the dissimilarity term and the regularization term defines, through the associated Euler–Lagrange equations, a set of coupled functional evolution equations. The conditions under which this type of evolution equation is well posed are known. We show that the matching functions that we obtain satisfy these conditions in the case where the error functional is derived from two different statistical similarity measures: global mutual information and local cross-covariance.

1.1. A generic registration problem. At a conceptual level, images are integrable real bounded functions defined on $\mathbb{R}^n$ (we restrict ourselves to $n = 2, 3$). These abstract images are not directly observable because of the physics of acquisition. What we call an *image* is the convolution of such a function with a C^∞ mollifier. We therefore view the *images* as belonging to the space of infinitely differentiable functions, $C^\infty(\mathbb{R}^n)$. They are bounded and Lipschitz continuous, as well as all their derivatives. These assumptions are not central to this article and can be greatly weakened. The reader will verify that the first image is required only to be C^2 while the second is required to be C^1 with a Lipschitz continuous first order derivative. This asymmetry is due to the fact that the two images do not play the same role in the registration problem. Weaker conditions may be possible but we have not investigated this specific problem. The range of the values of an image is the interval $[0, \mathcal{A}]$, $\mathcal{A} > 0$.

Let I_1 and I_2 be two images. Let $\mathbf{Id}$ be the identity mapping of $\mathbb{R}^n$ and $\mathbf{h} : \Omega \to \mathbb{R}^n$ a given vector field defined on a bounded and regular region of interest $\Omega \subset \mathbb{R}^n$. By regular we mean C^2. This technical assumption is required in the proofs of Lemma 2.3 and Proposition 2.4. Images are usually defined on a product of intervals. Ω is in practice chosen as a C^2 superset of this set (e.g., a disk). The image values are smoothly extended outside the product of intervals to reach the value of 0 on the boundary $\partial\Omega$ of Ω. The registration or matching problem may be defined as that of finding a vector field $\mathbf{h}^*$ minimizing an error criterion between I_1 and the warped image $I_2 \circ (\mathbf{Id} + \mathbf{h})$.

The search for this function is done within a set $\mathcal{F}$ of admissible functions such that it minimizes an energy functional $\mathcal{I} : \mathcal{F} \to \mathbb{R}$ of the form

$$\mathcal{I}(\mathbf{h}) = \mathcal{J}(\mathbf{h}) + \mathcal{R}(\mathbf{h}).$$

The term $\mathcal{J}(\mathbf{h})$ is designed to measure the "dissimilarity" between the reference image (I_1) and the $\mathbf{h}$-warped second image $(I_2(\mathbf{Id} + \mathbf{h}))$. The term $\mathcal{R}(\mathbf{h})$ is designed to penalize fast variations of the function $\mathbf{h}$. It is a regularization term introducing an a priori preference for smoothly varying functions.

Generally speaking, the set $\mathcal{F}$ is a dense linear subspace of a Hilbert space H, the scalar product of which is denoted by $(\cdot, \cdot)_H$. If $\mathcal{I}$ is sufficiently regular, its first variation[1] at $\mathbf{h} \in \mathcal{F}$ in the direction of $\mathbf{k} \in H$ is defined by (see, e.g., [15])

$$\delta_{\mathbf{k}}\mathcal{I}(\mathbf{h}) = \lim_{\varepsilon \to 0} \frac{\mathcal{I}(\mathbf{h} + \varepsilon\mathbf{k}) - \mathcal{I}(\mathbf{h})}{\varepsilon}. \tag{1.1}$$

If the mapping $\mathbf{k} \to \delta_{\mathbf{k}}\mathcal{I}(\mathbf{h})$ is linear and continuous, the Riesz representation theorem [19] guarantees the existence of a unique vector, denoted by $\nabla_H\mathcal{I}(\mathbf{h})$, and called the gradient of $\mathcal{I}$, which satisfies the equality

$$\delta_{\mathbf{k}}\mathcal{I}(\mathbf{h}) = (\nabla_H\mathcal{I}(\mathbf{h}), \mathbf{k})_H$$

[1] Also called the Gâteaux derivative.

for every $\mathbf{k} \in H$. The gradient depends on the choice of the scalar product $(\cdot,\cdot)_H$ though, a fact which explains our notation. If a minimizer $\mathbf{h}^*$ of $\mathcal{I}$ exists, then the set of equations $\delta_{\mathbf{k}}\mathcal{I}(\mathbf{h}^*) = 0$ must hold for every $\mathbf{k} \in H$, which is equivalent to $\nabla_H \mathcal{I}(\mathbf{h}^*) = 0$. These equations are called the Euler–Lagrange equations associated with the energy functional $\mathcal{I}$. They give necessary conditions for the existence of a minimizer, but they are not sufficient since they guarantee only the existence of a critical point of the functional $\mathcal{I}$. These critical points can be found in many ways, including methods for nonlinear equations. Rather than solving them directly, the search for a minimizer of $\mathcal{I}$ is done using a "gradient descent" strategy. Given an initial estimate $\mathbf{h}_0 \in \mathcal{F}$, a time-dependent differentiable function (also denoted by $\mathbf{h}$) from the interval $[0, +\infty[$ into H is computed as the solution of the following initial value problem:

$$\begin{cases} \dfrac{d\mathbf{h}}{dt} = -\Big(\nabla_H \mathcal{J}(\mathbf{h}) + \nabla_H \mathcal{R}(\mathbf{h})\Big), \\ \mathbf{h}(0)(\cdot) = \mathbf{h}_0(\cdot). \end{cases} \tag{1.2}$$

The asymptotic state (i.e., when $t \to \infty$) of $\mathbf{h}(t)$ is then chosen as the solution of the matching problem, provided that $\mathbf{h}(t) \in \mathcal{F}\ \forall t > 0$. We shall restrict ourselves to the case when $\nabla_H \mathcal{R}$ is an unbounded *linear* operator from its domain into H, whereas $\nabla_H \mathcal{J}$ may be a nonlinear function mapping H into H.

There are two (mostly theoretical) advantages in introducing the artificial time variable. The first is that, as shown in Theorem 1.2, we are able to prove the well-posedness of the initial value problem (1.2). The second is that, as shown in Proposition 2.5, we are able to prove that the asymptotic state of the solution of (1.2) is indeed a zero of the Euler–Lagrange equations and, in effect, a local minimum of $\mathcal{I}$. The third advantage is practical: our experiments (see section 6) have shown that when we discretize (1.2) we converge toward a local minimum at a reasonable speed.

1.2. Existence of a classical solution of (1.2). Equation (1.2) may be viewed as a first order ordinary differential equation with values in H. By borrowing tools from functional analysis and the theory of semigroups generated by unbounded linear operators on Hilbert spaces, we can prove the existence of a unique classical solution of (1.2) under fairly general hypotheses. We refer to the books of Brezis [7], Pazy [43], and Tanabe [52] for an in-depth study of these subjects. Because of our choice of the regularization term $\mathcal{R}$ (described in section 2), it turns out that the mapping $A : \mathbf{h} \to -\nabla_H \mathcal{R}(\mathbf{h})$ is linear from its domain, a subset $\mathcal{D}(A)$ of H, into H. Similarly, we denote by F the (nonlinear) mapping defined by $\mathbf{h} \to -\nabla_H \mathcal{J}$. F is called the *matching* function of the registration problem. The unknown of the initial-value problem (1.2) is an H-valued function $\mathbf{h} : [0, +\infty[\to H$ defined on $\mathbb{R}^+$. We now establish the properties required for A and F in order for (1.2), which is now written as a semilinear abstract initial value problem of the form

$$\begin{cases} \dfrac{d\mathbf{h}}{dt} - A\mathbf{h}(t) = F(\mathbf{h}(t)), \qquad t > 0, \\ \mathbf{h}(0) = \mathbf{h}_0 \in H, \end{cases} \tag{1.3}$$

to have a unique solution.

We first recall the following definition.

DEFINITION 1.1. *A function* $\mathbf{h} : [0, +\infty[\to H$ *is a classical solution of* (1.3) *if*

$$\mathbf{h} \in C([0, +\infty[; H) \cap C^1(]0, +\infty[; H) \cap C(]0, +\infty[; \mathcal{D}(A))$$

and (1.3) *is satisfied for* $t > 0$.

As a consequence of known results we have the following.

THEOREM 1.2. *If in* (1.3) *the linear operator* $-A$ *is strongly elliptic, invertible*[2] *and the nonlinear function* F *is bounded and Lipschitz continuous, then there exists a unique classical solution of* (1.3). *Moreover, the function* $t \to d\mathbf{h}/dt$ *from* $]0, +\infty[$ *into* H_α *(defined in the proof) is Hölder continuous.*

Proof. We use the notion of an analytical semigroup of operators on H defined, e.g., in [52, 43]. Theorem 3.6.1 in [52] or Theorem 2.7 in Chapter 7 of [43] show that if the operator $-A$ is strongly elliptic, then A generates an analytical semigroup $S_A(t)$, $t \geq 0$, on H. It follows from [43, Chapter 2, section 2.6], that $(-A)^\alpha$ can be defined for $0 < \alpha \leq 1$ and that $(-A)^\alpha$ is a closed linear invertible operator with domain $\mathcal{D}((-A)^\alpha)$ dense in H. It is invertible because $-A$ is. The closedness of $(-A)^\alpha$ implies that $\mathcal{D}((-A)^\alpha)$ endowed with the graph norm of $(-A)^\alpha$, i.e., the norm $|||\mathbf{h}||| = \|\mathbf{h}\|_H + \|(-A)^\alpha \mathbf{h}\|_H$, is a Banach space. Since $(-A)^\alpha$ is invertible, its graph norm is equivalent to the norm $\|\mathbf{h}\|_\alpha = \|(-A)^\alpha \mathbf{h}\|_H$. Thus, $\mathcal{D}((-A)^\alpha)$ equipped with the norm $\|\cdot\|_\alpha$ is a Banach space, which we denote by H_α. The space H_α is continuously embedded in H for all α's, which implies

$$\|\mathbf{h}\|_H \leq k_\alpha \|\mathbf{h}\|_{H_\alpha} \quad \forall \mathbf{h} \in H_\alpha,$$

for some $k_\alpha > 0$. Theorems 3.1 and 3.3 of [43] allow us to conclude. ∎

The remainder of the paper is divided into five additional sections. Section 2 discusses the regularization part of the initial-value problem (1.3). Section 3 is devoted to the definition of the two statistical dissimilarity measures we have considered and to the computations of the associated Euler–Lagrange equations. Section 4 contains the proofs of the Lipschitz continuity of the corresponding matching functions. Finally, sections 5 and 6 describe the implementation of the matching algorithms and present experimental results with real two- and three-dimensional (2D and 3D) images, respectively.

2. Regularization term. This section studies the regularization part of the initial-value problem (1.2), i.e., the term $\nabla_H \mathcal{R}(\mathbf{h})$. A one-parameter family of regularization operators is considered which encourages the preservation of edges of the displacement field along the edges of the reference image. In view of the results of the previous section, we choose concrete functional spaces $\mathcal{F}$ and H and specify the domain of the regularization operators. We then show that these operators satisfy the properties of A which are sufficient to assert the existence of a classical solution of (1.2) according to the main result of the previous section.

2.1. Function spaces and boundary conditions. We begin with a brief description of the functional spaces that will be appropriate for our purposes. In doing this, we will make reference to Sobolev spaces, denoted by $W^{k,p}(\Omega)$. We refer to the books of Evans [19] and Brezis [7] for formal definitions and in-depth studies of the properties of these functional spaces.

For the definition of $\nabla_H \mathcal{I}$, we use the Hilbert space

$$H = \mathbf{L}^2(\Omega) = \underbrace{L^2(\Omega) \times \cdots \times L^2(\Omega)}_{n \text{ terms}} = \left(W^{0,2}(\Omega)\right)^n.$$

[2]The invertibility of A is not required, as discussed in, e.g., [43, p. 195], but it makes the proofs simpler.

The regularization functionals that we consider are of the form

$$\mathcal{R}(\mathbf{h}) = \kappa \int_\Omega \varphi(D\mathbf{h}(\mathbf{x}))\, d\mathbf{x}, \tag{2.1}$$

where $D\mathbf{h}(\mathbf{x})$ is the Jacobian of $\mathbf{h}$ at $\mathbf{x}$, φ is a quadratic form of the elements of the matrix $D\mathbf{h}(\mathbf{x})$, and $\kappa > 0$. Therefore the set of admissible functions $\mathcal{F}$ will be contained in the space

$$\mathbf{H}^1(\Omega) = (W^{1,2}(\Omega))^n.$$

Additionally, the boundary conditions for $\mathbf{h}$ will be specified in $\mathcal{F}$. We consider Dirichlet conditions of the form $\mathbf{h} = 0$ almost everywhere on $\partial\Omega$ (in fact, because of the regularity of $\mathbf{h}$, proved in Proposition 2.4, this condition holds everywhere on $\partial\Omega$), and set

$$\mathcal{F} = \mathbf{H}_0^1(\Omega) = (W_0^{1,2}(\Omega))^n.$$

Because of the special form of $\mathcal{R}(\mathbf{h})$, the corresponding regularization operator is a second order differential one, and we therefore will need the space

$$\mathbf{H}^2(\Omega) = (W^{2,2}(\Omega))^n$$

for the definition of its domain.

2.2. Image-driven anisotropic diffusion. The family that we consider is obtained by defining φ in (2.1) by

$$\varphi(D\mathbf{h}) = \frac{1}{2}\mathrm{Tr}\big(D\mathbf{h}\ \mathbf{T}_{I_1}\ D\mathbf{h}^T\big), \tag{2.2}$$

where $\mathbf{T}_{I_1}$ is an $n \times n$ symmetric matrix defined at every point of Ω by the following expression:

$$\mathbf{T}_f = \frac{(\lambda + |\nabla f|^2)\mathbf{Id} - \nabla f \nabla f^T}{(n-1)|\nabla f|^2 + n\lambda} \quad \text{for} \quad f : \mathbb{R}^n \to \mathbb{R},\ \text{regular.}$$

This matrix is a regularized projector in the plane perpendicular to ∇f. It was first proposed by Nagel and Enkelmann [38] for computing optical flow while preserving the discontinuities of the deforming template. As pointed out by Alvarez, Weickert, and Sánchez [2], applying the smoothness constraint to the reference image (here I_1) instead of the deforming one (here I_2) allows us to avoid artifacts which appear when recovering large displacements. The matrix $\mathbf{T}_f$ has one eigenvector equal to ∇f, while the remaining eigenvectors span the plane perpendicular to ∇f. The eigenvalues λ_i of this matrix are all positive and verify $\sum_i \lambda_i = 1$, independently of ∇f.

It is straightforward to verify that the Euler–Lagrange equation corresponding to (2.1) in this case is

$$\mathbf{div}(D\varphi(D\mathbf{h})) = \begin{pmatrix} \mathrm{div}(\mathbf{T}_{I_1}\nabla h_1) \\ \vdots \\ \mathrm{div}(\mathbf{T}_{I_1}\nabla h_n) \end{pmatrix} = 0.$$

Thus, the regularization operator $\nabla_H \mathcal{R}(\mathbf{h})$ yields a linear diffusion term with $\mathbf{T}_{I_1}$ as diffusion tensor. In regions where $\nabla \mathbf{h}_i$ is small compared to the parameter λ in $\mathbf{T}_{I_1}$,

the diffusion tensor is almost isotropic and so is the regularization. At the edges of f, where $|\nabla I_1| \gg \lambda$, the diffusion takes place mainly along these edges. This operator is thus well suited for encouraging large variations of $\mathbf{h}$ along the edges of the reference image I_1.

We define the corresponding regularization operator as follows.

DEFINITION 2.1. *The linear operator $A : \mathcal{D}(A) \to H$ is defined as*

$$\begin{cases} \mathcal{D}(A) = \mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega), \\ \\ A\mathbf{h} = \begin{pmatrix} \operatorname{div}(\mathbf{T}_{I_1} \nabla h_1) \\ \vdots \\ \operatorname{div}(\mathbf{T}_{I_1} \nabla h_n) \end{pmatrix}. \end{cases}$$

We now check that $-A$ is strongly elliptic (or, in functional analytic language, is variational) by applying the standard variational approach [19].

PROPOSITION 2.2. *The operator $-A$ defines a bilinear form B on the space $\mathbf{H}_0^1(\Omega)$ which is continuous and coercive (strongly elliptic).*

Proof. The proof is quite standard, and we only sketch it here.

Because of the form of the operator A, it is sufficient to work on one of the coordinates, to consider the operator $a : \mathcal{D}(a) \to L^2(\Omega)$ defined by

$$a\, u = \operatorname{div}(\mathbf{T}_{I_1} \nabla u),$$

and to show that the operator $u \to -au$ defines a bilinear form b on the space $H_0^1(\Omega)$ which is continuous and coercive. Continuity is obtained from the fact that the coefficients of $\mathbf{T}_{I_1}$ are bounded by integrating by parts the expression of $b(u, v)$ and applying the Cauchy–Schwarz inequality. Coercivity is obtained from the fact that the eigenvalues of $\mathbf{T}_{I_1}$ are strictly positive and by using Poincaré's inequality. ∎

LEMMA 2.3. *The linear operator $-\kappa A$ is invertible for all $\kappa > 0$.*

Proof. It is sufficient to show that the equation $-\kappa A\mathbf{h} = \mathbf{f}$ has a unique solution for all $\mathbf{f} \in \mathbf{L}^2(\Omega)$. The proof of Proposition 2.2 shows that the bilinear form associated with the operator $-\kappa A$ is continuous and coercive in $\mathbf{H}^1(\Omega)$; hence the Lax–Milgram theorem tells us that the equation $-\kappa A\mathbf{h} = \mathbf{f}$ has a unique weak solution in $\mathbf{H}_0^1(\Omega)$ for all $\mathbf{f} \in \mathbf{L}^2(\Omega)$. Since Ω is regular (i.e., C^2), the weak solution is in $\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)$ and is a strong solution. ∎

2.3. Existence of a regular solution of (1.2). Now that the domain of the regularization operator is defined, we give a stronger result concerning the existence of a solution of (1.2): because the domain of A is defined in terms of Sobolev spaces, a classical solution of the abstract problem corresponds in fact to the notion of *weak* solution in PDE theory. The existence of a *regular* solution in this context may be shown, assuming regularity of the boundary $\partial\Omega$ of Ω. In effect we have the following result.

PROPOSITION 2.4. *The functions $(t, \mathbf{x}) \to \mathbf{h}(t, \mathbf{x})$ and $(t, \mathbf{x}) \to (\partial/\partial t)\mathbf{h}(t, \mathbf{x})$ are continuous on $]0, +\infty[\times\overline{\Omega}$, and for each $t > 0$ the function $\mathbf{x} \to \mathbf{h}(t, \mathbf{x})$ is in $\mathbf{C}^2(\Omega)$.*

Proof. It follows from a theorem due to Sobolev that $\mathbf{H}^2(\Omega) \subset \mathbf{C}(\overline{\Omega})$ for $n = 2, 3$; hence $\mathcal{D}(A) \subset \mathbf{H}^2(\Omega) \subset \mathbf{C}(\overline{\Omega})$ and, since $\mathbf{h}(t) \in \mathcal{D}(A)$ for $t > 0$, this proves the continuity of $\mathbf{h}$ on $]0, +\infty[\times\overline{\Omega}$. Next, because of Theorem 1.2, $t \to d\mathbf{h}/dt \in H_\alpha$ is Hölder continuous for $t > 0$. Moreover, a theorem due to Sobolev, which can be found, e.g., in Theorem 8.4.3 of [43], shows that if $\alpha > 3/4$ ($n = 3$) and $\alpha > 1/2$

$(n = 2)$, H_α is a set of Hölder continuous functions, and we have the continuity of $d\mathbf{h}/dt$ on $]0, +\infty[\times\overline{\Omega}$.

It remains to show that $\mathbf{h}(t,\cdot) \in \mathbf{C}^2(\Omega)$. First the function $\mathbf{x} \to F(\mathbf{h}(\mathbf{x}))$ from $\Omega \to \mathbb{R}^n$ is Hölder continuous if $\mathbf{h}$ is (this is proved in Propositions 4.12 and 4.22). Since $\mathbf{h}(t) \in \mathbf{H}^2(\Omega)$ for $t > 0$ and (Sobolev) $\mathbf{H}^2(\Omega) \subset \mathbf{C}^\gamma(\overline{\Omega})$ $(0 \le \gamma < 0.5$ for $n = 3$ and $0 \le \gamma < 1$ for $n = 2)$, $\mathbf{x} \to \mathbf{h}(t,\mathbf{x})$ is Hölder continuous for $t > 0$. Finally, since $(\partial/\partial t)\mathbf{h}(t,.)$ is Hölder continuous in Ω, it follows that $-A\mathbf{h} = F(\mathbf{h}) - d\mathbf{h}/dt$ is Hölder continuous in Ω, and by a classical regularity theorem for elliptic equations [16, 23], it follows that $\mathbf{h}(t,\cdot) \in \mathbf{C}^{2+\delta}(\Omega)$ for some $\delta > 0$, i.e., has second order Hölder continuous derivatives in the space variable and is thus a regular solution. ∎

2.4. Existence of minimizers. Having defined the regularization functional, we discuss in this section the existence of minimizers of the global energy functional

$$\mathcal{I}(\mathbf{h}) = \mathcal{J}(\mathbf{h}) + \kappa \int_\Omega \varphi(D\mathbf{h}(\mathbf{x}))\, d\mathbf{x}. \tag{2.3}$$

We assume that $\mathcal{J}(\mathbf{h})$ is continuous in $\mathbf{h}$ and bounded below. These properties are shown for the statistical dissimilarity functionals $\mathcal{J}(\mathbf{h})$ that we study in Proposition 3.1 and Theorem 4.18. In this case, a classical result (see, e.g., Chapter 8, Theorem 2, in [19]) shows that a minimizer exists if φ is convex and coercive. We readily check that φ given in (2.2) satisfies these hypotheses. As pointed out in [2], because of the smoothness of $\partial_i I_1$, $\mathbf{T}_{I_1}$ has strictly positive eigenvalues and therefore, clearly, the mapping $\varphi : \mathbf{X} \in \mathbb{R}^n \mapsto \mathbf{X}\mathbf{T}_{I_1}\mathbf{X}^T \in \mathbb{R}^+$ is convex. Then the functional $\mathcal{R}(\mathbf{h}) = \int_\Omega \varphi(D\mathbf{h}(\mathbf{x}))\, d\mathbf{x}$ satisfies the coercivity inequality, i.e., there exist $c_1 > 0$, $c_2 \ge 0$ such that $\varphi(D\mathbf{h}(\mathbf{x})) \ge c_1|D\mathbf{h}|^2 - c_2$, since we have $\nabla u^T \mathbf{T}_{I_1} \nabla u \ge \theta|\nabla u|^2$ for all $\mathbf{x} \in \Omega$, where $\theta > 0$ is the smallest eigenvalue of $\mathbf{T}_{I_1}$ in Ω.

2.5. Existence of a limiting steady state solution. Theorem 1.2 proves the existence of a solution of the initial-value problem (1.2) or (1.3). Proposition 2.4 gives some regularity properties of this solution. We also have to show that it satisfies the Euler–Lagrange equation in the limit. In order to see this, we introduce the function $L : \mathbb{R}^+ \to \mathbb{R}$ defined by

$$L(t) = \mathcal{I}(\mathbf{h}(t)). \tag{2.4}$$

The properties of L relevant to our goal are stated in the following.

PROPOSITION 2.5. *The function L is bounded below and differentiable. Its derivative is given by*

$$\frac{dL}{dt} = -\left\| \frac{d\mathbf{h}(t)}{dt} \right\|_H^2$$

and hence is continuous.

Proof. The boundedness of L follows from Proposition 3.1 in the mutual information case and from Theorem 4.18 in the local cross-covariance case. In order to prove differentiability we consider

$$\frac{L(t+\varepsilon) - L(t)}{\varepsilon} = \frac{\mathcal{I}(\mathbf{h}(t+\varepsilon)) - \mathcal{I}(\mathbf{h}(t))}{\varepsilon}. \tag{2.5}$$

We then use the first order Taylor expansion with integral remainder of the C^1 function $\mathbf{h}$,

$$\mathbf{h}(t+\varepsilon) = \mathbf{h}(t) + \varepsilon \int_0^1 \frac{d\mathbf{h}}{dt}(t + \zeta\varepsilon) d\zeta, \tag{2.6}$$

and replace $\mathbf{h}(t+\varepsilon)$ by the right-hand side of the previous equation on the right-hand side of (2.5). We denote by $\mathbf{k}(t,\varepsilon)$ the integral on the right-hand side of (2.6) and obtain

$$\frac{L(t+\varepsilon)-L(t)}{\varepsilon} = \frac{\mathcal{I}(\mathbf{h}(t)+\varepsilon\mathbf{k}(t,\varepsilon))-\mathcal{I}(\mathbf{h}(t))}{\varepsilon}. \tag{2.7}$$

In the proof of Proposition 2.4 we have shown the continuity of $d\mathbf{h}/dt$ on $]0,+\infty[\times\overline{\Omega}$; hence on $[t,t+\varepsilon]\times\overline{\Omega}$, $t>0$. This shows that $\mathbf{k}(t,\varepsilon)\in\mathbf{C}(\overline{\Omega})$, $t>0$. In fact, (2.5), Theorem 1.2, and Proposition 2.4 show that $\mathbf{k}(t,\varepsilon)\in\mathbf{C}^2(\Omega)\cap\mathcal{D}(A)$.

We would now like to take the limit of both sides when $\varepsilon\to 0$. Because of the similarity of the right-hand side of (2.7) with (1.1), we define

$$\tilde{\delta}_{\mathbf{k}}\mathcal{I}(\mathbf{h}(t)) = \lim_{\varepsilon\to 0}\frac{\mathcal{I}(\mathbf{h}(t)+\varepsilon\mathbf{k}(t,\varepsilon))-\mathcal{I}(\mathbf{h}(t))}{\varepsilon}. \tag{2.8}$$

The difference is that the function $\mathbf{k}$ here also depends upon ε, unlike in (1.1). We prove that this limit is well defined and equal to $\delta_{\mathbf{k}(t,0)}\mathcal{I}(\mathbf{h}(t))$.

First we show that $\lim_{\varepsilon\to 0}\mathbf{k}(t,\varepsilon)=\mathbf{k}(t,0)$ almost everywhere in Ω. Indeed,

$$\begin{aligned}\|\mathbf{k}(t,\varepsilon)-\mathbf{k}(t,0)\|_H &\le \int_0^1 \left\|\frac{d\mathbf{h}}{dt}(t+\zeta\varepsilon)-\frac{d\mathbf{h}}{dt}(t)\right\|_H d\zeta\\ &\le k_\alpha\int_0^1 \left\|\frac{d\mathbf{h}}{dt}(t+\zeta\varepsilon)-\frac{d\mathbf{h}}{dt}(t)\right\|_{H_\alpha} d\zeta\end{aligned}$$

for some α, $0<\alpha\le 1$ (Theorem 1.2); k_α is defined in the proof of the same theorem. Because of the same theorem, the function $t\to d\mathbf{h}/dt$ from $]0,+\infty[$ into H_α is Hölder continuous, say with exponent $\gamma>0$. This proves that

$$\int_0^1 \left\|\frac{d\mathbf{h}}{dt}(t+\zeta\varepsilon)-\frac{d\mathbf{h}}{dt}(t)\right\|_{H_\alpha} d\zeta \le C\varepsilon^\gamma,$$

where C is independent of ε and the convergence in H follows. Since $H=\mathbf{L}^2(\Omega)$ and $\mathbf{k}(t,\varepsilon)$ is continuous in $\overline{\Omega}$, the L^2 convergence implies the convergence almost everywhere.

The proof that $\tilde{\delta}_{\mathbf{k}}\mathcal{I}(\mathbf{h}(t))=\delta_{\mathbf{k}(t,0)}\mathcal{I}(\mathbf{h}(t))$ follows from this result, the computations done in sections 3.1.2 and 3.2, the fact that $\mathbf{k}(t,\varepsilon)$ converges almost everywhere toward $\mathbf{k}(t,0)$, and the fact that each coordinate of $\mathbf{k}(t,\varepsilon)$ is bounded and hence integrable. The last two points guarantee that such integrals as $\int_\Omega \mathbf{k}(t,\varepsilon)(x)\cdot\nabla I_2(x+\mathbf{h}(t,x))\,dx$ converge toward $\int_\Omega \mathbf{k}(t,0)(x)\cdot\nabla I_2(x+\mathbf{h}(t,x))\,dx$ by applying the dominated convergence theorem.

The boundedness of $\mathbf{k}(t,\varepsilon)$ follows from its continuity on the compact set $\overline{\Omega}$.

The continuity of dL/dt follows from the continuity of $d\mathbf{h}/dt$ (Theorem 1.2) and the continuity of the norm in H. □

We have therefore proved that our criterion $\mathcal{I}$ decreases along the "trajectory" $\mathbf{h}(t)$, solution of the initial-value problem (1.3). In order to prove that the solution asymptotically satisfies the Euler–Lagrange equations, we prove the following claim.

PROPOSITION 2.6. *If $\|d\mathbf{h}(t)/dt\|_H$ is bounded on $]0,+\infty[$, the limit when $t\to+\infty$ of the derivative dL/dt of the function $L(t)$ defined by (2.4) is equal to 0.*

Proof. We assume the converse and prove a contradiction. We assume that dL/dt does not go to zero when t goes to infinity. Therefore there exists $\varepsilon>0$ such that for

all $T > 0$ there exists $t > T$ such that $dL(t)/dt < -\varepsilon$. We can therefore construct an infinite, strictly increasing, sequence $\{t_n\}$, $n \geq 0$, $t_n > 0$, such that $dL(t_n)/dt < -\varepsilon$. Let us assume that there exists $\eta(t_n) > 0$ such that $dL(t_n + \eta(t_n))/dt = -\varepsilon/2$ and $dL(t)/dt < -\varepsilon/2$ for all $t_n \leq t < t_n + \eta(t_n)$. If there is no such t_n, we are done; otherwise we choose $t_{n+1} > t_n + \eta(t_n)$. We prove that $\eta(t_n)$ is bounded below. Indeed we write

$$\begin{aligned}\varepsilon/2 &\leq \left|\frac{dL(t_n+\eta(t_n))}{dt} - \frac{dL(t_n)}{dt}\right| = \left|\left\|\frac{d\mathbf{h}}{dt}(t_n+\eta(t_n))\right\|_H^2 - \left\|\frac{d\mathbf{h}}{dt}(t_n)\right\|_H^2\right| \\ &= \left(\left\|\frac{d\mathbf{h}}{dt}(t_n+\eta(t_n))\right\|_H + \left\|\frac{d\mathbf{h}}{dt}(t_n)\right\|_H\right)\left|\left\|\frac{d\mathbf{h}}{dt}(t_n+\eta(t_n))\right\|_H - \left\|\frac{d\mathbf{h}}{dt}(t_n)\right\|_H\right|.\end{aligned}$$

Let A be an upper bound on $\|d\mathbf{h}(t)/dt\|_H$; then the rightmost term in the above equation is less than or equal to

$$2A\left\|\frac{d\mathbf{h}}{dt}(t_n+\eta(t_n)) - \frac{d\mathbf{h}}{dt}(t_n))\right\|_H.$$

According to Theorem 1.2,

$$\left\|\frac{d\mathbf{h}}{dt}(t_n+\eta(t_n)) - \frac{d\mathbf{h}}{dt}(t_n))\right\|_H \leq k_\alpha \left\|\frac{d\mathbf{h}}{dt}(t_n+\eta(t_n)) - \frac{d\mathbf{h}}{dt}(t_n))\right\|_{H_\alpha} \leq k_\alpha(\eta(t_n))^\gamma,$$

and therefore

$$\eta(t_n) \geq \left(\frac{\varepsilon}{4Ak_\alpha}\right)^{1/\gamma}.$$

Consider now the sequence of intervals $[t_n, t_n + \eta(t_n)]$. On each such interval we have $dL/dt \leq -\varepsilon/2$, and therefore

$$L(t_n+\eta(t_n)) \leq L(t_n) - \eta(t_n)\frac{\varepsilon}{2} \leq L(t_n) - \frac{\varepsilon}{2}\left(\frac{\varepsilon}{4Ak_\alpha}\right)^{1/\gamma}.$$

This contradicts the fact that $L(t)$ is bounded below; hence we have

$$\lim_{t\to+\infty}\frac{dL}{dt} = 0. \qquad \square$$

This result shows that, asymptotically, $dL/dt = 0$, therefore $d\mathbf{h}(t)/dt = 0$, and hence the solution of the initial-value problem (1.2) satisfies the Euler–Lagrange equations associated with the energy functional $\mathcal{I}$.

3. Statistical similarity measures and their Euler–Lagrange equations. Our approach to matching multimodal images relies on regarding the intensity values of two different modalities as samples of two random processes. Within this probabilistic framework, the link between the two modalities is characterized by their joint probability density function (pdf). In this section, we compute the variational gradient of two statistical dissimilarity functionals. Among many possible criteria, the cross-covariance and the mutual information provide us with a convenient trade-off between robustness and generality. To be able to evaluate these criteria for a given field $\mathbf{h}$, we consider a nonparametric Parzen estimator [42] for the joint pdf as described

below. This estimator can be either *global*, assuming that the unknown relationship between the intensity values does not vary spatially, or *local*. We give only one example of each case because they are the ones we found to work best in our applications. However, the method and the proofs are quite generic. More examples can be found in [25]. We denote by X the random variable associated with the intensity values of I_1 and by $Y_{\mathbf{h}}$ that associated with the values of $I_2(\mathbf{Id}+\mathbf{h})$. For conciseness, we will use the notation $\mathbf{i}=(i_1,i_2)$ and $\mathbf{I_h}(\mathbf{x})=(I_1(\mathbf{x}),I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})))$.

3.1. Global mutual information. Our estimator is based on a one-dimensional (1D) normalized Gaussian kernel of variance β,

$$g_\beta(i)=\frac{1}{\sqrt{2\pi\beta}}\exp\left(\frac{-i^2}{2\beta}\right),$$

from which we construct $G_\beta(\mathbf{i})=g_\beta(i_1)\,g_\beta(i_2)$:

$$P(\mathbf{i},\mathbf{h})=\frac{1}{|\Omega|}\int_\Omega G_\beta(\mathbf{I_h}(\mathbf{x})-\mathbf{i})\,d\mathbf{x}.$$

Note that for each pair of intensities $\mathbf{i}\in\mathbb{R}^2$ the value of the estimated joint pdf is a nonlinear functional of $\mathbf{h}$. Also note that it is strictly positive. Its first variation is obtained by applying (1.1):

$$\frac{P(\mathbf{i},\mathbf{h}+\varepsilon\mathbf{k})-P(\mathbf{i},\mathbf{h})}{\varepsilon}$$
$$=\frac{1}{\varepsilon|\Omega|}\int_\Omega g_\beta(I_1(\mathbf{x})-i_1)\left(g_\beta(I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})+\varepsilon\mathbf{k}(\mathbf{x}))-i_2)-g_\beta(I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))-i_2)\right)d\mathbf{x}.$$

We use the first order Taylor expansion with integral remainder of the C^1 function g_β for the second factor within the integral:

$$\frac{P(\mathbf{i},\mathbf{h}+\varepsilon\mathbf{k})-P(\mathbf{i},\mathbf{h})}{\varepsilon}$$
$$=\frac{1}{|\Omega|}\int_\Omega\left[g_\beta(I_1(\mathbf{x})-i_1)\frac{I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})+\varepsilon\mathbf{k}(\mathbf{x}))-I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))}{\varepsilon}\right.$$
$$\left.\times\int_0^1 g'_\beta(I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))-i_2+\zeta(I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})+\varepsilon\mathbf{k}(\mathbf{x}))-I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))))\,d\zeta\right]d\mathbf{x}.$$

Taking the limit when $\varepsilon\to 0$, we obtain

$$\delta_{\mathbf{k}}P(\mathbf{i},\mathbf{h})=\frac{1}{|\Omega|}\int_\Omega \partial_2\,G_\beta(\mathbf{I_h}(\mathbf{x})-\mathbf{i})\nabla I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))\cdot\mathbf{k}(\mathbf{x})\,d\mathbf{x}, \tag{3.1}$$

where ∂_2 indicates the first order partial derivative with respect to the second variable.

Using this estimate, we first consider the maximization of mutual information, a concept which is borrowed from information theory. Given two random variables X and Y, their mutual information is defined as

$$\mathbf{MI}(X,Y)=\mathcal{H}(X)+\mathcal{H}(Y)-\mathcal{H}(X,Y),$$

where $\mathcal{H}$ stands for the differential entropy. The mutual information is positive and symmetric and measures how the intensity distributions of two images fail to be

independent. It can be defined in terms of the joint pdf P and its marginals $p(i_1) = \int_{\mathbb{R}} P(\mathbf{i},\mathbf{h})\, di_2$ and $p(i_2,\mathbf{h}) = \int_{\mathbb{R}} P(\mathbf{i},\mathbf{h})\, di_1$. These last two functions are also strictly positive. The following short notation will be useful:

$$E^{\mathrm{MI}}(\mathbf{i},\mathbf{h}) = -\log \frac{P(\mathbf{i},\mathbf{h})}{p(i_1)\, p(i_2,\mathbf{h})}. \tag{3.2}$$

The dissimilarity functional based on mutual information is then defined as the expected value of the function E^{MI}:

$$\mathcal{J}_{\mathrm{MI}}(\mathbf{h}) = -\mathbf{MI}(X, Y_{\mathbf{h}}) = \int_{\mathbb{R}^2} P(\mathbf{i},\mathbf{h})\, E^{\mathrm{MI}}(\mathbf{i},\mathbf{h})\, d\mathbf{i}.$$

3.1.1. Continuity of $\mathcal{J}_{\mathbf{MI}}(\mathbf{h})$. Recall that the existence of minimizers for $\mathcal{I}(\mathbf{h})$ was discussed by assuming the continuity and boundedness of $\mathcal{J}(\mathbf{h})$.

PROPOSITION 3.1. *Let $\mathbf{h}_n, n = 1,\dots,\infty$, be a sequence of functions of H such that $\mathbf{h}_n \to \mathbf{h}$ almost everywhere in Ω. Then $\mathcal{J}_{\mathrm{MI}}(\mathbf{h}_n) \to \mathcal{J}_{\mathrm{MI}}(\mathbf{h})$. Moreover, $|\mathcal{J}_{\mathrm{MI}}(\mathbf{h})|$ is bounded.*

Proof. Because I_2 and g_β are continuous, $G_\beta(\mathbf{I}_{\mathbf{h}_n}(\mathbf{x}) - \mathbf{i}) \to G_\beta(\mathbf{I}_{\mathbf{h}}(\mathbf{x}) - \mathbf{i})$ almost everyone in Ω for all $\mathbf{i}$. Since $G_\beta(\mathbf{I}_{\mathbf{h}_n}(\mathbf{x}) - \mathbf{i}) \le g_\beta(0)^2$, the dominated convergence theorem implies that $P(\mathbf{i},\mathbf{h}_n) \to P(\mathbf{i},\mathbf{h})$ for all $\mathbf{i} \in \mathbb{R}^2$. A similar reasoning shows that $p(i_2,\mathbf{h}_n) \to p(i_2,\mathbf{h})$ for all $i_2 \in \mathbb{R}$. Hence, the logarithm being continuous, and $p(i_1)$, $P(\mathbf{i},\mathbf{h}_n)$, $P(\mathbf{i},\mathbf{h})$, $p(i_2,\mathbf{h}_n)$, and $p(i_2,\mathbf{h})$ being > 0,

$$P(\mathbf{i},\mathbf{h}_n) \log \frac{P(\mathbf{i},\mathbf{h}_n)}{p(i_1)p(i_2,\mathbf{h}_n)} \to P(\mathbf{i},\mathbf{h}) \log \frac{P(\mathbf{i},\mathbf{h})}{p(i_1)p(i_2,\mathbf{h})} \quad \forall \mathbf{i} \in \mathbb{R}^2.$$

We next consider three cases to find an upper bound for $P(\mathbf{i},\mathbf{h}_n)\big|\log \frac{P(\mathbf{i},\mathbf{h}_n)}{p(i_1)p(i_2,\mathbf{h}_n)}\big|$.

We detail only the first one: $i_2 \le 0$ This is the case where

$$0 \le |i_2| \le |i_2 - I_2(\mathbf{x} + \mathbf{h}_n(\mathbf{x}))| \ \le |i_2 - \mathcal{A}| \ \ n \ge 1.$$

Hence

$$g_\beta(i_2 - \mathcal{A}) \le g_\beta(i_2 - I_2(\mathbf{x} + \mathbf{h}_n(\mathbf{x}))) \ \le g_\beta(i_2) \ \ n \ge 1.$$

This yields

$$\frac{g_\beta(i_2 - \mathcal{A})}{g_\beta(i_2)} \le \frac{P(\mathbf{i},\mathbf{h}_n)}{p(i_1)p(i_2,\mathbf{h}_n)} \le \frac{g_\beta(i_2)}{g_\beta(i_2 - \mathcal{A})}$$

and

$$\left|\log \frac{P(\mathbf{i},\mathbf{h}_n)}{p(i_1)p(i_2,\mathbf{h}_n)}\right| \le \log \frac{g_\beta(i_2)}{g_\beta(i_2 - \mathcal{A})},$$

and therefore

$$P(\mathbf{i},\mathbf{h}_n)\left|\log \frac{P(\mathbf{i},\mathbf{h}_n)}{p(i_1)p(i_2,\mathbf{h}_n)}\right| \le g_\beta(i_2)p(i_1) \log \frac{g_\beta(i_2)}{g_\beta(i_2 - \mathcal{A})}.$$

The function on the right-hand side is continuous and integrable in $\mathbb{R}\times]-\infty,\mathcal{A}]$.

The next two cases, $0 \le i_2 \le \mathcal{A}$ and $i_2 \ge \mathcal{A}$, are left to the reader. Combining the three cases, the dominated convergence theorem implies that

$$\mathcal{J}_{\mathrm{MI}}(\mathbf{h}_n) = -\int_{\mathbb{R}^2} P(\mathbf{i},\mathbf{h}_n) \log \frac{P(\mathbf{i},\mathbf{h}_n)}{p(i_1)p(i_2,\mathbf{h}_n)}\, d\mathbf{i} \to \mathcal{J}_{\mathrm{MI}}(\mathbf{h}) = -\int_{\mathbb{R}^2} P_{\mathbf{h}}(\mathbf{i}) \log \frac{P_{\mathbf{h}}(\mathbf{i})}{p(i_1)p_{\mathbf{h}}(i_2)}\, d\mathbf{i}.$$

The proof also shows that $|\mathcal{J}_{\mathrm{MI}}(\mathbf{h})|$ is bounded. ∎

3.1.2. Euler–Lagrange equation. We do an explicit computation of the first variation of $\mathcal{J}_{\mathrm{MI}}$ by applying (1.1). First we obtain

$$\delta_{\mathbf{k}} E^{\mathrm{MI}}(\mathbf{i},\mathbf{h}) = -\frac{\delta_{\mathbf{k}} P(\mathbf{i},\mathbf{h})}{P(\mathbf{i},\mathbf{h})} + \frac{\delta_{\mathbf{k}} p(i_2,\mathbf{h})}{p(i_2,\mathbf{h})}.$$

We then write

$$\begin{aligned}\delta_{\mathbf{k}} \mathcal{J}_{\mathrm{MI}}(\mathbf{h}) &= \int_{\mathbb{R}^2} \left[\delta_{\mathbf{k}} P(\mathbf{i},\mathbf{h})\ E^{\mathrm{MI}}(\mathbf{i},\mathbf{h}) + P(\mathbf{i},\mathbf{h})\ \delta_{\mathbf{k}}\ E^{\mathrm{MI}}(\mathbf{i},\mathbf{h})\right] d\mathbf{i} \\ &= \int_{\mathbb{R}^2} \left[(E^{\mathrm{MI}}(\mathbf{i},\mathbf{h}) - 1)\,\delta_{\mathbf{k}} P(\mathbf{i},\mathbf{h})\ + \frac{P(\mathbf{i},\mathbf{h})}{p(i_2,\mathbf{h})}\delta_{\mathbf{k}} p(i_2,\mathbf{h})\right] d\mathbf{i}.\end{aligned}$$

Using the fact that

$$\int_{\mathbb{R}^2} \frac{P(\mathbf{i},\mathbf{h})}{p(i_2,\mathbf{h})}\delta_{\mathbf{k}} p(i_2,\mathbf{h})\, d\mathbf{i} = \int_{\mathbb{R}} \delta_{\mathbf{k}} p(i_2,\mathbf{h})\, di_2 = 0,$$

this yields

$$\delta_{\mathbf{k}} \mathcal{J}_{\mathrm{MI}}(\mathbf{h}) = \int_{\mathbb{R}^2} (E^{\mathrm{MI}}(\mathbf{i},\mathbf{h}) - 1)\,\delta_{\mathbf{k}} P(\mathbf{i},\mathbf{h})\, d\mathbf{i}.$$

We then apply (3.1) to obtain

$$\delta_{\mathbf{k}} \mathcal{J}_{\mathrm{MI}}(\mathbf{h}) = \frac{1}{|\Omega|}\int_{\mathbb{R}^2}\int_{\Omega} (E^{\mathrm{MI}}(\mathbf{i},\mathbf{h}) - 1)\,\partial_2 G_\beta(\mathbf{I_h}(\mathbf{x}) - \mathbf{i})\nabla I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))\cdot\mathbf{k}(\mathbf{x})\ d\mathbf{x}\, d\mathbf{i}.$$

A convolution with respect to the intensity variable $\mathbf{i}$ appears in this expression. It commutes with the derivative ∂_2 with respect to the second intensity variable i_2, and therefore

$$\delta_{\mathbf{k}} \mathcal{J}_{\mathrm{MI}}(\mathbf{h}) = \frac{1}{|\Omega|}\int_{\Omega} (G_\beta \star \partial_2\, E^{\mathrm{MI}})\big(\mathbf{I_h}(\mathbf{x}),\mathbf{h}\big)\nabla I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))\cdot\mathbf{k}(\mathbf{x})\ d\mathbf{x}.$$

To simplify notation, we denote by L_{MI} the function $\frac{1}{|\Omega|}\partial_2\, E^{\mathrm{MI}}$. We then define the function $f_{\mathrm{MI}} : \mathbb{R}^2 \times H \to \mathbb{R}$:

$$f_{\mathrm{MI}}(\mathbf{i},\mathbf{h}) = G_\beta \star L_{\mathrm{MI}}(\mathbf{i},\mathbf{h}). \tag{3.3}$$

It is easily verified that

$$L_{\mathrm{MI}}(\mathbf{i},\mathbf{h}) = -\frac{1}{|\Omega|}\ \left(\frac{\partial_2 P}{P}(\mathbf{i},\mathbf{h}) - \frac{p'}{p}(i_2,\mathbf{h})\right) = \frac{1}{\beta}\ (r(i_2,\mathbf{h}) - R(\mathbf{i},\mathbf{h})), \tag{3.4}$$

where

$$r(i_2,\mathbf{h}) = \frac{1}{p(i_2,\mathbf{h})}\frac{1}{\beta|\Omega|}\int_{\Omega} I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))\ g_\beta(I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})) - i_2)\ d\mathbf{x}, \tag{3.5}$$

$$R(\mathbf{i},\mathbf{h}) = \frac{1}{P(\mathbf{i},\mathbf{h})}\frac{1}{\beta|\Omega|}\int_{\Omega} I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))\ G_\beta(\mathbf{I_h}(\mathbf{x}) - \mathbf{i})\ d\mathbf{x}. \tag{3.6}$$

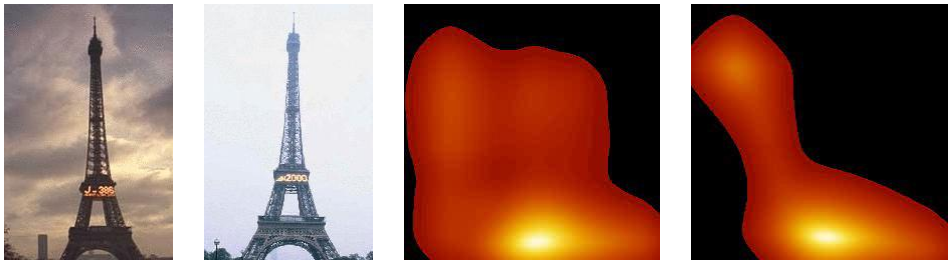


FIG. 1. *Effect on $P(\mathbf{i},\mathbf{h})$ of minimizing $\mathcal{J}_{MI}(\mathbf{h})$ with respect to $\mathbf{h}$: the two images on the right show the estimated joint pdf between the two images on the left before and after optimization. Notice that the joint pdf has been clustered, but the final shape of its support remains nonfunctional.*

Since $I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})) \in [0,\, \mathcal{A}]$ for all $\mathbf{h} \in H$ and for all $\mathbf{x} \in \Omega$, we have

$$0 \leq r(i_2,\mathbf{h}),\ R(\mathbf{i},\mathbf{h}) \leq \frac{\mathcal{A}}{\beta} \quad \forall \mathbf{h} \in H. \tag{3.7}$$

For each $\mathbf{h} \in H$, the mapping $\mathbf{k} \to \delta_{\mathbf{k}}\mathcal{J}_{\mathrm{MI}}(\mathbf{h})$ is clearly linear. To show that it is continuous it is sufficient, according to Schwarz inequality, to show that the function $x \to f_{\mathrm{MI}}(\mathbf{I_h}(\mathbf{x}),\mathbf{h})\nabla I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))$ is bounded, and this is a consequence of Theorem 4.11. The variational gradient $\nabla_H \mathcal{J}_{\mathrm{MI}}(\mathbf{h})$ of $\mathcal{J}_{\mathrm{MI}}$ can therefore be defined, and its expression is given by

$$\nabla_H \mathcal{J}_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) = f_{\mathrm{MI}}\big(\mathbf{I_h}(\mathbf{x}),\mathbf{h}\big)\ \nabla I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})).$$

The function L_{MI} plays the role of an intensity comparison function. Its first term $\partial_2 P(\mathbf{i},\mathbf{h})/P(\mathbf{i},\mathbf{h})$ tends to cluster the joint pdf, while the term $-p'(i_2,\mathbf{h})/p(i_2,\mathbf{h})$ tries to prevent the marginal law $p(i_2,\mathbf{h})$ from becoming too clustered; i.e., it keeps the intensities of $I_2(\mathbf{Id}+\mathbf{h})$ as unpredictable as possible. For an example of the effect on $P(\mathbf{i},\mathbf{h})$ of minimizing $\mathcal{J}_{\mathrm{MI}}(\mathbf{h})$ with respect to $\mathbf{h}$, see Figure 1. The reader will notice that the resulting joint pdf is more localized than the original one but that the minimization has not imposed a strong functional relation of the type $i_1 = f(i_2)$, unlike the method described in the next section.

According to the notation introduced in section 1.2, we define $F_{\mathrm{MI}} : H \to H$ by

$$F_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) = -f_{\mathrm{MI}}\big(\mathbf{I_h}(\mathbf{x}),\mathbf{h}\big)\ \nabla I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})). \tag{3.8}$$

3.2. Local cross-covariance. An interesting generalization is to make the probability density estimator local, since it allows one to take into account nonstationarities in the relation between intensities. To do this, we build an estimate in the neighborhood of each point $\mathbf{x}_0$ in Ω. This is achieved by weighting our previous estimate with a normalized spatial Gaussian of variance γ:

$$G_\gamma(\mathbf{x}) = \frac{1}{2\pi\gamma} e^{-\frac{|\mathbf{x}|^2}{2\gamma}}.$$

This means that to each point $\mathbf{x}_0$ of Ω we associate a joint pdf defined by

$$P(\mathbf{i},\mathbf{h},\mathbf{x}_0) = \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega G_\beta(\mathbf{I_h}(\mathbf{x}) - \mathbf{i}) G_\gamma(\mathbf{x}-\mathbf{x}_0)\, d\mathbf{x},$$

where $\mathcal{G}_\gamma(\mathbf{x}_0) = \int_\Omega G_\gamma(\mathbf{x}-\mathbf{x}_0)\ d\mathbf{x}$. This new pdf is along the line of the ideas discussed in [30], except that we consider a bidimensional local histogram at each

TABLE 3.1

Definitions of the local mean μ_1 and variance v_1 of X, the local mean μ_2 and variance v_2 of $Y_{\mathbf{h}}$, and the local correlation $v_{1,2}$ of X and $Y_{\mathbf{h}}$.

$\mu_1(\mathbf{x}) = \int_{\mathbb{R}} i_1\, p(i_1, \mathbf{x})\, di_1$	$v_1(\mathbf{x}) = \int_{\mathbb{R}} i_1^2\, p(i_1, \mathbf{x})\, di_1 - \mu_1^2(\mathbf{x})$
$\mu_2(\mathbf{h}, \mathbf{x}) = \int_{\mathbb{R}} i_2\, p(i_2, \mathbf{h}, \mathbf{x})\, di_2$	$v_2(\mathbf{h}, \mathbf{x}) = \int_{\mathbb{R}} i_2^2\, p(i_2, \mathbf{h}, \mathbf{x})\, di_2 - \mu_2(\mathbf{h}, \mathbf{x})^2$
$v_{1,2}(\mathbf{h}, \mathbf{x}) = \int_{\mathbb{R}^2} i_1\, i_2\, P(\mathbf{i}, \mathbf{h}, \mathbf{x})\, d\mathbf{i} - \mu_1(\mathbf{x})\, \mu_2(\mathbf{h}, \mathbf{x})$	

point. Its marginals are now given by $p(i_1, \mathbf{x}) = \int_{\mathbb{R}} P(\mathbf{i}, \mathbf{h}, \mathbf{x})\ di_2$ and $p(i_2, \mathbf{h}, \mathbf{x}) = \int_{\mathbb{R}} P(\mathbf{i}, \mathbf{h}, \mathbf{x})\ di_1$. Using these local estimates, we consider the case of the cross-covariance, which has been widely used as a robust comparison function for image matching. Within recent energy-minimization approaches relying on the computation of its gradient, we can mention, for instance, the works of Faugeras and Keriven [21], Cachier and Pennec [11], and Netsch et al. [39]. The cross-covariance, being a measure of the *affine* dependency between the intensities, is more constraining than the mutual information. A dissimilarity functional is obtained as a function of the quantities defined in Table 3.1 by averaging the local cross-covariance

$$\mathcal{J}_{\mathrm{CC}}(\mathbf{h}) = \int_\Omega \mathcal{J}_{\mathrm{CC}}(\mathbf{h}, \mathbf{x})\ d\mathbf{x} = -\int_\Omega \frac{v_{1,2}(\mathbf{h}, \mathbf{x})^2}{v_1(\mathbf{x})\, v_2(\mathbf{h}, \mathbf{x})}\ d\mathbf{x}.$$

The minus sign simply reflects the fact that we want to minimize the dissimilarity. The continuity and boundedness of this criterion (needed for proving the existence of a minimizer) are a consequence of Theorem 4.18. In a way similar to the case of the mutual information, its first order variation is well defined and defines a gradient given by

$$\nabla_{\!H} \mathcal{J}_{\mathbf{CC}}(\mathbf{h})(\mathbf{x}) = f_{\mathrm{CC}}\big(\mathbf{I}_{\mathbf{h}}(\mathbf{x}), \mathbf{h}, \mathbf{x}\big)\ \nabla I_2(\mathbf{x} + \mathbf{h}(\mathbf{x})),$$

where

$$f_{\mathrm{CC}}(\mathbf{i}, \mathbf{h}, \mathbf{x}) = G_\gamma \star G_\beta \star L_{\mathrm{CC}}(\mathbf{i}, \mathbf{h}, \mathbf{x}),$$

$$L_{\mathrm{CC}}(\mathbf{i}, \mathbf{h}, \mathbf{x}) = \frac{1}{\mathcal{G}_\gamma(\mathbf{x})} \partial_2\, E^{\mathrm{CC}}(\mathbf{i}, \mathbf{h}, \mathbf{x}),$$

and

$$E^{\mathrm{CC}}(\mathbf{i}, \mathbf{h}, \mathbf{x}) = -\frac{1}{v_1(\mathbf{x})\ v_2(\mathbf{h}, \mathbf{x})} \Big(2\ v_{\ 1,2}(\mathbf{h}, \mathbf{x})\, i_2\, (i_1 - \mu_1(\mathbf{x})) + \ \mathcal{J}_{\mathbf{CC}}(\mathbf{h}, \mathbf{x})\, v_1(\mathbf{x})\, i_2\, (i_2 - 2\, \mu_2(\mathbf{h}, \mathbf{x})) \Big).$$

Hence

(3.9)

$$L_{\mathrm{CC}}(\mathbf{i}, \mathbf{h}, \mathbf{x}) = -\frac{2}{\mathcal{G}_\gamma(\mathbf{x})} \left(\frac{v_{1,2}(\mathbf{h}, \mathbf{x})}{v_2(\mathbf{h}, \mathbf{x})} \left(\frac{i_1 - \mu_1(\mathbf{x})}{v_1(\mathbf{x})} \right) + \mathcal{J}_{\mathrm{CC}}(\mathbf{h}, \mathbf{x}) \left(\frac{i_2 - \mu_2(\mathbf{h}, \mathbf{x})}{v_2(\mathbf{h}, \mathbf{x})} \right) \right).$$

Notice that L_{CC} is an affine function of $\mathbf{i}$, and therefore

$$G_\beta \star L_{\mathrm{CC}}(\mathbf{i},\mathbf{h},\mathbf{x}) = L_{\mathrm{CC}}(\mathbf{i},\mathbf{h},\mathbf{x}).$$

Thus

$$\begin{aligned} f_{\mathrm{CC}}(\mathbf{i},\mathbf{h},\mathbf{x}) &= G_\gamma \star L_{\mathrm{CC}}(\mathbf{i},\mathbf{h},\mathbf{x}) \\ &= -G_\gamma \star \frac{2}{\mathcal{G}_\gamma(\mathbf{x})} \left(\frac{v_{1,2}(\mathbf{h},\mathbf{x})}{v_2(\mathbf{h},\mathbf{x})} \left(\frac{i_1 - \mu_1(\mathbf{x})}{v_1(\mathbf{x})} \right) + \mathcal{J}_{\mathrm{CC}}(\mathbf{h},\mathbf{x}) \left(\frac{i_2 - \mu_2(\mathbf{h},\mathbf{x})}{v_2(\mathbf{h},\mathbf{x})} \right) \right). \end{aligned}$$

As in the mutual information case, the function L_{CC} compares intensities in the two images. It shows that minimizing $\mathcal{J}_{\mathbf{CC}}$ with respect to the field $\mathbf{h}$ amounts to attempting to make the pair of intensities at corresponding pixels lie on a straight line in $\mathbb{R}^2$.

According to the notation introduced in section 1.1, we define $F_{\mathrm{CC}} : H \to H$ by

$$F_{\mathrm{CC}}(\mathbf{h})(\mathbf{x}) = -f_{\mathrm{CC}}(\mathbf{I}_{\mathbf{h}}(\mathbf{x}),\mathbf{h},\mathbf{x}) \nabla I_2(\mathbf{x} + \mathbf{h}(\mathbf{x})). \tag{3.10}$$

For an example of the effect of minimizing $\mathcal{J}_{\mathrm{CC}}(\mathbf{h})$ with respect to $\mathbf{h}$ on $P(\mathbf{i},\mathbf{h})$, see Figure 2. The reader will notice that the resulting joint pdf is more localized than the original one and that the minimization has imposed an affine relation between the intensities i_1 and i_2, unlike the method described in the previous section.

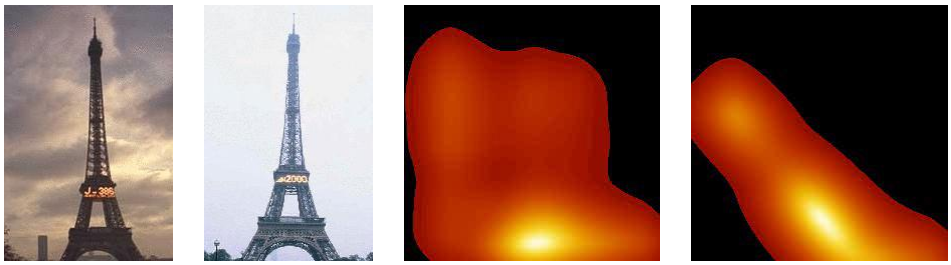


FIG. 2. *Effect on $P(\mathbf{i},\mathbf{h})$ of minimizing $\mathcal{J}_{\mathrm{CC}}(\mathbf{h})$ with respect to $\mathbf{h}$: the two images on the right show the estimated joint pdf between the two images on the left before and after optimization. Notice that the joint pdf has been clustered, and the final shape of its support approaches an affine function. The global estimation of the cross-covariance has been used to illustrate this effect. In the local case, each locally estimated pdf would undergo this effect.*

4. Lipschitz continuity of the matching functions. This section is devoted to proving that the functions F_{MI} and F_{CC} are Lipschitz continuous. We begin with some elementary results on Lipschitz continuous functions that will be used very often in what follows. We state them here without proof.

PROPOSITION 4.1. *Let $\mathcal{H}$ be a Banach space, and let us denote its norm by $\|.\|_{\mathcal{H}}$. Let f_i, $i = 1,2 : \mathcal{H} \to \mathbb{R}$, be two Lipschitz continuous functions. We have the following:*

1. *$f_1 + f_2$ is Lipschitz continuous.*
2. *If f_1 and f_2 are bounded, then the product $f_1 f_2$ is Lipschitz continuous.*
3. *If $f_2 > 0$ and if f_1 and f_2 are bounded, then the ratio $\frac{f_1}{f_2}$ is Lipschitz continuous.*

In what follows, we will need the following definitions and notations.

DEFINITION 4.2. *We denote by* $\mathcal{H}_1 = [0, \mathcal{A}] \times H$ *and* $\mathcal{H}_2 = [0, \mathcal{A}]^2 \times H$ *the Banach spaces equipped with the norms* $\|(z, \mathbf{h})\|_{\mathcal{H}_1} = z + \|\mathbf{h}\|_H$ *and* $\|(z_1, z_2, \mathbf{h})\|_{\mathcal{H}_2} = z_1 + z_2 + \|\mathbf{h}\|_H$*, respectively.*

We will use several times the following (obvious) result.

LEMMA 4.3. *Let* $f : \mathcal{H}_2 \to \mathbb{R}$ *be such that* $(z_1, z_2) \to f(z_1, z_2, \mathbf{h})$ *is Lipschitz continuous for all* $\mathbf{h}$ *with a Lipschitz constant* l_f *independent of* $\mathbf{h}$ *and such that* $\mathbf{h} \to f(z_1, z_2, \mathbf{h})$ *is Lipschitz continuous for all* (z_1, z_2) *with a Lipschitz constant* L_f *independent of* (z_1, z_2)*. Then* f *is Lipschitz continuous.*

4.1. Global mutual information. In the following, we will use the function $L_{\mathrm{MI}} : [0, \mathcal{A}]^2 \times H \to \mathbb{R}$ defined in (3.4)–(3.6) We then consider the result f_{MI} of convolving L_{MI} with G_β, i.e., the two functions $s : \mathcal{H}_1 \to \mathbb{R}$, defined as

$$s(z_2, \mathbf{h}) = (g_\beta \star r)(z_2, \mathbf{h}) = \int_{\mathbb{R}} g_\beta(z_2 - i_2) r(i_2, \mathbf{h})\, di_2, \tag{4.1}$$

and $S : \mathcal{H}_2 \to \mathbb{R}$, defined as

$$S(z_1, z_2, \mathbf{h}) = (G_\beta \star R)(\mathbf{z}, \mathbf{h}) = \int_{\mathbb{R}^2} G_\beta(\mathbf{z} - \mathbf{i}) R(\mathbf{i}, \mathbf{h})\, d\mathbf{i}. \tag{4.2}$$

We prove a series of propositions.

PROPOSITION 4.4. *For each* $\mathbf{h} \in H$*, the function* $[0, \mathcal{A}] \to \mathbb{R}^+$ *defined by* $z_2 \to s(z_2, \mathbf{h})$ *is Lipschitz continuous with a Lipschitz constant* l_s*, which is independent of* $\mathbf{h}$*. Moreover, it is bounded by* $\frac{\mathcal{A}}{\beta}$*.*

Proof. The second part follows from (3.7). We then prove that the magnitude of the derivative of the function is bounded independently of $\mathbf{h}$. Indeed

$$\begin{aligned} |s'(z_2, \mathbf{h})| &= \frac{1}{\beta} \left| \int_{-\infty}^{+\infty} (z_2 - i_2) g_\beta(z_2 - i_2) r(i_2, \mathbf{h}) di_2 \right| \\ &\leq \frac{\mathcal{A}}{\beta} \int_{-\infty}^{+\infty} |z_2 - i_2| g_\beta(z_2 - i_2) di_2 = 2\frac{\mathcal{A}}{\beta^2}. \end{aligned}$$

∎

PROPOSITION 4.5. *For each* $z_2 \in [0, \mathcal{A}]$*, the function* $\mathbf{h} \to s(z_2, \mathbf{h}) : H \to \mathbb{R}$ *is Lipschitz continuous on* H *with Lipschitz constant* L_s*, which is independent of* $z_2 \in [0, \mathcal{A}]$*.*

Proof. We consider

$$s(z_2, \mathbf{h}_1) - s(z_2, \mathbf{h}_2) = \int_{\mathbb{R}} g_\beta(z_2 - i_2) \left(r(i_2, \mathbf{h}_1) - r(i_2, \mathbf{h}_2) \right) di_2. \tag{4.3}$$

According to (3.5), $r(i_2, \mathbf{h})$ is proportional to the ratio $N(i_2, \mathbf{h})/p(i_2, \mathbf{h})$ of the two functions

$$N(i_2, \mathbf{h}) = \int_\Omega I_2(\mathbf{x} + \mathbf{h}(\mathbf{x})) g_\beta(I_2(\mathbf{x} + \mathbf{h}(\mathbf{x})) - i_2)\, d\mathbf{x}$$

and

$$p(i_2, \mathbf{h}) = \int_\Omega g_\beta(I_2(\mathbf{x} + \mathbf{h}(\mathbf{x})) - i_2)\, d\mathbf{x}.$$

We write

$$|s(z_2,\mathbf{h}_1)-s(z_2,\mathbf{h}_2)| \le \int_{\mathbb{R}} g_\beta(z_2-i_2)\frac{N(i_2,\mathbf{h}_2)\,|p(i_2,\mathbf{h}_2)-p(i_2,\mathbf{h}_1)|}{p(i_2,\mathbf{h}_1)p(i_2,\mathbf{h}_2)}\,di_2 + \int_{\mathbb{R}} g_\beta(z_2-i_2)\frac{p(i_2,\mathbf{h}_2)\,|N(i_2,\mathbf{h}_2)-N(i_2,\mathbf{h}_1)|}{p(i_2,\mathbf{h}_1)p(i_2,\mathbf{h}_2)}\,di_2, \tag{4.4}$$

and consider the first term on the right-hand side:

$$p(i_2,\mathbf{h}_2)-p(i_2,\mathbf{h}_1)=\int_\Omega \left(g_\beta(i_2-\ I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x}))\,)-g_\beta(i_2-\ I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))\,)\right)\,d\mathbf{x}.$$

We use the first order Taylor expansion with integral remainder of the C^1 function g_β:

$$g_\beta(i+t)=g_\beta(i)+t\int_0^1 g'_\beta(i+t\alpha)\,d\alpha.$$

We can therefore write

$$\begin{aligned} &g_\beta(i_2-\ I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x})))-g_\beta(i_2-\ I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x})))\\ &=(I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))\ -\ I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x})))\int_0^1 g'_\beta\Big(i_2-\big(\alpha\,I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x}))\\ &\qquad\qquad +(1-\alpha)\,I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))\big)\Big)\,d\alpha. \end{aligned}$$

We use the fact that I_2 is Lipschitz continuous and write

$$\begin{aligned} &|p(i_2,\mathbf{h}_2)-p(i_2,\mathbf{h}_1)|\\ &\le Lip(I_2)\int_\Omega\Bigg(|\mathbf{h}_1(\mathbf{x})-\mathbf{h}_2(\mathbf{x})|\\ &\qquad\qquad\times\left|\int_0^1 g'_\beta(i_2-(\alpha\,I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x}))\ +(1-\alpha)\,I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))\,)\,d\alpha\right|\Bigg)\,d\mathbf{x}. \end{aligned}$$

The Schwarz inequality implies

$$\begin{aligned} &|p(i_2,\mathbf{h}_2)-p(i_2,\mathbf{h}_1)|\\ &\le Lip(I_2)\|\mathbf{h}_1-\mathbf{h}_2\|_H\Bigg(\int_\Omega\Bigg(\int_0^1 g'_\beta(i_2-(\alpha\,I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x}))\\ &\qquad\qquad+\ (1-\alpha)\,I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))\,)\,d\alpha\Bigg)^2\,d\mathbf{x}\Bigg)^{\frac12}. \end{aligned}$$

We introduce the function

$$\begin{aligned} c(i_2,\mathbf{h}_1,\mathbf{h}_2)=\Bigg(\int_\Omega\Bigg(\int_0^1\Big|i_2-\Big(\alpha\,I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x}))\ +(1-\alpha)\,I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))\Big)\Big|\\ \cdot\ g_\beta\Big(i_2-\big(\alpha\,I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x}))\ +(1-\alpha)\,I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))\,\big)\Big)\,d\alpha\Bigg)^2\,d\mathbf{x}\Bigg)^{\frac12} \end{aligned}$$

and notice that

$$\left(\int_\Omega \left(\int_0^1 g'_\beta(i_2 - (\alpha\, I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x})) + (1-\alpha)\, I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))))\, d\alpha\right)^2 d\mathbf{x}\right)^{\frac{1}{2}} \leq \frac{1}{\beta} c(i_2, \mathbf{h}_1, \mathbf{h}_2).$$

So far we have

$$\int_{\mathbb{R}} g_\beta(z_2 - i_2) \frac{N(i_2,\mathbf{h}_2)\,|p(i_2,\mathbf{h}_2) - p(i_2,\mathbf{h}_1)|}{p(i_2,\mathbf{h}_1)p(i_2,\mathbf{h}_2)}\, di_2 \leq \frac{Lip(I_2)}{\beta}\|\mathbf{h}_1 - \mathbf{h}_2\|_H \left(\int_{\mathbb{R}} g_\beta(z_2 - i_2) \frac{N(i_2,\mathbf{h}_2)\; c(i_2,\mathbf{h}_1,\mathbf{h}_2)}{p(i_2,\mathbf{h}_1)p(i_2,\mathbf{h}_2)}\, di_2\right). \tag{4.5}$$

We study the function of z_2 that is on the right-hand side of this inequality. First we note that the function is well defined since no problems occur when i_2 goes to infinity because "there are three Gaussians in the numerator and two in the denominator." We then show that this function is bounded independently of $\mathbf{h}_1$ and $\mathbf{h}_2$ for all $z_2 \in [0, \mathcal{A}]$.

As in the proof of Proposition 3.1, we consider three cases and detail only the first one: $i_2 \leq 0$. This is the case where

$$\begin{aligned} 0 \leq |i_2| &\leq |i_2 - I_2(\mathbf{x}+\mathbf{h}_j(\mathbf{x}))| \leq |i_2 - \mathcal{A}|, \quad j = 1,2, \\ 0 \leq |i_2| &\leq |i_2 - (\alpha I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x})) + (1-\alpha) I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x})))| \leq |i_2 - \mathcal{A}|, \quad 0 \leq \alpha \leq 1. \end{aligned}$$

Hence

$$\begin{aligned} g_\beta(i_2 - \mathcal{A}) &\leq g_\beta(i_2 - I_2(\mathbf{x}+\mathbf{h}_j(\mathbf{x}))) \leq g_\beta(i_2), \quad j = 1,2, \\ g_\beta(i_2 - \mathcal{A}) &\leq g_\beta(i_2 - (\alpha I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x})) + (1-\alpha) I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x})))) \leq g_\beta(i_2), \; 0 \leq \alpha \leq 1. \end{aligned}$$

This yields

$$\begin{aligned} &\int_{-\infty}^0 g_\beta(z_2 - i_2) \frac{N(i_2,\mathbf{h}_2)\; c(i_2,\mathbf{h}_1,\mathbf{h}_2)}{p(i_2,\mathbf{h}_1)p(i_2,\mathbf{h}_2)}\, di_2 \\ &\leq |\Omega|^{1/2}\mathcal{A} \int_{-\infty}^0 g_\beta(z_2 - i_2)|i_2 - \mathcal{A}| \left(\frac{g_\beta(i_2)}{g_\beta(i_2 - \mathcal{A})}\right)^2 di_2. \end{aligned}$$

The integral on the right-hand side is well defined and defines a continuous function of z_2.

The remaining two cases, $0 \leq i_2 \leq \mathcal{A}$ and $i_2 \geq \mathcal{A}$, are left to the reader. In all three cases, the functions of z_2 appearing on the right-hand side are continuous, independent of $\mathbf{h}_1$ and $\mathbf{h}_2$, and therefore upperbounded on $[0, \mathcal{A}]$ by a constant independent of $\mathbf{h}_1$ and $\mathbf{h}_2$. Returning to inequality (4.5), we have proved that there existed a positive constant C independent of z_2 such that

$$\int_{\mathbb{R}} g_\beta(z_2 - i_2) \frac{N(i_2,\mathbf{h}_2)\,|p(i_2,\mathbf{h}_2) - p(i_2,\mathbf{h}_1)|}{p(i_2,\mathbf{h}_1)p(i_2,\mathbf{h}_2)}\, di_2 \leq C\|\mathbf{h}_1 - \mathbf{h}_2\|_H \quad \forall z_2 \in [0,\mathcal{A}], \quad \forall \mathbf{h}_1,\, \mathbf{h}_2 \in H.$$

A similar proof can be developed for the second term on the right-hand side of inequality (4.4). In conclusion, we have proved that there exists a constant L_s, independent of z_2, such that

$$|s(z_2,\mathbf{h}_1) - s(z_2,\mathbf{h}_2)| \leq L_s\|\mathbf{h}_1 - \mathbf{h}_2\|_H \quad \forall z_2 \in [0,\mathcal{A}], \quad \forall \mathbf{h}_1,\, \mathbf{h}_2 \in H. \qquad \square$$

Thus, we can state the following.

PROPOSITION 4.6. *The function $s : \mathcal{H}_1 \to \mathbb{R}$ is Lipschitz continuous.*

Proof. The proof follows from Propositions 4.4 and 4.5 and Lemma 4.3. ∎

We now proceed with showing the same kind of properties for the function S.

PROPOSITION 4.7. *For all $\mathbf{h} \in H$, the function $[0, \mathcal{A}]^2 \to \mathbb{R}^+$ defined by $(z_1, z_2) \to S(z_1, z_2, \mathbf{h})$ is Lipschitz continuous with a Lipschitz constant l_S, which is independent of $\mathbf{h}$. Moreover, it is bounded by $\frac{\mathcal{A}}{\beta}$.*

Proof. The proof follows the same pattern as the proof of Proposition 4.4. ∎

PROPOSITION 4.8. *For all $(z_1, z_2) \in [0, \mathcal{A}]^2$, the function $\mathbf{h} \to S(z_1, z_2, \mathbf{h})$, $H \to \mathbb{R}$ is Lipschitz continuous with a Lipschitz constant L_S, which is independent of $(z_1, z_2) \in [0, \mathcal{A}]^2$.*

Proof. The proof follows the same pattern as that of Proposition 4.5. ∎

Therefore we can state the following result.

PROPOSITION 4.9. *The function $S : \mathcal{H}_2 \to \mathbb{R}$ is Lipschitz continuous.*

Proof. The proof follows those of Propositions 4.7 and 4.8 and Lemma 4.3. ∎

From Propositions 4.6, 4.9, and 4.1 we obtain the following.

COROLLARY 4.10. *The function $f_{\mathrm{MI}} : \mathcal{H}_2 \to \mathbb{R}$ defined by $(z_1, z_2, \mathbf{h}) \to s(z_2, \mathbf{h}) - S(z_1, z_2, \mathbf{h})$ is Lipschitz continuous and bounded by $2\frac{\mathcal{A}}{\beta}$. We denote by $Lip(f_{\mathrm{MI}})$ the corresponding Lipschitz constant.*

We can now state the main result of this section, as follows.

THEOREM 4.11. *The function $F_{\mathrm{MI}} : H \to H$, defined by*

$$\begin{aligned} F_{\mathrm{MI}}(\mathbf{h}) &= -f_{\mathrm{MI}}(I_1(\mathbf{Id}), I_2(\mathbf{Id}+\mathbf{h}), \mathbf{h})\nabla I_2(\mathbf{Id}+\mathbf{h}) \\ &= (S(I_1(\mathbf{Id}), I_2(\mathbf{Id}+\mathbf{h}), \mathbf{h}) - s(I_2(\mathbf{Id}+\mathbf{h}), \mathbf{h}))\nabla I_2(\mathbf{Id}+\mathbf{h}), \end{aligned}$$

is Lipschitz continuous and bounded.

Proof. Boundedness comes from Corollary 4.10 and the fact that $|\nabla I_2|$ is bounded. This implies that $F_{\mathrm{MI}}(\mathbf{h}) \in H = \mathbf{L}^2(\Omega)$ for all $\mathbf{h} \in H$.

We consider the ith component F^i_{MI} of F_{MI}:

$$F^i_{\mathrm{MI}}(\mathbf{h}_1)(\mathbf{x}) - F^i_{\mathrm{MI}}(\mathbf{h}_2)(\mathbf{x}) = U_1 T^i_1 - U_2 T^i_2, \qquad i = 1, \ldots, n,$$

with

$$\begin{aligned} U_j &= S(I_1(\mathbf{x}), I_2(\mathbf{x}+\mathbf{h}_j(\mathbf{x})), \mathbf{h}_j) - s(I_2(\mathbf{x}+\mathbf{h}_j(\mathbf{x})), \mathbf{h}_j), \\ T^i_j &= \partial_i I_2(\mathbf{x}+\mathbf{h}_j(\mathbf{x})), \quad i = 1, \ldots, n, \end{aligned}$$

and $j = 1, 2$. We continue with

$$|F^i_{\mathrm{MI}}(\mathbf{h}_1)(\mathbf{x}) - F^i_{\mathrm{MI}}(\mathbf{h}_2)(\mathbf{x})| \le |U_1 - U_2||T_1| + |U_2||T_1 - T_2|.$$

Because $\partial_i I_2$ is bounded, $|T^i_j| \le \|\partial_i I_2\|_\infty$. Because of Corollary 4.10, $|U_2| \le 2\frac{\mathcal{A}}{\beta}$. Because $\partial_i I_2$ is Lipschitz continuous, $|T_1 - T_2| \le Lip(\partial_i I_2)|\mathbf{h}_1(\mathbf{x}) - \mathbf{h}_2(\mathbf{x})|$.

Finally, because of Corollary 4.10 and the fact that I_2 is Lipschitz continuous,

$$|U_1 - U_2| \le Lip(f_{\mathrm{MI}})\left(Lip(I_2)|\mathbf{h}_1(\mathbf{x}) - \mathbf{h}_2(\mathbf{x})| + \|\mathbf{h}_1 - \mathbf{h}_2\|_H\right).$$

Collecting all terms, we obtain

$$|F^i_{\mathrm{MI}}(\mathbf{h}_1)(\mathbf{x}) - F^i_{\mathrm{MI}}(\mathbf{h}_2)(\mathbf{x})| \le C^i(|\mathbf{h}_1(\mathbf{x}) - \mathbf{h}_2(\mathbf{x})| + \|\mathbf{h}_1 - \mathbf{h}_2\|_H)$$

for some positive constant C^i, $i = 1, \ldots, n$. The last inequality yields, through the application of the Cauchy–Schwarz method,

$$\|F_{\mathrm{MI}}(\mathbf{h}_1) - F_{\mathrm{MI}}(\mathbf{h}_2)\|_H \leq L_F \|\mathbf{h}_1 - \mathbf{h}_2\|_H$$

for some positive constant L_F, and this completes the proof. □

The following proposition is needed in the proof of Proposition 2.4.

PROPOSITION 4.12. *The function* $\Omega \to \mathbb{R}^n$ *such that* $\mathbf{x} \to F_{\mathrm{MI}}(\mathbf{h})(\mathbf{x})$ *satisfies*

$$|F_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) - F_{\mathrm{MI}}(\mathbf{h})(\mathbf{y})| \leq K(|\mathbf{x} - \mathbf{y}| + |\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{y})|)$$

for some constant $K > 0$.

Proof. We write

$$\begin{aligned} F_{\mathrm{MI}}(\mathbf{h})(\mathbf{y}) - F_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) = {} & f_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) \nabla I_2(\mathbf{x} + \mathbf{h}(\mathbf{x})) - f_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) \nabla I_2(\mathbf{y} + \mathbf{h}(\mathbf{y})) \\ & + f_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) \nabla I_2(\mathbf{y} + \mathbf{h}(\mathbf{y})) - f_{\mathrm{MI}}(\mathbf{h})(\mathbf{y}) \nabla I_2(\mathbf{y} + \mathbf{h}(\mathbf{y})). \end{aligned}$$

Hence

$$\begin{aligned} |F_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) - F_{\mathrm{MI}}(\mathbf{h})(\mathbf{y})| \leq {} & + |f_{\mathrm{MI}}(\mathbf{h})(\mathbf{x})| \, |\nabla I_2(\mathbf{x} + \mathbf{h}(\mathbf{x})) - \nabla I_2(\mathbf{y} + \mathbf{h}(\mathbf{y}))| \\ & |\nabla I_2(\mathbf{y} + \mathbf{h}(\mathbf{y}))| \, |f_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) - f_{\mathrm{MI}}(\mathbf{h})(\mathbf{y})|. \end{aligned}$$

Corollary 4.10 and the fact that the functions I_1, I_2, and its first order derivative are Lipschitz continuous imply

$$\begin{aligned} & |F_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) - F_{\mathrm{MI}}(\mathbf{h})(\mathbf{y})| \\ & \leq 2 \frac{\mathcal{A}}{\beta} Lip(\nabla I_2)(|\mathbf{x} - \mathbf{y}| + |\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{y})|) + \| |\nabla I_2| \|_\infty Lip(f_{\mathrm{MI}}) |\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{y})|, \end{aligned}$$

and hence the result. □

4.2. Local cross-covariance. In this section we prove the Lipschitz-continuity of the mapping $H \to H$ defined by $F_{\mathrm{CC}}(\mathbf{h})$ (see (3.10)). The reasoning is analogous to that for the global case. We start with some estimates of the bounds of the means and variances of the relevant random variables.

LEMMA 4.13. *Let* $\mathrm{diam}(\Omega)$ *be the diameter of the open bounded set* Ω*:*$\mathrm{diam}(\Omega) = \sup_{\mathbf{x},\mathbf{y} \in \Omega} \|\mathbf{x} - \mathbf{y}\|$. *We denote by* $G_\gamma(\mathrm{diam}(\Omega))$ *the value* $\inf_{\mathbf{x},\mathbf{y} \in \Omega} G_\gamma(\mathbf{x} - \mathbf{y})$ *and define*

$$K_\Omega = \frac{G_\gamma(\mathbf{0})}{G_\gamma(\mathrm{diam}(\Omega))}. \tag{4.6}$$

We have $\frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \geq |\Omega| G_\gamma(\mathrm{diam}(\Omega)) > 0$ *and* $\int_\Omega G_\gamma(\mathbf{x} - \mathbf{x}_0) \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \, d\mathbf{x}_0 \leq K_\Omega$ *for all* $\mathbf{x} \in \Omega$.

Proof. Since $\mathcal{G}_\gamma(\mathbf{x}_0) = \int_\Omega G_\gamma(\mathbf{y} - \mathbf{x}_0) \, d\mathbf{y}$, we have $\mathcal{G}_\gamma(\mathbf{x}_0) \geq |\Omega| G_\gamma(\mathrm{diam}(\Omega))$. Therefore

$$\begin{aligned} \int_\Omega G_\gamma(\mathbf{x} - \mathbf{x}_0) \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \, d\mathbf{x}_0 & \leq \frac{1}{|\Omega| G_\gamma(\mathrm{diam}(\Omega))} \int_\Omega G_\gamma(\mathbf{x} - \mathbf{x}_0) d\mathbf{x}_0 \\ & \leq \frac{1}{|\Omega| G_\gamma(\mathrm{diam}(\Omega))} \times |\Omega| G_\gamma(\mathbf{0}) = K_\Omega. \end{aligned}$$

□

LEMMA 4.14. *For all* $\mathbf{x}_0 \in \Omega$, *the following inequalities are verified:*

$$0 \leq \mu_1(\mathbf{x}_0) \leq \mathcal{A} \quad \textit{and} \quad \beta \leq v_1(\mathbf{x}_0) \leq \beta + \mathcal{A}^2.$$

Proof. We have

$$\mu_1(\mathbf{x}_0) = \int_{\mathbb{R}^2} i_1 \left(\frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega G_\beta(\mathbf{I_h}(\mathbf{x}) - \mathbf{i})\, G_\gamma(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x} \right) di_1\, di_2 \tag{4.7}$$
$$= \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega I_1(\mathbf{x})\, G_\gamma(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x},$$

and the first inequalities follow from the fact that $I_1(\mathbf{x}) \in [0, \mathcal{A}]$. Similarly, for $v_1(\mathbf{x}_0)$, we have

$$v_1(\mathbf{x}_0) = \int_{\mathbb{R}^2} i_1^2 \left(\frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega G_\beta(\mathbf{I_h}(\mathbf{x}) - \mathbf{i})\, G_\gamma(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x} \right) di_1\, di_2 - \mu_1^2(\mathbf{x}_0) \tag{4.8}$$
$$= \beta + \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega I_1(\mathbf{x})^2\, G_\gamma(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x} - \left(\frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega I_1(\mathbf{x})\, G_\gamma(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x} \right)^2,$$

from which the right-hand side of the second inequality follows. The application of the Cauchy–Schwarz inequality to

$$\int_\Omega I_1(\mathbf{x})\, G_\gamma(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x} = \int_\Omega \left(I_1(\mathbf{x}) \sqrt{G_\gamma(\mathbf{x} - \mathbf{x}_0)} \right) \sqrt{G_\gamma(\mathbf{x} - \mathbf{x}_0)}\, d\mathbf{x}$$

yields the left-hand side. ∎

We then characterize the mean $\mu_2(\mathbf{x}_0, \mathbf{h})$; see Table 3.1.

LEMMA 4.15. *The function $\Omega \times H \longrightarrow \mathbb{R}$ defined by $(\mathbf{x}_0, \mathbf{h}) \longrightarrow \mu_2(\mathbf{x}_0, \mathbf{h})$ is bounded and Lipschitz continuous in H uniformly in Ω.*

Proof. According to the definition of μ_2 in Table 3.1 we have

$$\mu_2(\mathbf{x}_0, \mathbf{h}) = \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega I_2(\mathbf{x} + \mathbf{h}(\mathbf{x})) G_\gamma(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x}.$$

We have immediately

$$\mu_2(\mathbf{x}_0, \mathbf{h}) \leq \mathcal{A}.$$

Moreover,

$$|\mu_2(\mathbf{x}_0, \mathbf{h}_1) - \mu_2(\mathbf{x}_0, \mathbf{h}_2)| \leq \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega |I_2(\mathbf{x} + \mathbf{h}_1(\mathbf{x})) - I_2(\mathbf{x} + \mathbf{h}_2(\mathbf{x}))| G_\gamma(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x}.$$

Since I_2 is Lipschitz, a combination of Lemma 4.13 and the Cauchy–Schwarz inequality yields

$$|\mu_2(\mathbf{x}_0, \mathbf{h}_1) - \mu_2(\mathbf{x}_0, \mathbf{h}_2)| \leq Lip(I_2) K_\Omega \|\mathbf{h}_1 - \mathbf{h}_2\|_H. \qquad \blacksquare$$

We have similar properties for the variance $v_2(\mathbf{x}_0, \mathbf{h})$; see Table 3.1.

LEMMA 4.16. *The function $\Omega \times H \longrightarrow \mathbb{R}$ defined by $(\mathbf{x}_0, \mathbf{h}) \longrightarrow v_2(\mathbf{x}_0, \mathbf{h})$ is strictly positive, bounded, and Lipschitz continuous in H uniformly in Ω.*

Proof. According to the definition of v_2 in Table 3.1 we have

$$v_2(\mathbf{h}, \mathbf{x}_0) = \int_{\mathbb{R}^2} i_2^2 \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \left(\int_\Omega g_\beta(I_2(\mathbf{x} + \mathbf{h}(\mathbf{x})) - i_2)\, G_\gamma(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x} \right) di_2 - \mu_2(\mathbf{x}_0, \mathbf{h})^2.$$

Hence

$$v_2(\mathbf{h},\mathbf{x}_0) = \beta + \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))^2 \, G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} - \mu_2(\mathbf{h},\mathbf{x}_0)^2.$$

Therefore $v_2(\mathbf{h},\mathbf{x}_0) \leq \beta + \mathcal{A}^2$. To prove that it is strictly positive, we write

$$\begin{aligned} v_2(\mathbf{h},\mathbf{x}_0) = \beta &+ \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))^2 \, G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} \\ &- \left(\frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})) G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} \right)^2, \end{aligned}$$

and the same reasoning as in Lemma 4.14 shows that $\beta \leq v_2(\mathbf{h},\mathbf{x}_0)$. For the second part, since $\mu_2(\mathbf{h},\mathbf{x}_0)$ is bounded and Lipschitz continuous uniformly in Ω (Lemma 4.15), it suffices to show the Lipschitz continuity of the first term on the right-hand side. For this term we have, using again a combination of Lemma 4.13 and the Cauchy–Schwarz inequality,

$$\begin{aligned} &\frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \left| \int_\Omega I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))^2 \, G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} - \int_\Omega I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x}))^2 \, G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} \right| \\ &\leq \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega (I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))+I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x}))) \, |I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x}))-I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x}))| \, G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} \\ &\leq 2\mathcal{A} \, Lip(I_2) \, K_\Omega \|\mathbf{h}_1 - \mathbf{h}_2\|_H. \qquad \square \end{aligned}$$

We now show the boundedness and Lipschitz continuity of the correlation $v_{1,2}(\mathbf{x}_0,\mathbf{h})$; see Table 3.1.

PROPOSITION 4.17. *The function $H \times \Omega \to \mathbb{R}$ defined by $(\mathbf{h},\mathbf{x}_0) \mapsto v_{1,2}(\mathbf{h},\mathbf{x}_0)$ is bounded and Lipschitz continuous in H, uniformly in Ω.*

Proof. According to the definition of $v_{1,2}$ in Table 3.1 we have

$$\begin{aligned} v_{1,2}(\mathbf{h},\mathbf{x}_0) = \int_{\mathbb{R}^2} i_1 \, i_2 \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \left(\int_\Omega G_\beta(\mathbf{I}_\mathbf{h}(\mathbf{x}) - \mathbf{i}) \, G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} \right) \, di_1 \, di_2 \\ - \mu_1(\mathbf{x}_0)\mu_2(\mathbf{x}_0,\mathbf{h}). \end{aligned}$$

Hence

$$v_{1,2}(\mathbf{h},\mathbf{x}_0) = \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega I_1(\mathbf{x}) \, I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})) \, G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} - \mu_1(\mathbf{x}_0) \, \mu_2(\mathbf{h},\mathbf{x}_0).$$

Thus we have, for all $\mathbf{x}_0 \in \Omega$, $|v_{1,2}(\mathbf{h},\mathbf{x}_0)| \leq 2\mathcal{A}^2$, which proves the first part of the proposition. For the second part, since $\mu_2(\mathbf{h},\mathbf{x}_0)$ is Lipschitz continuous uniformly in Ω (Lemma 4.15) and μ_1 is bounded, it suffices to show the Lipschitz continuity of the first term on the right-hand side. For this term we have, using again a combination of Lemma 4.13 and the Cauchy–Schwarz inequality,

$$\begin{aligned} &\frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \left| \int_\Omega I_1(\mathbf{x}) I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x})) \, G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} - \int_\Omega I_1(\mathbf{x}) \, I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x})) \, G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} \right| \\ &\leq \frac{1}{\mathcal{G}_\gamma(\mathbf{x}_0)} \int_\Omega |I_1(\mathbf{x})| \, |I_2(\mathbf{x}+\mathbf{h}_1(\mathbf{x})) - I_2(\mathbf{x}+\mathbf{h}_2(\mathbf{x}))| \, G_\gamma(\mathbf{x}-\mathbf{x}_0) \, d\mathbf{x} \\ &\leq \mathcal{A} \, Lip(I_2) \, K_\Omega \|\mathbf{h}_1 - \mathbf{h}_2\|_H. \qquad \square \end{aligned}$$

THEOREM 4.18. *The function $H \times \Omega \to \mathbb{R}$ defined by $(\mathbf{h}, \mathbf{x}) \mapsto \mathcal{J}_{\mathrm{CC}}(\mathbf{h}, \mathbf{x})$ is bounded and Lipschitz continuous in H, uniformly in Ω.*

Proof. Indeed, by definition we have $-1 \leq \mathcal{J}_{\mathrm{CC}}(\mathbf{h}, \mathbf{x}) \leq 0$ for all $\mathbf{h} \in H$. Moreover, we have

$$\mathcal{J}_{\mathrm{CC}}(\mathbf{h}, \mathbf{x}) = -\frac{v_{1,2}(\mathbf{h}, \mathbf{x})^2}{v_1(\mathbf{x})\, v_2(\mathbf{h}, \mathbf{x})}, \tag{4.9}$$

and the following properties hold true:

- $v_{1,2}(\mathbf{h}, \mathbf{x})$ is bounded and Lipschitz-continuous in H uniformly in Ω (Proposition 4.17).
- $v_2(\mathbf{h}, \mathbf{x})$ is strictly positive, bounded, and Lipschitz-continuous in H, uniformly in Ω (Lemma 4.16).
- $v_1(\mathbf{x})$ is bounded and > 0 (Lemma 4.14).

We may therefore apply Proposition 4.1. ∎

THEOREM 4.19. *The function $\mathcal{H}_2 \times \Omega \to \mathbb{R}$ defined by*

$$(\mathbf{z}, \mathbf{h}, \mathbf{x}) \mapsto L_{\mathrm{CC}}(\mathbf{z}, \mathbf{h}, \mathbf{x})$$

is bounded and Lipschitz continuous in $\mathcal{H}_2$, uniformly in Ω.

Proof. We have

$$L_{\mathrm{CC}}(\mathbf{z}, \mathbf{h}, \mathbf{x}) = -\frac{2}{\mathcal{G}_\gamma(\mathbf{x})} \left[\frac{v_{1,2}(\mathbf{h}, \mathbf{x})}{v_2(\mathbf{h}, \mathbf{x})} \left(\frac{z_1 - \mu_1(\mathbf{x})}{v_1(\mathbf{x})} \right) + \mathcal{J}_{\mathrm{CC}}(\mathbf{h}, \mathbf{x}) \left(\frac{z_2 - \mu_2(\mathbf{h}, \mathbf{x})}{v_2(\mathbf{h}, \mathbf{x})} \right) \right].$$

This can be rewritten as

$$L_{\mathrm{CC}}(\mathbf{z}, \mathbf{h}, \mathbf{x}) = f_1(\mathbf{h}, \mathbf{x})\, z_1 + f_2(\mathbf{h}, \mathbf{x})\, z_2 + f_3(\mathbf{h}, \mathbf{x}).$$

The functions $H \times \Omega \to \mathbb{R}$, f_1, f_2, and f_3, are bounded and Lipschitz continuous in H uniformly in Ω because of Lemmas 4.13, 4.14, 4.15, 4.16, Proposition 4.17, and Theorem 4.18. The result readily follows. ∎

THEOREM 4.20. *The function $f_{\mathrm{CC}} : \mathcal{H}_2 \times \Omega \to \mathbb{R}$, defined by*

$$f_{\mathrm{CC}}(\mathbf{z}, \mathbf{h}, \mathbf{x}) = G_\gamma \star L_{\mathrm{CC}}(\mathbf{z}, \mathbf{h}, \mathbf{x}),$$

is bounded and Lipschitz continuous in $\mathcal{H}_2$ uniformly in Ω.

Proof. The fact that it is bounded follows from Theorem 4.19. To obtain the Lipschitz continuity, we write

$$\begin{aligned} &|f_{\mathrm{CC}}(\mathbf{z}, \mathbf{h}, \mathbf{x}) - f_{\mathrm{CC}}(\mathbf{z}', \mathbf{h}', \mathbf{x})| \\ &= \left| \int_\Omega G_\gamma(\mathbf{x} - \mathbf{x}_0) \left(L_{\mathrm{CC}}(\mathbf{z}, \mathbf{h}, \mathbf{x}_0) - L_{\mathrm{CC}}(\mathbf{z}', \mathbf{h}', \mathbf{x}_0) \right) d\mathbf{x}_0 \right| \\ &\leq G_\gamma(0)\, |\Omega|\, Lip(L_{\mathrm{CC}}) \left(|\mathbf{z} - \mathbf{z}'| + \|\mathbf{h} - \mathbf{h}'\|_H \right) \end{aligned}$$

and

$$\begin{aligned} &|G_\gamma \star f(\mathbf{z}, \mathbf{h}, \mathbf{x}) - G_\gamma \star f(\mathbf{z}, \mathbf{h}, \mathbf{y})| \\ &\qquad \leq \int_\Omega |G_\gamma(\mathbf{x} - \mathbf{x}_0) - G_\gamma(\mathbf{y} - \mathbf{x}_0)||f(\mathbf{z}, \mathbf{h}, \mathbf{x}_0)|\, d\mathbf{x}_0 \leq B_f\, Lip(G_\gamma) |\Omega|\, |\mathbf{x} - \mathbf{y}|, \end{aligned}$$

and hence obtain the result. ∎

THEOREM 4.21. *The function* $F_{\mathrm{CC}} : H \to H$ *defined by*

$$F_{\mathrm{CC}}(\mathbf{h}) = -f_{\mathrm{CC}}(I_1(\mathbf{Id}), I_2(\mathbf{Id}+\mathbf{h}), \mathbf{h}, \mathbf{Id})\nabla I_2(\mathbf{Id}+\mathbf{h})$$

is Lipschitz continuous and bounded.

Proof. Since f_{CC} is bounded (Theorem 4.20) as well as ∇I_2, so is F_{CC}, which therefore belongs to H. For the Lipschitz continuity, we write

$$\begin{aligned} F_{\mathrm{CC}}(\mathbf{h}_1) - F_{\mathrm{CC}}(\mathbf{h}_2) &= f_{\mathrm{CC}}(I_1(\mathbf{Id}), I_2(\mathbf{Id}+\mathbf{h}_1), \mathbf{h}_1, \mathbf{Id})\,(\nabla I_2(\mathbf{Id}+\mathbf{h}_2) - \nabla I_2(\mathbf{Id}+\mathbf{h}_1)) \\ &+ \nabla I_2(\mathbf{Id}+\mathbf{h}_2)\,(f_{\mathrm{CC}}(I_1(\mathbf{Id}), I_2(\mathbf{Id}+\mathbf{h}_2), \mathbf{h}_2, \mathbf{Id}) - f_{\mathrm{CC}}(I_1(\mathbf{Id}), I_2(\mathbf{Id}+\mathbf{h}_1), \mathbf{h}_1, \mathbf{Id}))\,. \end{aligned}$$

The Lipschitz continuity follows from that of ∇I_2, of L_{CC} (Theorem 4.20), and the use of the Cauchy–Schwarz inequality as in the proof of Theorem 4.11. □

PROPOSITION 4.22. *The function* $\Omega \to \mathbb{R}^n$ *such that* $\mathbf{x} \to F_{\mathrm{CC}}(\mathbf{h})(\mathbf{x})$ *satisfies*

$$|F_{\mathrm{CC}}(\mathbf{h})(\mathbf{x}) - F_{\mathrm{CC}}(\mathbf{h})(\mathbf{y})| \leq K(|\mathbf{x}-\mathbf{y}| + |\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{y})|),$$

for some constant $K > 0$.

Proof. The proof is similar to that of Proposition 4.12; it follows from Theorem 4.20 and the fact that the functions I_1, I_2 and all the derivatives of I_2 are Lipschitz continuous. □

5. Numerical implementation. The numerical implementation of the previously described continuous matching flows involves estimating the matching term, which depends on one of the two intensity-comparison functions F_{MI} and F_{CC} and the regularization operator, which is $\mathbf{div}(\mathbf{T}_{I_1} D\mathbf{h})$. For the discretization in time, we adopt an explicit forward scheme. Implicit schemes are difficult to devise due to the high nonlinearity of the matching functions. The way in which the time step is chosen is discussed in the following section. Alvarez, Weickert, and Sánchez [2] propose a very efficient scheme for discretizing the Nagel–Enkelmann divergence operator which we adopt in our experiments. We use a schematic notation for the description of the finite-difference schemes. For instance, let us denote by $h_p^{i,j}$ and $L_p^{i,j}$ the components $(p = 1, 2)$ of $\mathbf{h}$ and its Laplacian $\Delta\mathbf{h}$ at the grid point (i, j) in the discrete image domain. The voxel size in all directions is assumed to be equal to one. A possible scheme for $\alpha\Delta\mathbf{h}$ is

$$L_p^{i,j} = \alpha\Big(h_p^{i+1,j} + h_p^{i-1,j} + h_p^{i,j-1} + h_p^{i,j+1} - 4\,h_p^{i,j}\Big), \tag{5.1}$$

which we write schematically as

$$L_p^{i,j} = \alpha * \begin{array}{|c|c|c|} \hline & 1 & \\ \hline 1 & -4 & 1 \\ \hline & 1 & \\ \hline \end{array}\ h_p.$$

In this notation, the tables represent the discrete grid and contain the weights associated with each pixel (zero if none). The function to which the grid corresponds is written at the bottom. The scheme proposed in [2] is the following. Let $A = \mathbf{div}(\mathbf{T}_{I_1} D\mathbf{h})$, where

$$\mathbf{T}_{I_1} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

Then, for $p = 1, 2$,

$$A_p^{i,j} = \tfrac{1}{2} * \underset{a}{\begin{array}{|c|c|c|}\hline & & \\ \hline & 1 & 1 \\ \hline & & \\ \hline\end{array}} * \underset{h_p}{\begin{array}{|c|c|c|}\hline & & \\ \hline & -1 & 1 \\ \hline & & \\ \hline\end{array}} + \tfrac{1}{2} * \underset{a}{\begin{array}{|c|c|c|}\hline & & \\ \hline 1 & 1 & \\ \hline & & \\ \hline\end{array}} * \underset{h_p}{\begin{array}{|c|c|c|}\hline & & \\ \hline 1 & -1 & \\ \hline & & \\ \hline\end{array}}$$

$$+ \tfrac{1}{2} * \underset{c}{\begin{array}{|c|c|c|}\hline & 1 & \\ \hline & 1 & \\ \hline & & \\ \hline\end{array}} * \underset{h_p}{\begin{array}{|c|c|c|}\hline & 1 & \\ \hline & -1 & \\ \hline & & \\ \hline\end{array}} + \tfrac{1}{2} * \underset{c}{\begin{array}{|c|c|c|}\hline & & \\ \hline & 1 & \\ \hline & 1 & \\ \hline\end{array}} * \underset{h_p}{\begin{array}{|c|c|c|}\hline & & \\ \hline & -1 & \\ \hline & 1 & \\ \hline\end{array}}$$

$$+ \tfrac{1}{4} * \underset{b}{\begin{array}{|c|c|c|}\hline & & \\ \hline & 1 & \\ \hline & & 1 \\ \hline\end{array}} * \underset{h_p}{\begin{array}{|c|c|c|}\hline & & \\ \hline & -1 & \\ \hline & & 1 \\ \hline\end{array}} + \tfrac{1}{4} * \underset{b}{\begin{array}{|c|c|c|}\hline 1 & & \\ \hline & 1 & \\ \hline & & \\ \hline\end{array}} * \underset{h_p}{\begin{array}{|c|c|c|}\hline 1 & & \\ \hline & -1 & \\ \hline & & \\ \hline\end{array}}$$

$$- \tfrac{1}{4} * \underset{b}{\begin{array}{|c|c|c|}\hline & & 1 \\ \hline & 1 & \\ \hline & & \\ \hline\end{array}} * \underset{h_p}{\begin{array}{|c|c|c|}\hline & & 1 \\ \hline & -1 & \\ \hline & & \\ \hline\end{array}} - \tfrac{1}{4} * \underset{b}{\begin{array}{|c|c|c|}\hline & & \\ \hline & 1 & \\ \hline 1 & & \\ \hline\end{array}} * \underset{h_p}{\begin{array}{|c|c|c|}\hline & & \\ \hline & -1 & \\ \hline 1 & & \\ \hline\end{array}} .$$

The 3*D case.* This scheme generalizes readily to the 3D case. In order to write the 3D scheme explicitly in a compact way, we take advantage of its very simple form and introduce a more compact notation. We write

$$S_{\frac{1}{2}}(a, x^+) \equiv \frac{1}{2} * \underset{a}{\begin{array}{|c|c|c|}\hline & & \\ \hline & 1 & 1 \\ \hline & & \\ \hline\end{array}} * \underset{h_p}{\begin{array}{|c|c|c|}\hline & & \\ \hline & -1 & 1 \\ \hline & & \\ \hline\end{array}},$$

where x^+ indicates the direction defined by the voxels with non-null weights, starting at the center. With this notation, we write the 2D Nagel–Enkelmann operator above as

$$\begin{aligned} A_p^{i,j} = {} & S_{\frac{1}{2}}(a, x^+) + S_{\frac{1}{2}}(a, x^-) + S_{\frac{1}{2}}(c, y^+) + S_{\frac{1}{2}}(c, y^-) \\ & + S_{\frac{1}{4}}(b, x^+y^+) + S_{\frac{1}{4}}(b, x^-y^-) - S_{\frac{1}{4}}(b, x^+y^-) - S_{\frac{1}{4}}(b, x^-y^+). \end{aligned}$$

In the 3D case, we have

$$\mathbf{T}_{I_1} = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix},$$

and the corresponding scheme is, for $p = 1, 2, 3$,

$$\begin{aligned} A_p^{i,j,k} = {} & S_{\frac{1}{2}}(a, x^+) + S_{\frac{1}{2}}(a, x^-) + S_{\frac{1}{2}}(d, y^+) + S_{\frac{1}{2}}(d, y^-) + S_{\frac{1}{2}}(f, z^+) \\ & + S_{\frac{1}{2}}(f, z^-) + S_{\frac{1}{4}}(b, x^+y^+) + S_{\frac{1}{4}}(b, x^-y^-) - S_{\frac{1}{4}}(b, x^+y^-) - S_{\frac{1}{4}}(b, x^-y^+) \\ & + S_{\frac{1}{4}}(c, x^+z^+) + S_{\frac{1}{4}}(c, x^-z^-) - S_{\frac{1}{4}}(c, x^+z^-) - S_{\frac{1}{4}}(c, x^-z^+) \\ & + S_{\frac{1}{4}}(e, y^+z^+) + S_{\frac{1}{4}}(e, y^-z^-) - S_{\frac{1}{4}}(e, y^+z^-) - S_{\frac{1}{4}}(e, y^-z^+). \end{aligned}$$

5.1. Intensity comparison functions. Concerning the intensity comparison functions, we approximate convolutions with a Gaussian kernel by recursive filtering using the smoothing operator introduced in [17]. Terms of the form $\nabla I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))$ and $I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))$ are calculated by trilinear interpolation. The global function F_{MI} is estimated by explicitly computing the global density estimate $P(\mathbf{i},\mathbf{h})$ through recursive smoothing of the discrete joint histogram of intensities, as detailed in section 5.1.2. For the local function F_{CC}, a special implementation has been developed, as detailed in section 5.1.4.

5.1.1. Convolutions. The convolutions by a Gaussian kernel are approximated by recursive filtering using the smoothing operator introduced by Deriche [17]. Given a discrete 1D input sequence $x(n), n = 1,\dots,M$, its convolution by the smoothing operator $S_\alpha(n) = k\,(\alpha|n|+1)\,e^{-\alpha|n|}$ is calculated efficiently as (see [17])

$$y(n) = (S_\alpha \star x)(n) = y_1(n) + y_2(n),$$

where

$$\begin{cases} y_1(n) \;=\; k\big(x(n) \;+\; e^{-\alpha}\,(\alpha-1)\,x(n-1)\big) \\ \qquad\qquad\qquad\qquad + 2\,e^{-\alpha}\,y_1(n-1) \;-\; e^{-2\alpha}\,y_1(n-2), \\ \\ y_2(n) \;=\; k\big(e^{-\alpha}\,(\alpha+1)\,x(n+1) \;-\; e^{-2\alpha}\,x(n+2)\big) \\ \qquad\qquad\qquad\qquad + 2\,e^{-\alpha}\,y_2(n+1) \;-\; e^{-2\alpha}\,y_2(n+2). \end{cases}$$

The normalization constant k is chosen by requiring that $\int_{\mathbb{R}} S_\alpha(t)\;dt = 1$, which yields $k = \alpha/4$. This scheme is very efficient since the number of operations required is independent of the smoothing parameter α. The smoothing filter can be readily generalized to n dimensions by defining the separable filter $T_\alpha(\mathbf{x}) = \prod_{i=1}^{n} S_\alpha(x_i)$.

5.1.2. Density estimation. Parzen density estimates are obtained by smoothing the discrete joint histogram of intensities. We define the piecewise constant function $\mathbf{v} : \Omega \to [0,N]^2 \subset \mathbb{N}^2$ by quantification of $\mathbf{I_h}(\mathbf{x})$ into $N+1$ intensity levels (bins):

$$\mathbf{v}(\mathbf{x}) = \begin{pmatrix} \lfloor \zeta I_1(\mathbf{x}) \rfloor \\ \\ \lfloor \zeta I_2(\mathbf{x}+\mathbf{h}(\mathbf{x})) \rfloor \end{pmatrix} = \begin{cases} (0,0)^T & \text{on} \quad \Omega_{0,0}, \\ \vdots & \\ (N,N)^T & \text{on} \quad \Omega_{N,N}, \end{cases}$$

where $\zeta = N/\mathcal{A}$, $\lfloor\cdot\rfloor$ denotes the floor operator in $\mathbb{R}^+$, i.e., the function $\mathbb{R}^+ \to \mathbb{N}$ such that $\lfloor x \rfloor = \max\{n \in \mathbb{N} : n \le x\}$ and $\{\Omega_{k,l}\}_{(k,l)\in[0,N]^2}$ is a partition of Ω. We then compute, setting $\beta' = \zeta^2\beta$,

$$\begin{aligned} P(\mathbf{i},\mathbf{h}) &= \frac{1}{|\Omega|}\int_\Omega G_\beta(\mathbf{I_h}(\mathbf{x}) - \mathbf{i})\;d\mathbf{x} = \frac{\zeta}{|\Omega|}\int_\Omega G_{\beta'}\big(\zeta\;(\mathbf{I_h}(\mathbf{x}) - \mathbf{i})\big)\;d\mathbf{x} \\ &\simeq \frac{\zeta}{|\Omega|}\int_\Omega G_{\beta'}\big(\mathbf{v}(\mathbf{x}) - \zeta\mathbf{i}\big)\;d\mathbf{x} = \frac{\zeta}{|\Omega|}\;\sum_{k=0}^{N}\sum_{l=0}^{N}\;\int_{\Omega_{k,l}} G_{\beta'}(k-\zeta i_1, l-\zeta i_2)\;d\mathbf{x} \\ &= \zeta\sum_{k=0}^{N}\sum_{l=0}^{N}\;\underbrace{|\Omega_{k,l}|/|\Omega|}_{H(k,l)}\;G_{\beta'}(k-\zeta i_1, l-\zeta i_2) = \zeta\;\big(H \star G_{\beta'}\big)(\zeta\mathbf{i}), \end{aligned}$$

H being the discrete joint histogram. The convolution is performed by recursive filtering, as described in the previous section. Note that this way of computing P is quite efficient since only one pass on the images is required, followed by the convolution.

5.1.3. Implementation of the computation of $\boldsymbol{F}_{\mathrm{MI}}$. The global mutual information function is then estimated as follows:

- Estimate $P(\mathbf{i},\mathbf{h})$ and its marginals.
- Estimate $L_{\mathrm{MI}}(\mathbf{i},\mathbf{h})$ (equation (3.4)) using centered finite-differences for the derivatives.
- Estimate $f_{\mathrm{MI}}(\mathbf{i},\mathbf{h}) = G_\beta \star L_{\mathrm{MI}}(\mathbf{i},\mathbf{h})$ (equation (3.3)) by recursive smoothing.
- Estimate $F_{\mathrm{MI}}(\mathbf{h})(\mathbf{x}) = -f_{\mathrm{MI}}(\mathbf{I}_{\mathbf{h}}(\mathbf{x}),\mathbf{h})\,\nabla I_2(\mathbf{x}+\mathbf{h}(\mathbf{x}))$ (equation (3.8)).

5.1.4. Implementation of the computation of $\boldsymbol{F}_{\mathrm{CC}}$. The function f_{CC} is estimated as

$$L_{\mathrm{CC}}(\mathbf{i},\mathbf{x}) = (G_\gamma \star f_1)(\mathbf{x})\; i_1 + (G_\gamma \star f_2)(\mathbf{x})\; i_2 + (G_\gamma \star f_3)(\mathbf{x}),$$

where

$$f_1(\mathbf{x}) = -2\; v_{1,2}(\mathbf{h},\mathbf{x})/(\mathcal{G}_\gamma(\mathbf{x})\; v_1(\mathbf{x})\; v_2(\mathbf{h},\mathbf{x})),$$

$$f_2(\mathbf{x}) = -2\; \mathcal{J}_{\mathrm{CC}}(\mathbf{h},\mathbf{x})/(\mathcal{G}_\gamma(\mathbf{x})\; v_2(\mathbf{h},\mathbf{x})),$$

and

$$f_3(\mathbf{x}) = -\big(f_1(\mathbf{x})\; \mu_1(\mathbf{x}) + f_2(\mathbf{x})\; \mu_2(\mathbf{h},\mathbf{x})\big).$$

All the required space dependent quantities like $\mu_1(\mathbf{x})$ are computed through recursive spatial smoothing. This algorithm is similar to the one proposed in [11].

5.2. Parameters. We now discuss the way in which the different parameters of the algorithms are determined.

- γ: This is the variance of the spatial Gaussian for local density estimates. Its value does not affect the computation time since the local statistics are calculated using the recursive smoothing filter. Thanks to this, we have conducted some experiments with different values of this parameter, which have shown that the algorithms are not very sensitive to it. Qualitatively speaking, the local window has to be large enough for the statistics to be significant and small enough to account for nonstationarities of the density. We set γ to 5 in all our experiments.
- β: This is the variance of the Gaussian for the Parzen estimates. Unlike γ, determining a good value for β is important for obtaining good results. In our case, it is determined automatically as follows (we refer to [6] for a recent comprehensive study on nonparametric density estimation). We adopt a cross-validation technique based on an empirical maximum likelihood method. We denote by $\{\mathbf{i}_k\}$ a set of m intensity pair samples ($k = 1,\ldots,m$) and take the value of β which maximizes the empirical likelihood:

$$L(\beta) = \prod_{k=1}^{m} \hat{P}_{\beta,k}(\mathbf{i}_k),$$

where

$$\hat{P}_{\beta,k}(\mathbf{i}_k) = \frac{1}{m-n_k} \sum_{\{s:\ \mathbf{i}_s \neq \mathbf{i}_k\}} G_\beta(\mathbf{i}_k - \mathbf{i}_s)$$

and n_k is the number of data samples for which $\mathbf{i}_s = \mathbf{i}_k$.
For the experiments shown in this paper, the optimal value of β varies in the range 15–20.

- κ: This parameter determines the coefficient of the regularization term in the energy functional. Since the range of the different matching functions F varies considerably, we normalize it by dividing by the maximum value $\kappa_0 = \|F(\mathbf{h}_0)\|_\infty$, $\mathbf{h}_0$ being the initial field. This is equivalent to replacing κ by C such that

$$C = \kappa\,\kappa_0.$$

  The behavior of the algorithms is much more stable with respect to C than to κ. κ has been kept to the fixed value of 10 in all our experiments except the one with the Eiffel tower, where the expected displacement is close to a translation, a very smooth transformation indeed. In this experiment the value of κ has been fixed to 100.
- Δt: The time step is chosen such that the coefficient of the regularization term (i.e., $C\Delta t$) is less than a specified value. It is known, for instance, that the scheme (5.1) is stable for values of $C\Delta t$ no larger than 0.25. In our experiments, we fix $C\Delta t$ to a value of 0.1. The resulting Δt is different at each resolution due to the fact that C is recomputed at the beginning of each level. In some of the examples, we have computed the value of Δt by doing a line-search in the gradient direction, looking for the minimum of the criterion. At each resolution, the line-search consistently produces values which are asymptotically (a) almost constant for that particular resolution and (b) close to a value such that $C\Delta t$ is approximately 0.15–0.2. This is true only asymptotically for each resolution. The first optimal steps are much larger, and this has suggested to us to use the line-search in conjunction with more sophisticated minimization algorithms such as the conjugate gradient. We discuss this point in one of the examples of section 6.
- σ: This is the scale parameter. We adopt a multiresolution approach (see the next section), smoothing the images at each stage by a small amount. Within each stage of the multiresolution pyramid, the parameter σ is fixed to a small value, typically 0.5 voxels/pixels.

One extra parameter is needed for the regularization operator.

- λ: This is the parameter controlling the anisotropic behavior of the Nagel–Enkelmann tensor. We adopt the method proposed by Alvarez, Weickert, and Sánchez [2]. Given q, which in practice is fixed to 0.1, we take the value of λ such that

$$q = \int_0^\lambda \mathcal{H}_{|\nabla I_1|}(z)\,dz,$$

  where $\mathcal{H}_{|\nabla I_1|}(z)$ is the normalized histogram of $|\nabla I_1|$.

6. Experimental results. We present results of experiments using the previously described algorithms. To recover large deformations, we use a multiresolution approach by applying the gradient descent to a set of smoothed and subsampled images. Since the functionals considered are not convex, this coarse-to-fine strategy helps avoid irrelevant extrema while reducing the computational cost of the algorithms. The first stage of the multiresolution pyramid is initialized with $\mathbf{h}_0 = 0$. The field resulting from the computation at a coarse level is multiplied by two and resampled at the

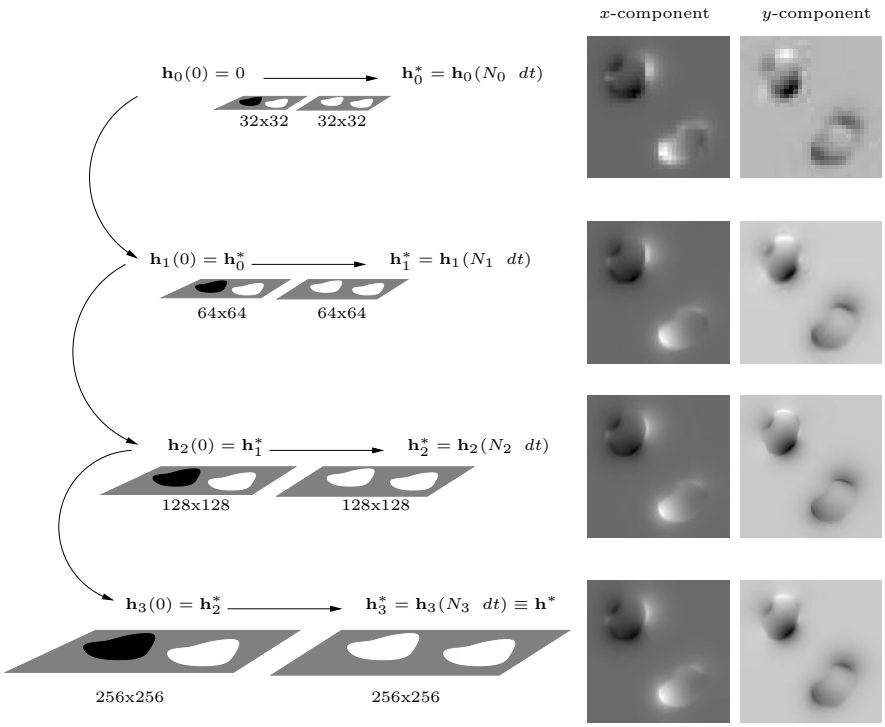


FIG. 3. *Multiresolution strategy: the first stage of the multiresolution pyramid is initialized with* $\mathbf{h}_0 = 0$*, and the resulting field from a coarse level is scaled by two and resampled at the next finer scale (by (tri)-linear interpolation) at each step of the pyramid descent where it is used as the initial value of the corresponding evolution equation. The left column shows the pyramid; the right shows the computed deformation field at each level. This example illustrates the result of the experiment in Figure* 4.

finer scale (by (tri)-linear interpolation) at each step of the pyramid descent, where it is used as the initial value of the corresponding evolution equation. The number of iterations at each level is automatically determined based on the reduction of the criterion. Iterations at each level are stopped when the criterion changes by an amount less than a specified threshold. The number of resolution levels in the pyramid is chosen so that at the coarsest level the image has approximately n pixels (voxels) in its smallest dimension, n being manually fixed to a small value such as 8, 16, or 32. (see Figure 3).

Experiment with synthetic data (Figure 4). In this example, we use the mutual information criterion to compute the displacement field between two synthetic images. Despite its simplicity, this example illustrates the main difficulties of the problem we consider: the distortion is nonrigid, and we are unable to directly compare intensities. Besides, this particular example is especially difficult since the discrete histogram (i.e., before Gaussian smoothing) contains only six nonzero entries, which underlines the importance of the Parzen-window regularization.

T2-PD registration (Figure 5). We use the mutual information criterion to realign two slices from a proton density (PD) and a T2-weighted magnetic resonance image (MRI) volume (same patient). A good introduction to medical image acquisition can be found in [8, 40]. An artificial geometric distortion has been applied to the

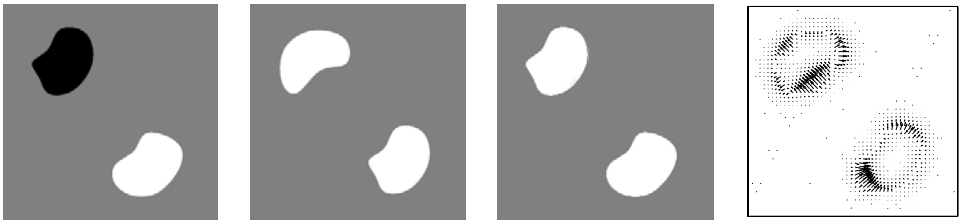


FIG. 4. *Mutual information criterion with a synthetic example. From left to right:* I_1 *(*256×256*),* I_2*,* $I_2(\mathbf{Id} + \mathbf{h}^*)$*, and a plot of the estimated optimal displacement field* $\mathbf{h}^*$*.*

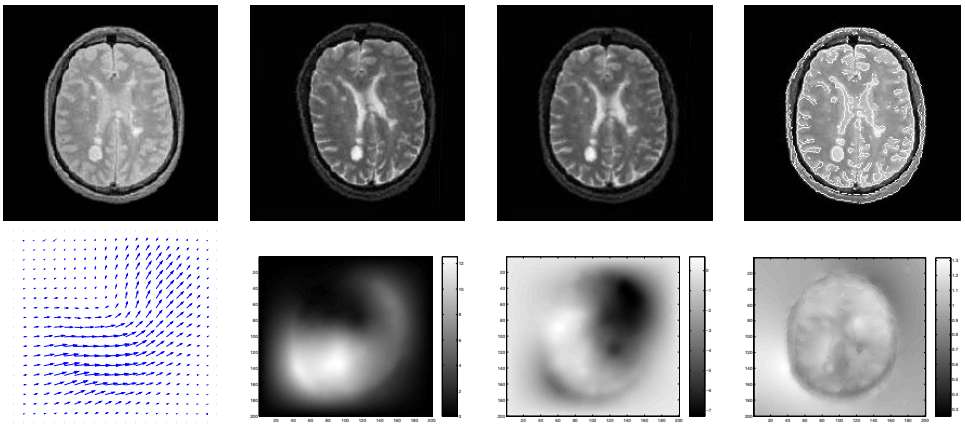


FIG. 5. *PD/T2 MRI registration using mutual information. First row, from left to right:* I_1*,* I_2*,* $I_2(\mathbf{Id} + \mathbf{h}^*)$*, and the contours of* $I_2(\mathbf{Id} + \mathbf{h}^*)$ *superimposed on* I_1*. The second row shows, from left to right: an arrow-plot of* $\mathbf{h}^*$ *(one every three pixels shown), the dense* x *and* y *components of* $\mathbf{h}^*$*, and the determinant of the Jacobian of the transformation* $\mathbf{Id} + \mathbf{h}^*$*.*

original preregistered dataset. In order to evaluate the accuracy of the realignment, we superimposed some contours of the T2 image (initial and recovered pose) over the reference image (PD). It gives a good qualitative indication of the quality of the registration. Most of the anatomical structures seem correctly realigned.

MRI-fMRI registration (Figure 6). This example shows an experiment with MR data of the brain of a macaque monkey. The reference image is a T1-weighted anatomical volume, and the image to register is a functional, mion (monocrystalline iron oxide nanoparticle) contrast, MRI (fMRI). The contrast in this modality is related to blood oxygenation level. This registration was obtained using mutual information. Each image in this figure shows the intersection of the image volume with three orthogonal planes. The crosses in each plane image show a point of interest. The observation of these points shows that the correspondence has been improved by the registration process.

Matching of anatomical versus diffusion-tensor–related MRI (Figure 7). Our next experiment shows the results of an experiment with real 3D MR data of a human brain. This registration was also obtained using mutual information. The reference image is a T1-weighted anatomical MRI of a human brain. The target image is a T2-weighted anatomical MRI from the same patient, which was acquired as part of the process of obtaining an image of the water diffusion tensor at each voxel (dif-

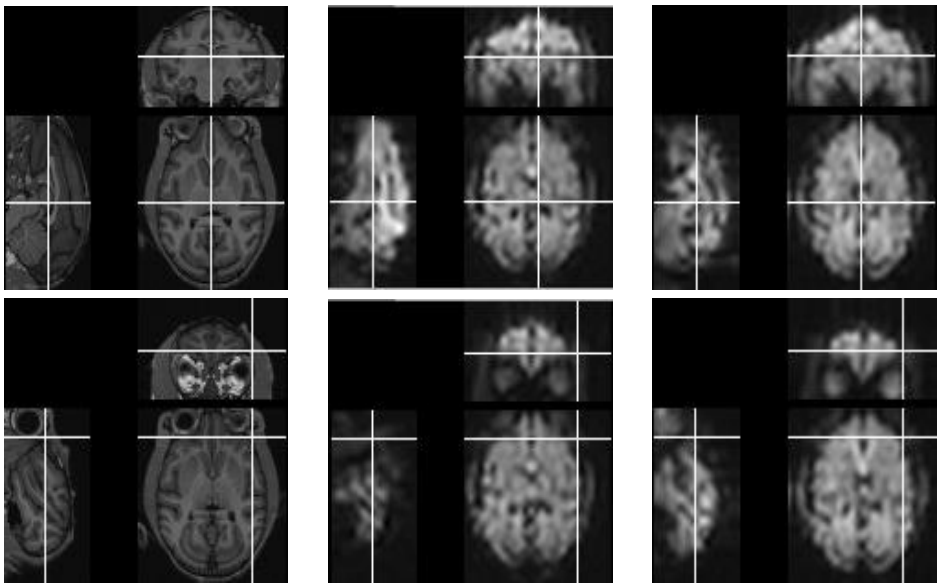


FIG. 6. *MRI-fMRI registration using the mutual information. The first and second images in each row show the reference anatomical MRI and the initial fMRI, respectively. The third image shows the registered fMRI. The two rows show two different points of interest in the volume.*

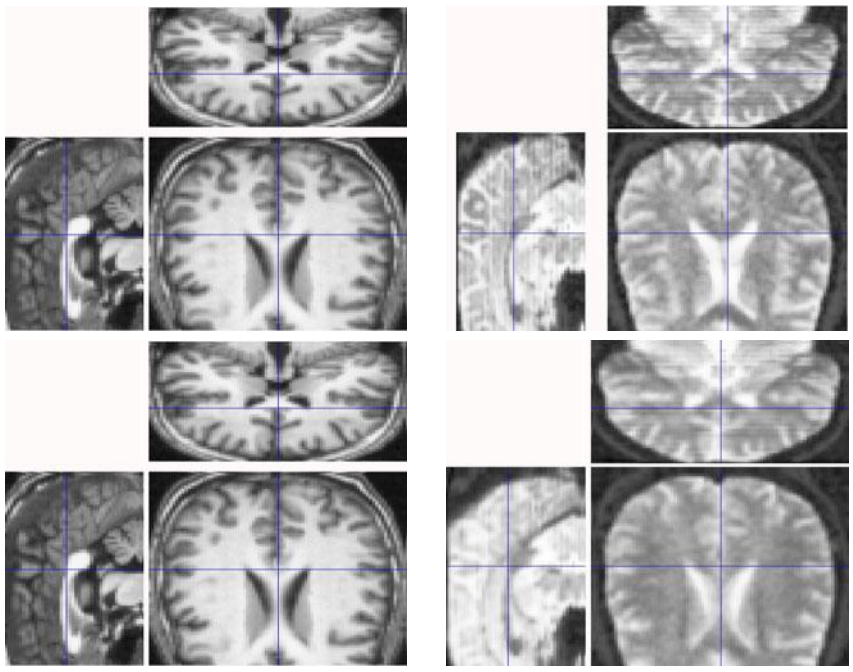


FIG. 7. *Matching of anatomical vs. diffusion-tensor-related MRI, using mutual information (see text).*

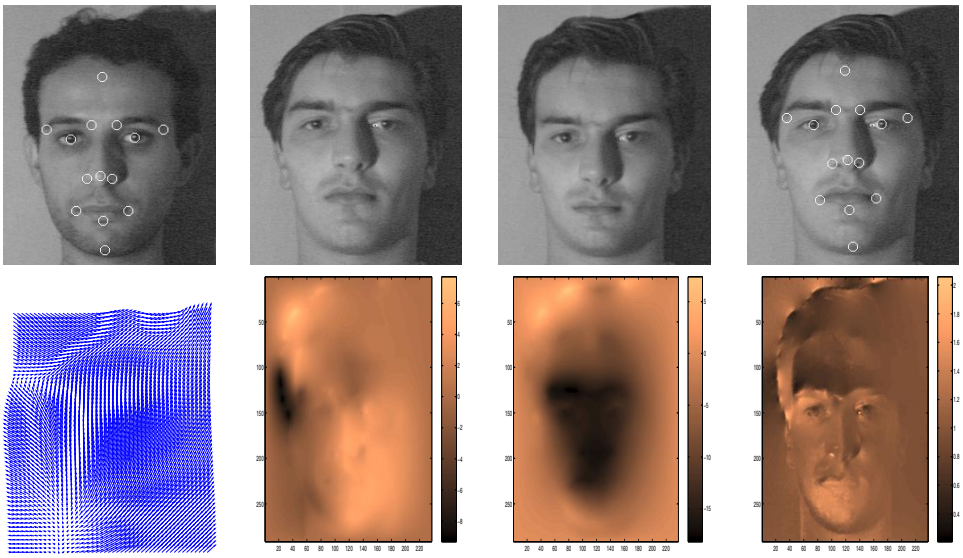


FIG. 8. *Human face template matching. First row from left to right:* I_1 *with some reference points marked,* I_2*,* $I_2(\mathbf{Id} + \mathbf{h}^*)$*, and* I_2 *with the corresponding reference points according to the estimated displacement field* $\mathbf{h}^*$*. The second row shows, from left to right: an arrow-plot of* $\mathbf{h}^*$ *(one every five pixels shown), its dense* x *and* y *components, and the determinant of the Jacobian of the transformation* $\mathbf{Id} + \mathbf{h}^*$*. The local cross-covariance was used as similarity criterion.*

fusion MR). Notice that the intensities in this modality are related in a noninvertible way, i.e., the correspondence $i_1 \to i_2$ is not a monotonous function, thereby justifying the use of mutual information. The estimated deformation field has a dominant y component, a property which is physically coherent with the applied gradient. The observation of the points pointed at by the crosses shows that the correspondence has been improved by the registration process.

Face template matching (Figure 8). This experiment shows template matching of human faces. The different albedos (fractions of incident light reflected by the surfaces) of the two skins create a "multimodal" situation, and the transformation is truly nonrigid due to the different shapes of the noses and mouths. Notice the excellent matching of the different features. This result was obtained using local cross-covariance. The running time was approximately five minutes on a PC at 900MHz. With the correspondences, one can interpolate the displacement field and the texture to perform fully automatic morphing.

The Eiffel tower example (Figure 9). This last experiment shows matching under varying illumination conditions. The Eiffel tower is taken at nearly one year distance under very different weather conditions. The result was obtained using local cross-covariance.

Conjugate gradient minimization. The explicit time discretization using a fixed time step corresponds to a steepest descent method without line-search, which is generally quite inefficient. In this section, we present results obtained with a modified version of the algorithm, which performs line-searching and uses a Fletcher–Reeves conjugate gradient minimization routine as described in [46]. Figure 10 shows plots of the decrease of the functional for the face template matching experiment of Figure 8

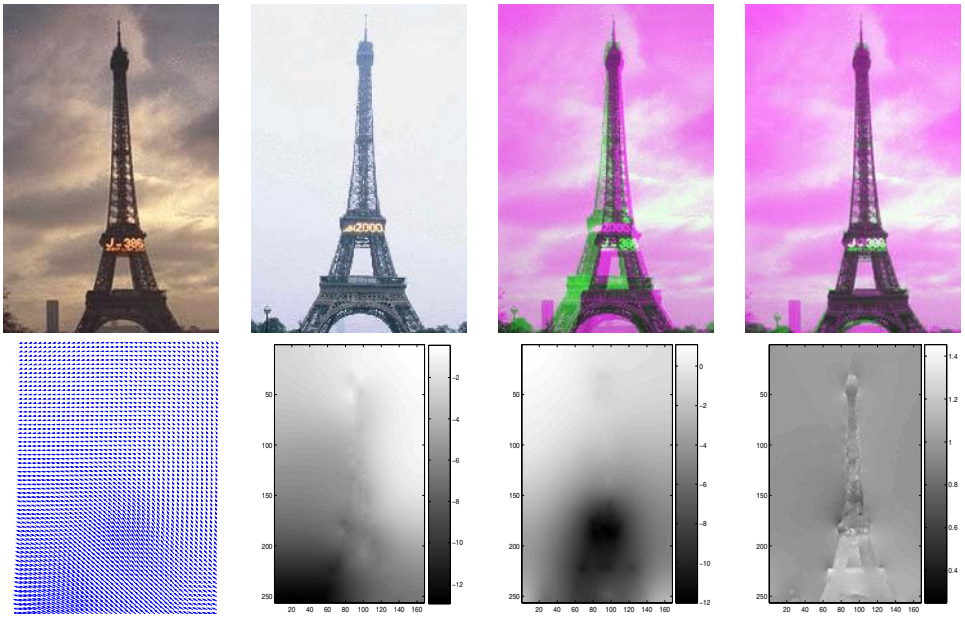


FIG. 9. *Matching under varying illumination conditions. First row from left to right:* I_1, I_2, I_2 *superimposed to* I_1, *and* $I_2(\mathbf{Id}+\mathbf{h}^*)$ *superimposed to* I_1. *The second row shows, from left to right: an arrow-plot of* $\mathbf{h}^*$ *(one every three pixels shown), its dense* x *and* y *components, and the determinant of the Jacobian of the transformation* $\mathbf{Id}+\mathbf{h}^*$. *The local cross-covariance was used as similarity criterion.*

using both the standard steepest descent and the modified conjugate gradient version. Plots (b) and (c) show the advantage of the multiresolution approach over the single resolution plot in (a), as a local minimum is avoided using a five-level multiresolution pyramid (the three coarsest levels are shown in (b), the finest two in (c)). Plots (d) and (e) show that the conjugate gradient method allows about one order of magnitude reduction in the total number of iterations required. In this example the gain in speed is much higher since the number of iterations at the finest level in (e) is very small, despite the fact that each iteration is slightly more costly. The result of Figure 8 is obtained in less than two seconds on a PC at 2.4GHz using the conjugate gradient version, instead of four minutes using the multiresolution approach. The same criterion was used to stop iterations in every case.

7. Conclusion. Image registration is best posed as an optimization problem defined over some functional space. The optimization functional is in general the sum of two terms, a data term and a regularization term. When one deals with different image modalities, the data term has to rely on image statistics rather than directly on the image intensities. This has the effect of making the data term a bit complicated. We have considered two such terms, a global and a local one. We have considered only one regularization term, but it is in effect quite representative. In order to prove that the registration problem is well posed, we have shown the existence of minimizers, computed the Euler–Lagrange equations induced by our functionals, and shown that the corresponding evolution functional equations had a unique and regular solution for a given initial condition.

In order to prove this last point we have used well-known theorems in functional

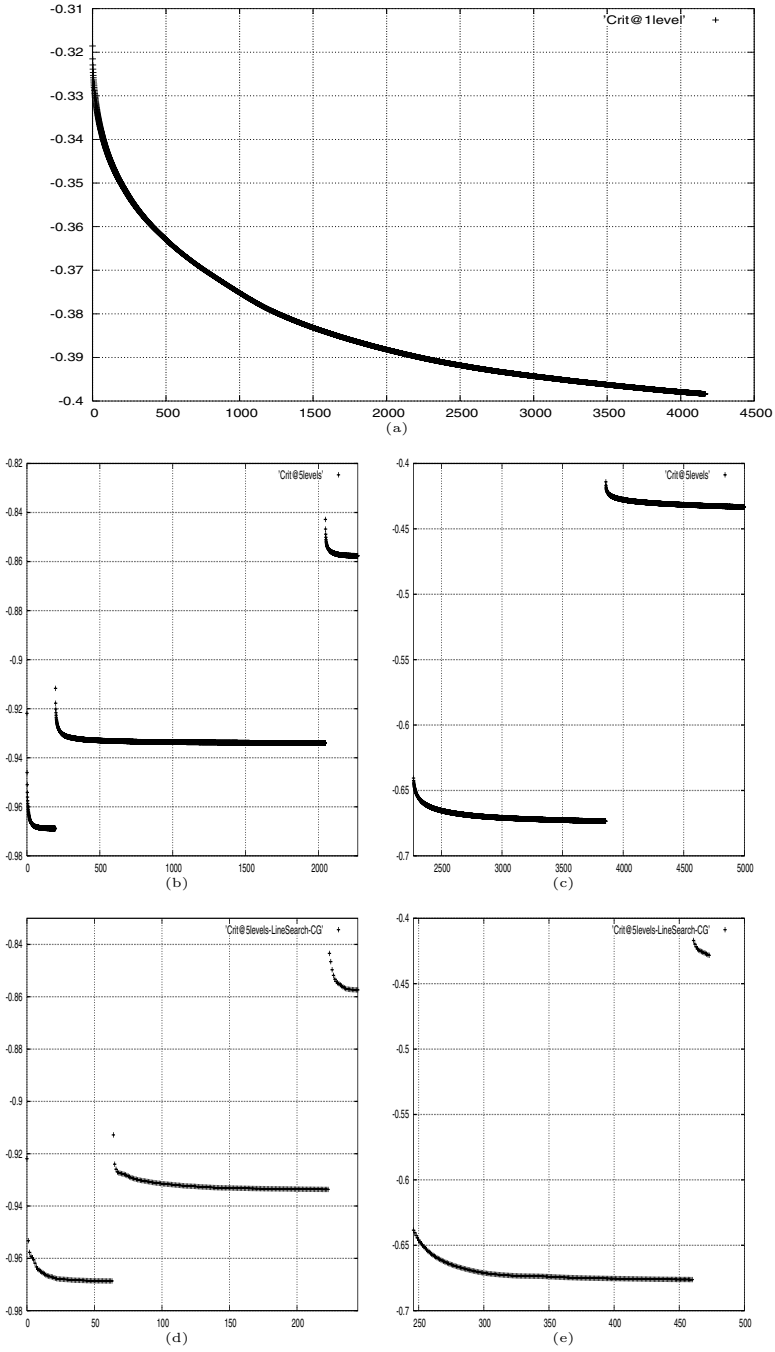


FIG. 10. *Plots of the decrease of the functional for the face template matching experiment of Figure* 8 *as a function of the number of iterations using both the standard steepest descent and the modified conjugate gradient version. Plots* (b) *and* (c) *show the advantage of the multiresolution approach over the single resolution plot in* (a)*, as a local minimum is avoided by using a five-level multiresolution pyramid (the three coarsest levels are shown in* (b)*). Plots* (d) *and* (e) *show that the conjugate gradient method allows about one order of magnitude reduction in the total number of iterations required. In this example the gain in speed is much higher since the number of iterations at the finest level in* (e) *is very small. The same criterion was used to stop iterations in every case.*

analysis. These theorems say that if the differential operator defined by the regularization term is strictly elliptic and the gradient of the data term is Lipschitz continuous, then the functional evolution equation has a unique classical solution. We have proved that these hypotheses are satisfied in our two examples. Since they are "generic" in image registration, this suggests that most of these problems are well posed. Our numerical implementation for computing the solutions of the Euler–Lagrange equations has allowed us to demonstrate the power and the flexibility of the approach for solving a wide range of difficult image registration problems.

REFERENCES

[1] L. Alvarez, R. Deriche, J. Weickert, and J. Sànchez. *Dense disparity map estimation respecting image discontinuities: A PDE and scale-space based approach*, Int. J. Visual Communication and Image Representation, Special Issue on Partial Differential Equations in Image Processing, Computer Vision, and Computer Graphics, 13 (2002), pp. 3–21.

[2] L. Alvarez, J. Weickert, and J. Sánchez, *Reliable estimation of dense optical flow fields with large displacements*, Int. J. Computer Vision, 39 (2000), pp. 41–56.

[3] G. Aubert, R. Deriche, and P. Kornprobst, *Computing optical flow via variational techniques*, SIAM J. Appl. Math., 60 (1999), pp. 156–182.

[4] G. Aubert and P. Kornprobst, *A mathematical study of the relaxed optical flow problem in the space BV (Ω)*, SIAM J. Math. Anal., 30 (1999), pp. 1282–1308.

[5] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations*, Appl. Math. Sci. 147, Springer-Verlag, New York, 2002.

[6] D. Bosq, *Nonparametric Statistics for Stochastic Processes*, 2nd ed., Lecture Notes in Statistics 110, Springer-Verlag, New York, 1998.

[7] H. Brezis, *Analyse Fonctionnelle. Théorie et Applications*, Masson, Paris, 1983.

[8] R. B. Buxton, *Introduction to Functional Magnetic Resonance Imaging*, Cambridge University Press, Cambridge, UK, 2002.

[9] P. Cachier and N. Ayache, *Regularization in Non-Rigid Registration:* I. *Trade-off between Smoothness and Intensity Similarity*, Technical report 4188, INRIA, Rocquencourt, France, 2001.

[10] P. Cachier and N. Ayache, *Regularization in Non-Rigid Registration*: II. *Isotropic Energies, Filters, and Splines*, Technical report 4243, INRIA, Rocquencourt, France, 2001.

[11] P Cachier and X. Pennec, 3*d non-rigid registration by gradient descent on a Gaussian weighted similarity measure using convolutions*, in Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2000), Hilton Head, SC, 2000, IEEE Press, Los Alamitos, CA, 20000, pp. 182–189.

[12] C. Chefd'hotel, G. Hermosillo, and O. Faugeras, *Flows of diffeomorphisms for multimodal image registration*, in Proceedings of the International Symposium on Biomedical Imaging (ISBI 2002), Washington, DC, IEEE, Los Alamitos, CA, 2002.

[13] G. Christensen, M. I. Miller, and M. W. Vannier, *A* 3*d deformable magnetic resonance textbook based on elasticity*, in Proceedings of the American Association for Artificial Intelligence Symposium: Applications of Computer Vision in Medical Image Processing, Stanford, CA, 1994.

[14] U. Clarenz, S. Henn, M. Rumpf, and K. Witsch, *Relations between optimization and gradient flow methods with applications to image registration*, in Multigrid and Related Methods for Optimization Problems, W. Hackbusch and M. Griebel, eds., Proceedings of the Gesellschaft für Angewondte Mathematik und Mechanik (GAMM) seminar, Leipzig, 2002.

[15] R. Courant, *Calculus of Variations*, New York, 1946.

[16] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, 6 volumes, Springer, New York, 2000.

[17] R. Deriche, *Fast algorithms for low-level vision*, IEEE Trans. Pattern Analysis and Machine Intelligence, 1 (1990), pp. 78–88.

[18] R. Deriche, P. Kornprobst, and G. Aubert, *Optical flow estimation while preserving its discontinuities: A variational approach*, in Proceedings of the 2nd Asian Conference on Computer Vision, Singapore, 1995, Vol. 2, Lecture Notes in Comput. Sci. 1035, Springer-Verlag, New York, 1996, pp. 71–80.

[19] L. C. Evans, *Partial Differential Equations*, Graduate Studies in Math. 19, 1998.

[20] O. Faugeras, B. Hotz, H. Mathieu, T. Viéville, Z. Zhang, P. Fua, E. Théron, L. Moll, G. Berry, J. Vuillemin, P. Bertin, and C. Proy, *Real Time Correlation Based Stereo: Algorithm Implementations and Applications*, Technical report 2013, INRIA, Sophia-Antipolis, France, 1993.

[21] O. Faugeras and R. Keriven, *Variational principles, surface evolution, PDEs, level set methods and the stereo problem*, IEEE Trans. Image Process., 7 (1998), pp. 336–344.

[22] T. Gaens, F. Maes, D. Vandermeulen, and P. Suetens, *Non-rigid multimodal image registration using mutual information*, in Proceedings of the Comput. Sci. First International Conference on Medical Image Computing and Computer-Assisted Intervention, J. van Leeuwen, G. Goos, and J. Hartmanis, eds., Lecture Notes in Comput. Sci. 1496, Springer, New York, 1998.

[23] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss. 8, Springer-Verlag, New York, 1983.

[24] N. Hata, T. Dohi, S. Warfield, W. Wells, III, R. Kikinis, and F. A. Jolesz, *Multimodality deformable registration of pre-and intra-operative images for MRI-guided brain surgery*, in Proceedings of the Comput. Sci. First International Conference on Medical Image Computing and Computer-Assisted Intervention, J. van Leeuwen, G. Goos, and J. Hartmanis, eds., Lecture Notes in Comput. Sci. 1496, Springer, New York, 1998.

[25] G. Hermosillo, *Variational Methods for Multimodal Image Matching*, Ph.D. thesis, INRIA, Rocquencourt, France, 2002; available online from ftp://ftp-sop.inria.fr/robotvis/html/Papers/hermosillo:02.ps.gz.

[26] D. Hill, *Combination of 3D Medical Images from Multiple Modalities*, Ph.D. thesis, Department of Radiological Sciences, University of London, London, 1993.

[27] W. Hinterberger, O. Scherzer, C. Schnörr, and J. Weickert, *Analysis of optical flow models in the framework of calculus of variations*, Numer. Funct. Anal. Optim., 23 (2002), pp. 69–89.

[28] B. K. Horn and B. G. Schunk, *Determining optical flow*, Artificial Intelligence, 17 (1981), pp. 185–203.

[29] T. Kanade and M. Okutomi, *A stereo matching algorithm with an adaptive window: Theory and experiment*, IEEE Trans. Pattern Anal. and Machine Intelligence, 16 (1994), pp. 920–932.

[30] J. J. Koenderink and A. J. van Doorn, *Blur and disorder*, in Proceedings of the Scale-Space Theories in Computer Vision, Second International Conference (Scale-Space'99), M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert, eds., Lecture Notes in Comput. Sci. 1682, Springer, 1999, pp. 1–9.

[31] M. E. Leventon and W. E. L. Grimson, *Multi-modal volume registration using joint intensity distributions*, in Medical Image Computing and Computer-Assisted Intervention-MICCAI'98, W. M. Wells, A. Colchester, and S. Delp, eds., Lecture Notes in Comput. Sci. 1496, Springer-Verlag, Cambridge, MA, 1998.

[32] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, *Multimodality image registration by maximization of mutual information*, IEEE Trans. Medical Imaging, 16 (1997), pp. 187–198.

[33] J. B. A. Maintz, H. W. Meijering, and M. A. Viergever, *General multimodal elastic registration based on mutual information*, in Medical Imaging 1998, Image Processing 3338, SPIE, Bellingham, WA, 1998, pp. 144–154.

[34] E. Mémin and P. Pérez, *A multigrid approach for hierarchical motion estimation*, in Proceedings of the 6th International Conference on Computer Vision, Bombay, India, 1998, IEEE Computer Society Press, Los Alamitos, CA, pp. 933–938.

[35] E. Mémin and P. Pérez, *Dense/parametric estimation of fluid flows*, in Proceedings of the IEEE International Conference on Image Processing (ICIP'99), Kobe, Japan, 1999, IEEE Press, Los Alamitos, CA.

[36] C. Meyer, J. Boes, B. Kim, and P. Bland, *Evaluation of control point selection in automatic, mutual information driven, 3d warping*, in Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention, J. van Leeuwen, G. Goos, and J. Hartmanis, eds., Lecture Notes in Comput. Sci. 1496, 1998.

[37] M. Miller and L. Younes, *Group actions, homeomorphisms, and matching: A general framework*, Int. J. Computer Vision, 41 (2001), pp. 61–84.

[38] H. H. Nagel and W. Enkelmann, *An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences*, IEEE Trans. Pattern Anal. and Machine Intelligence, 8 (1986), pp. 565–593.

[39] T. Netsch, P. Rosch, A. van Muiswinkel, and J. Weese, *Towards real-time multi-modality 3d medical image registration*, in Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada, IEEE Computer Society Press, Los Alamitos, 2001.

[40] W. W. Orrison, J. D. Lewine, J. A. Sanders, and M. F. Hartshorne, *Functional Brain Imaging*, Mosby-Year Book, 1995.

[41] S. Ourselin, A. Roche, S. Prima, and N. Ayache, *Block matching: A general framework to improve robustness of rigid registration of medical images*, in Medical Image Computing and Computer-Assisted Intervention-MICCAI 2000, Lecture Notes in Comput. Sci. 1935, Springer-Verlag, New York, 2000.

[42] E. Parzen, *On the estimation of probability density function*, Ann. Math. Statist., 33 (1962), pp. 1065–1076.

[43] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[44] X. Pennec, P. Cachier, and N. Ayache, *Understanding the "demon's algorithm": 3d nonrigid registration by gradient descent*, in Proceedings of the Second International Conference on Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Comput. Sci. 1679, Springer, 1999, pp. 597–605.

[45] G. Penney, J. Weese, J. A. Little, P. Desmedt, D. L. G. Hill, and D. J. Hawkes, *A comparison of similarity measures for use in 2d-3d medical image registration*, in Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention, J. van Leeuwen, G. Goos, and J. Hartmanis, eds., Lecture Notes in Comput. Sci. 1496, Springer-Verlag, New York, 1998.

[46] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK, 1988.

[47] M. Proesmans, L. Van Gool, E. Pauwels, and A. Oosterlinck, *Determination of optical flow and its discontinuities using non-linear diffusion*, in Proceedings of the 3rd ECCV, II, Lecture Notes in Comput. Sci. 801, Springer-Verlag, New York, 1994, pp. 295–304.

[48] A. Roche, A. Guimond, J. Meunier, and N. Ayache, *Multimodal elastic matching of brain images*, in Proceedings of the 6th European Conference on Computer Vision, Dublin, 2000, Lecture Notes in Comput. Sci. 1842–1843, Springer-Verlag, Berlin, 2000.

[49] A. Roche, G. Malandain, X. Pennec, and N. Ayache, *The correlation ratio as new similarity metric for multimodal image registration*, in Medical Image Computing and Computer-Assisted Intervention-MICCAI'98, W. M. Wells, A. Colchester, and S. Delp, eds., Lecture Notes in Comput. Sci. 1496, Springer, Cambridge, MA, 1998, pp. 1115–1124.

[50] D. Rückert, C. Hayes, C. Studholme, P. Summers, M. Leach, and D. J. Hawkes, *Nonrigid registration of breast MR images using mutual information*, in Medical Image Computing and Computer-Assisted Intervention-MICCAI'98, W. M. Wells, A. Colchester, and S. Delp, eds., Lecture Notes in Comput. Sci. 1496, Springer, Cambridge, MA, 1998.

[51] D. Scharstein and R. Szeliski, *Stereo matching with nonlinear diffusion*, Int. J. Computer Vision, 28 (1998), pp. 155–174.

[52] H. Tanabe, *Equations of Evolution*, Pitman, Boston, 1975.

[53] J. P. Thirion, *Image matching as a diffusion process: An analogy with Maxwell's demons*, Medical Image Analysis, 2 (1998), pp. 243–260.

[54] A. Trouvé, *Diffeomorphisms groups and pattern matching in image analysis*, Int. J. Computer Vision, 28 (1998), pp. 213–221.

[55] P. Viola, *Alignment by Maximization of Mutual Information*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1995.

[56] P. Viola and W. M. Wells III, *Alignment by maximization of mutual information*, Int. J. Computer Vision, 24 (1997), pp. 137–154.

[57] J. Weickert and C. Schnörr, *A theoretical framework for convex regularizers in PDE-based computation of image motion*, Int. J. Computer Vision, 45 (2001), pp. 245–264.

[58] W. M. Wells, III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, *Multi-modal volume registration by maximization of mutual information*, Medical Image Anal., 1 (1996), pp. 35–51.

[59] R. P. Woods, J. C. Maziotta, and S. R. Cherry, *MRI-PET registration with automated algorithm*, J. Computer Assisted Tomography, 17 (1993), pp. 536–546.

[60] Z. Zhang, R. Deriche, O. Faugeras, and Q. T. Luong, *A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry*, Artificial Intelligence Journal, 78 (1995), pp. 87–119.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1588–1608

A RADIAL BIPHASIC MODEL FOR LOCAL CELL-MATRIX MECHANICS IN ARTICULAR CARTILAGE*

MANSOOR A. HAIDER†

Abstract. Analytical and numerical solutions are presented for an interface problem that models deformation in the local cell-matrix unit (chondron) of articular cartilage. The cell and its protective pericellular matrix layer are modeled as isotropic biphasic continua deforming in small strain. A spherical geometry with purely radial deformation is assumed. Enforcement of the boundary and interface conditions results in an eigenvalue problem that is self-adjoint when the permeabilities of the cell and the layer are the same. In this case, a series solution of the interface problem is presented for a time-varying displacement prescribed at the boundary of the pericellular layer. The case of nonuniform permeability is considered via a numerical finite difference solution. The analytical and numerical solutions are used to conduct a parametric analysis of mechanical signal transmission due to an applied sinusoidal displacement. The dual role of the pericellular matrix as a mechanical signal transmitter and a protective layer is analyzed. For frequencies in the range 0-3Hz, transmission of transient-free radial displacement, solid stress, and strain are evaluated with varying pericellular stiffness and permeability in biphasic models of normal and osteoarthritic chondrons.

Key words. articular cartilage, chondrocyte, chondron, pericellular matrix, cartilage mechanics, mechanical signal transduction, biphasic theory, interface problem, eigenvalue problem

AMS subject classifications. 92, 74, 35, 41

DOI. 10.1137/S0036139902417700

1. Introduction.

Articular cartilage is a hydrated biological soft tissue that lines the surfaces of diarthroidal joints such as the knee, shoulder, and hip. The primary function of cartilage is to distribute stresses in load-bearing and to provide a low friction surface for joint motion [15]. While cartilage can perform these functions over a lifetime, the degeneration of cartilage, called osteoarthritis (OA), is a common condition that progresses with age. The structure of cartilage arises from an *extracellular matrix (ECM)* that consists predominantly of cross-linked type-II collagen fibers and entrapped proteoglycan macromolecules. Embedded in the ECM are specialized cells called *chondrocytes* (Figure 1.1). The metabolic activity of the chondrocytes dictates maintenance and turnover of the ECM constituents and hence the structural integrity of the tissue [22]. Within the ECM, the chondrocytes are locally isolated and make up less than 10% of the tissue volume in adult cartilage [21]. Articular cartilage is an avascular and aneural biological soft tissue. Consequently, chondrocyte metabolism depends not only on inherent genetic and biochemical factors but also on mechanical and physicochemical factors in the local environment of a single cell. Such factors include matrix stress, fluid pressurization, fixed-charge density of the ECM, and ionic composition of the interstitial fluid [8]. Describing the complex relationships between components of the local cell environment and external loading of a joint requires the formulation and solution of biomechanical models at several scales. Knowledge of these relationships is an important step in describing the sequence of events that cause changes in the local cellular environment to alter the biochemical response of a chondrocyte.

*Received by the editors November 11, 2002; accepted for publication (in revised form) October 31, 2003; published electronically June 22, 2004. This work was supported by funding from the NSF (DMS-0211154) and the NIH (AG-15768).

http://www.siam.org/journals/siap/64-5/41770.html

†Department of Mathematics, North Carolina State University, Box 8205, Raleigh, NC 27695-8205 (m_haider@ncsu.edu).

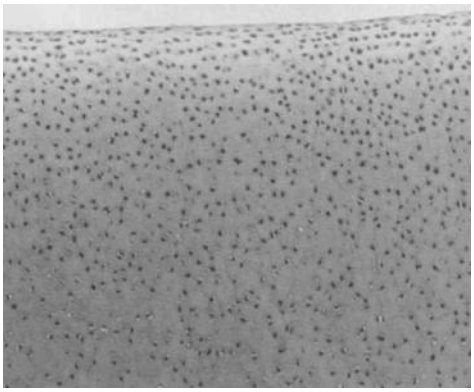


FIG. 1.1. *A photomicrograph of articular cartilage. The chondrocytes (black) are cellular inclusions that populate the extracellular matrix (grey).* Reprinted from F. Guilak, A. Ratcliffe, and V. C. Mow, *Chondrocyte deformation and local tissue strain in articular cartilage: A confocal microscopy study,* J. Orthop. Res., 13 (1995), pp. 410–421, with permission from Orthopaedic Research Society[7].

The local mechanical environment of a single cell is strongly influenced by its *pericellular matrix (PCM)*. The PCM is a protective layer that encapsulates the chondrocyte and is believed to play a significant role in regulating transmission of mechanical signals to the cell. Together, the cell and its PCM are termed a *chondron*. As a joint undergoes loading, mechanical signals are transmitted via the ECM to each chondron and, via the PCM, to each cell which, in response, can alter its metabolic activity. The composition of the PCM differs from that of the ECM in that the PCM is dominated by type-VI collagen and has a higher proteoglycan density. To quantify the effect of the PCM on mechanical signal transmission using biomechanical models, material parameter values are required for the chondrocyte and the PCM.

Using video microscopy, elastic parameter values for the chondrocyte are typically obtained by employing solutions of static contact problems for an elastic half-space [23] or sphere [9] that model micropipette aspiration testing of isolated cells. In a study of human chondrocytes [12], cells extracted via enzymatic digestion were found to be nearly incompressible, with a mean modulus on the order of 1kPa. No significant difference was observed between cells in the healthy and osteoarthritic sample groups. Recently, the solution of a layered elastic contact problem was used to determine elastic properties of the PCM via micropipette aspiration tests performed on intact chondrons that were mechanically extracted from human articular cartilage [1]. The mean Young's modulus of chondron PCM from healthy human articular cartilage was found to be 66.5kPa, whereas a 38% decrease in the mean modulus to 41.2kPa was observed in the osteoarthritic group [1]. In a multiscale finite element analysis [6], the macroscopic solution for transient deformation of a cartilage layer under a step load was computed and used to solve a separate microscale problem to determine the mechanical environment of a single chondrocyte. In this study, the inclusion of a PCM layer in the microscale model significantly altered the mechanical environment of a single cell. Together, these findings support the hypotheses that the PCM acts as a stiff protective layer that strongly influences the local environment of a chondrocyte, and that the mechanical properties of the PCM can be altered significantly with OA.

In the current study, analytical and numerical solutions are presented for an

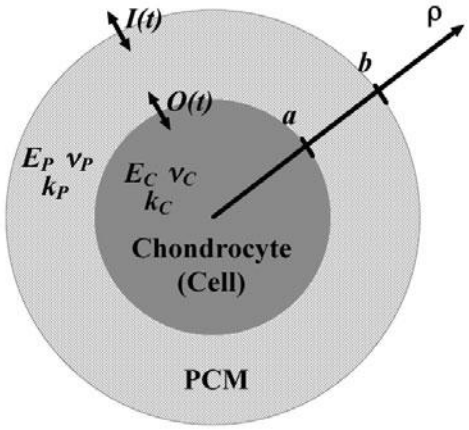


FIG. 1.2. *Spherical model of a chondron. The chondrocyte is modeled as a biphasic sphere with an attached biphasic layer that represents the pericellular matrix. Of primary interest is the effect of the PCM layer on transmission of a mechanical signal $I(t)$ from the chondron boundary to the cell-PCM interface.*

interface problem that models local transmission of mechanical signals in a single chondron. The chondron is idealized to consist of a spherical chondrocyte with an attached spherical layer that represents the PCM (Figure 1.2). Both regions are assumed to be biphasic continua that are isotropic and deforming in small strain. At present, experimental values are available for only the "drained" elastic moduli associated with equilibrated deformation states (where fluid motion has ceased) of isolated chondrocytes [12] and chondrons [1]. Hence, the primary focus of the current study is the effect of the large PCM stiffness, relative to the cell, on mechanical signal transmission through the PCM layer. Given the lack of direct measurements of permeability in the chondron, ranges of permeability in the cell and PCM are estimated (see section 4.2). Enforcement of the boundary and interface conditions results in an eigenvalue problem that is self-adjoint when the permeability of the chondron is assumed to be uniform (i.e., $k_C = k_P$). The resulting characteristic equation is analyzed in the limit of a very stiff PCM layer. Via Duhamel's principle, a series solution is obtained for a time-varying displacement prescribed at the boundary of the chondron. The case $k_C \neq k_P$ is considered using a numerical finite difference solution. These solutions are evaluated for the case of an applied sinusoidal displacement and used to assess the dual role of the PCM as a mechanical signal transmitter and a protective layer via a parametric analysis of radial displacement, solid stress, and strain. Since dissipation time scales are rapid, the transient-free oscillatory response is evaluated as a function of forcing frequency in a range that is representative of dynamic human motion (0–3Hz). Scaled amplitudes of the transient-free displacement and stress signals at the cell-PCM interface are computed as functions of frequency for several values of the material and geometric parameters. The analytical biphasic solution developed herein can also be used to verify the accuracy of biphasic finite element methods that model deformation of soft tissues with material interfaces via penalty methods (e.g., joint contact problems [5]).

2. Reduction of biphasic theory for the case of radial deformation.

2.1. Biphasic governing equations. The modern mixture theories [3], [25] provide a foundation for modeling the mechanics of articular cartilage and other biological soft tissues. In these mixture models, variations in structure and composition within a tissue are quantified via material parameters. The biphasic model [18], [16], based on Bowen's theory of incompressible mixtures [3], has been widely employed in modeling the mechanics of articular cartilage and other orthopedic tissues (e.g., intervertebral disc [11], bone [14], meniscus [20]).

In biphasic models of articular cartilage, the ECM is idealized as a solid phase that is saturated by a second phase of interstitial fluid. This continuum model is valid on scales for which the two phases can be treated as superimposed continua with an interphase drag mechanism [13]. Since the diameter of a chondrocyte is several orders of magnitude larger than the characteristic length scale of the ECM constituents and the cell cytoskeleton, this continuum approach is appropriate for modeling local biphasic cell-matrix mechanics. While biphasic models of cartilage ignore physicochemical effects, they capture essential load-bearing mechanisms including deformation of the tissue matrix, pressurization of the interstitial fluid, and dissipation due to interphase drag.

In this study, flow and deformation in the chondron are modeled using linear biphasic theory [18]. The momentum balance laws for the solid and fluid phases are, respectively,

$$\nabla \cdot \sigma^s + \Pi = \mathbf{0}, \qquad \nabla \cdot \sigma^f - \Pi = \mathbf{0}, \tag{2.1}$$

where σ^s and σ^f are partial Cauchy stress tensors that measure the force per unit mixture area on each phase. The symbol Π denotes a momentum exchange vector that accounts for the interphase drag force as fluid flows past solid in the mixture. Note that, in biphasic models of cartilage, the contribution of inertial terms to the momentum balance equations is negligible as the motion is dominated by elastic deformation and diffusive drag, and occurs at relatively low frequencies. The mixture is assumed to be intrinsically incompressible and saturated, so that

$$\phi \nabla \cdot (\partial_t \mathbf{u}) + (1-\phi)(\nabla \cdot \mathbf{v}) = 0, \tag{2.2}$$

where $\mathbf{u}$ is the solid displacement, $\mathbf{v}$ is the fluid velocity, and ϕ is the solid volume fraction, which is constant in the linear model and typically less than 20%. The solid phase is assumed to be linear elastic and isotropic, the fluid phase is assumed to be inviscid, and the momentum exchange is assumed to be due to Darcy's Law. The resulting constitutive laws are

$$\sigma^s = -\phi p \mathbf{I} + \lambda \mathrm{tr}(\mathbf{e})\mathbf{I} + 2\mu \mathbf{e}, \qquad \sigma^f = -(1-\phi)p\mathbf{I}, \qquad \Pi = K(\mathbf{v} - \partial_t \mathbf{u}), \tag{2.3}$$

where $\mathbf{I}$ is the identity tensor, p is a pore pressure used to enforce the incompressibility constraint (2.2), $\mathbf{e}(= 1/2(\nabla \mathbf{u} + \nabla \mathbf{u}^T))$ is the infinitesimal strain tensor, λ, μ are Lamé coefficients for the solid phase, and K is a diffusive drag coefficient. The Lamé coefficients λ, μ are associated with drained elastic equilibrium states that occur under static loading when all fluid flow has ceased in the mixture. An alternate set of elastic moduli are the Young's modulus E and Poisson ratio ν $(0 \le \nu < 0.5)$, where $\mu = \frac{E}{2(1+\nu)}$ and $\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}$.

By substituting (2.3) into (2.1), the governing equations (2.1) and (2.2) constitute a system of seven equations in the seven unknowns $\mathbf{u}, \mathbf{v}, p$. The fluid velocity $\mathbf{v}$ is

eliminated to yield the reduced system of four equations in the four unknowns $\mathbf{u}, p$:

$$\partial_t(\nabla\cdot\mathbf{u}) = k\nabla^2 p, \qquad \mu\left[\frac{1}{1-2\nu}\nabla(\nabla\cdot\mathbf{u}) + \nabla^2\mathbf{u}\right] = \nabla p, \tag{2.4}$$

where $k = \frac{(1-\phi)^2}{K}$ is the *permeability*. The fluid velocity is then given by

$$\mathbf{v} = \partial_t\mathbf{u} - \frac{1-\phi}{K}\nabla p. \tag{2.5}$$

Based on in situ testing of intact cartilage from knee joints, typical mean values for the ECM material parameters in the linear biphasic model are known to vary with both site and species. In [2], mean parameter values in normal human knees were found to lie in the ranges $H^A \equiv \lambda + 2\mu = 0.53$-$0.70$MPa, $\nu = 0.00$-0.10 and $k = 1.18$-$2.17 \times 10^{-15}\mathrm{m}^4/\mathrm{N}\cdot\mathrm{s}$.

2.2. Boundary and interface conditions. On a bounded biphasic domain Ω, a boundary value problem is obtained by solving (2.4) subject to boundary conditions that prescribe values of the pore pressure p, and solid displacement $\mathbf{u}$ or mixture traction vector $\mathbf{t} = (\sigma^s + \sigma^f)\cdot\mathbf{n}$ on the boundary $\partial\Omega$ with unit outward normal $\mathbf{n}$. When Ω contains an internal interface Γ with unit outward normal $\mathbf{n}$, the following biphasic interface conditions are enforced on Γ [10]:

$$[[\mathbf{u}]] = \mathbf{0}, \qquad [[p]] = 0, \qquad [[\lambda\mathrm{tr}(\mathbf{e})\mathbf{I} + 2\mu\mathbf{e}]]\cdot\mathbf{n} = \mathbf{0}, \qquad [[k\nabla p]]\cdot\mathbf{n} = 0, \tag{2.6}$$

where $[[\bullet]] \equiv (\bullet)_+ - (\bullet)_-$. Since the first two relations in (2.6) require continuous displacement and pressure, respectively, the last two relations provide four equations for the four unknown quantities on the interface Γ.

2.3. Reduced equations for the case of radial deformation. Let $\mathbf{r} = (\rho, \theta, \phi)$ denote a spherical coordinate system. A radial biphasic model for chondron deformation is developed by assuming that the displacement $\mathbf{u}(\mathbf{r}, t)$ has only a radial component and that the displacement and pressure $p(\mathbf{r}, t)$ vary only in the radial direction ρ. Hence,

$$\mathbf{u}(\mathbf{r}, t) = (u(\rho, t), 0, 0), \qquad p(\mathbf{r}, t) = p(\rho, t). \tag{2.7}$$

Substituting into the first relation of (2.4) and integrating, we arrive at

$$\partial_\rho p = k^{-1}\partial_t u - \rho^{-2}c(t), \tag{2.8}$$

where $c(t)$ is an arbitrary function of time. Substituting (2.7) into the second relation of (2.4) and using (2.8) yield

$$\partial_t u = kH^A\left(\rho^{-2}\partial_\rho(\rho^2\partial_\rho u) - 2\rho^{-2}u\right) + k\rho^{-2}c(t), \tag{2.9}$$

where $H^A = \lambda + 2\mu$ is called the *aggregate modulus*. For the radial chondron model, equations (2.8)–(2.9) are the reduced governing equations for the unknown displacement $u(\rho, t)$ and pore pressure $p(\rho, t)$.

In order to integrate (2.8), first take the divergence of the second relation in (2.4) to obtain

$$\nabla^2 p = H^A\nabla^2(\nabla\cdot\mathbf{u}). \tag{2.10}$$

It then follows that $p = H^A(\nabla\cdot\mathbf{u})+\psi$, where ψ is harmonic. Substituting this relation for p into (2.8) and using (2.9), one obtains

$$H^A\partial_\rho(\rho^{-2}\partial_\rho(\rho^2 u)) + \partial_\rho\psi = H^A\left(\rho^{-2}\partial_\rho(\rho^2\partial_\rho u) - 2\rho^{-2}u\right)$$

which simplifies to $\partial_\rho\psi = 0$. Hence the pore pressure is given by

$$p = H^A(2\rho^{-1}u + \partial_\rho u) + f(t), \tag{2.11}$$

where $f(t)$ is an arbitrary function of time.

3. Radial biphasic model of a chondron. A chondron consists of a single cell (chondrocyte) that is encapsulated by a PCM layer. For this initial model of chondron mechanics, a spherical geometry is assumed in which a biphasic cell is surrounded by an attached biphasic layer that represents the PCM (Figure 1.2). The displacement $\mathbf{u}$ and velocity $\mathbf{v}$ are assumed to be nonzero in only the radial direction and depend only on the radial coordinate ρ and time t. Consequently, the reduced biphasic equations of section 2.3 apply.

This study focuses on the use of biphasic solutions to model the effect of the large PCM-to-cell modulus ratio on signal transmission from the ECM to the cell via the PCM layer. Recent micropipette experiments on isolated chondrons indicated that the PCM was much stiffer than the chondrocyte in healthy human cartilage (mean $H_C^A/H_P^A = 0.044$), but that the PCM stiffness decreased significantly in the osteoarthritic group (mean $H_C^A/H_P^A = 0.071$) [1]. The biphasic solutions considered herein enable a parametric analysis that can provide insight into the effect of tissue degradation, and subsequent loss of PCM stiffness, on mechanical signal transmission in the chondron. A secondary benefit is that the analytical solution in the case $k_C = k_P$ can be used to verify the accuracy of biphasic finite element codes that employ penalty methods (e.g., [5]) to enforce the interface conditions (2.6).

3.1. Governing equations. The radial biphasic interface problem for mechanical signal transmission in a chondron is now formulated. In the cell ($0 \le \rho \le a$), the requirement that the solid velocity $\partial_t u$ is bounded at $\rho = 0$ implies that $c(t) \equiv 0$ in (2.9). In addition, the first relation in (2.6) implies that $\partial_t u$ is continuous at $\rho = a$. By (2.8), the last relation in (2.6) will then be satisfied provided that $c(t) \equiv 0$ for $a < \rho \le b$. Hence the reduced governing equations (2.9) and (2.11) are, respectively,

$$\partial_t u = \begin{cases} k_C H_C^A\left(\rho^{-2}\partial_\rho(\rho^2\partial_\rho u) - 2\rho^{-2}u\right), & 0 \le \rho \le a, \\ k_P H_P^A\left(\rho^{-2}\partial_\rho(\rho^2\partial_\rho u) - 2\rho^{-2}u\right), & a < \rho \le b, \end{cases} \quad t > 0, \tag{3.1}$$

$$p = \begin{cases} H_C^A(2\rho^{-1}u + \partial_\rho u) + f_C(t), & 0 \le \rho \le a, \\ H_P^A(2\rho^{-1}u + \partial_\rho u) + f_P(t), & a < \rho \le b, \end{cases} \quad t > 0. \tag{3.2}$$

The subscripts C and P denote quantities associated with the chondrocyte and PCM, respectively.

At this point, it is observed that the pressure $p(\rho, t)$ and displacement $u(\rho, t)$ are uncoupled in the reduced governing equations (3.1)–(3.2). The diffusion equation (3.1) can be solved on $0 \le \rho \le b$ subject to the first and third jump conditions of (2.6) at $\rho = a$, and a time-varying displacement boundary condition at $\rho = b$. The pressure is then obtained from (3.2), where $f_C(t)$ and $f_P(t)$ are used to satisfy the second jump condition of (2.6) at $\rho = a$, and to satisfy a pressure boundary condition at $\rho = b$. Note that satisfaction of the fourth jump condition in (2.6) follows from (2.8) by displacement continuity at $\rho = a$ and the fact that $c(t) = 0$ for $0 \le \rho \le b$.

3.2. Reduction to an eigenvalue problem in the case $k_C = k_P$. Consider separable solutions $u(\rho,t) = e^{-\gamma t}\phi(\rho)$ of (3.1). Substitution and separation of variables leads to the following eigenvalue problem for the eigenvalue-eigenfunction pair (γ,ϕ):

$$L\phi = \gamma\phi, \quad \text{where} \quad L \equiv \begin{cases} k_C H_C^A \rho^{-2}\left[\partial_\rho(\rho^2\partial_\rho) - 2\right], & 0 \le \rho \le a, \\ k_P H_P^A \rho^{-2}\left[\partial_\rho(\rho^2\partial_\rho) - 2\right], & a < \rho \le b. \end{cases} \tag{3.3}$$

In this study, the signal transmission problem is considered for a time-varying displacement $u(b,t)$ prescribed at the outer PCM boundary. The prescribed displacement models the arrival of a mechanical signal from the ECM at the boundary of the chondron. Since displacement is prescribed, a homogeneous boundary condition is employed for the eigenfunction ϕ at $\rho = b$ (time-dependence is incorporated via Duhamel's principle in section 3.3.1). Along with the homogeneous boundary condition, the first and third jump conditions in (2.6) give

$$\phi(b) = 0, \qquad \phi(a^+) = \phi(a^-), \qquad H_P^A \phi'(a^+) - H_C^A \phi'(a^-) = \frac{2(\lambda_C - \lambda_P)}{a}\phi(a). \tag{3.4}$$

Note that the third relation in (3.4) has been obtained from the third condition in (2.6) by employing the second relation in (3.4), where $\phi(a) \equiv \phi(a^+) = \phi(a^-)$.

It is straightforward to show that the operator L in (3.3) is self-adjoint in the special case where the permeability of the cell k_C is equal to the permeability of the PCM k_P (see Appendix A). In this case, all eigenvalues $\gamma_j (j = 1, 2, \dots)$ of the eigenvalue problem (3.3)–(3.4) are real and distinct, the corresponding eigenfunctions $\phi_j (j = 1, 2, \dots)$ are orthogonal, and the common value of the permeability is denoted by k.

Eigenfunctions for the eigenvalue problem (3.3)–(3.4) are constructed via the following representation based on spherical Bessel functions that satisfy (3.3):

$$\phi_j = \begin{cases} \frac{kH_C^A}{\gamma_j \rho^2}\left[\sin\left(\sqrt{\frac{\gamma_j}{kH_C^A}}\rho\right) - \sqrt{\frac{\gamma_j}{kH_C^A}}\rho\cos\left(\sqrt{\frac{\gamma_j}{kH_C^A}}\rho\right)\right], & 0 \le \rho \le a, \\ \frac{\delta_j kH_P^A}{\gamma_j \rho^2}\left[\sin\left(\sqrt{\frac{\gamma_j}{kH_P^A}}(\rho+\kappa_j)\right) - \sqrt{\frac{\gamma_j}{kH_P^A}}\rho\cos\left(\sqrt{\frac{\gamma_j}{kH_P^A}}(\rho+\kappa_j)\right)\right], & a < \rho \le b. \end{cases} \tag{3.5}$$

In (3.5), κ_j and δ_j are constants that are used to satisfy the interface conditions (3.4). Using (3.5) in the first condition of (3.4) gives

$$\kappa_j = \sqrt{\frac{kH_P^A}{\gamma_j}}\tan^{-1}\left(\sqrt{\frac{\gamma_j}{kH_P^A}}b\right) - b. \tag{3.6}$$

The second condition in (3.4) gives

$$\delta_j = \frac{H_C^A}{H_P^A}\frac{\sin\left(\sqrt{\frac{\gamma_j}{kH_C^A}}a\right) - \sqrt{\frac{\gamma_j}{kH_C^A}}a\cos\left(\sqrt{\frac{\gamma_j}{kH_C^A}}a\right)}{\sin\left(\sqrt{\frac{\gamma_j}{kH_P^A}}(a+\kappa_j)\right) - \sqrt{\frac{\gamma_j}{kH_P^A}}a\cos\left(\sqrt{\frac{\gamma_j}{kH_P^A}}(a+\kappa_j)\right)}. \tag{3.7}$$

Lastly, (3.5) is substituted into the last condition of (3.4). Using (3.6) and (3.7), the constants κ_j and δ_j are eliminated and, after some manipulation, the following nondimensional characteristic equation for the scaled eigenvalues $\bar{\gamma}_j = \frac{\gamma_j a^2}{kH_C^A}$ is obtained:

$$\begin{aligned}&\left[\bar{F}_1(\bar{\gamma})\sin\left(\sqrt{\bar{\gamma}}\right) + \bar{F}_3(\bar{\gamma})\cos\left(\sqrt{\bar{\gamma}}\right)\right]\sin\left(\frac{b-a}{a}\sqrt{\epsilon_1\bar{\gamma}}\right)\\ &\quad + \left[\bar{F}_2(\bar{\gamma})\cos\left(\sqrt{\bar{\gamma}}\right) + \bar{F}_4(\bar{\gamma})\sin\left(\sqrt{\bar{\gamma}}\right)\right]\cos\left(\frac{b-a}{a}\sqrt{\epsilon_1\bar{\gamma}}\right) = 0,\end{aligned} \tag{3.8}$$

where

$$\begin{aligned}\bar{F}_1(\bar{\gamma}) &= -4\epsilon_1(\beta_P - \epsilon_2) - \frac{4b}{a}\bar{\gamma}\epsilon_1^2(\beta_P - \epsilon_2) - \frac{b}{a}\bar{\gamma}^2\epsilon_1^3,\\ \bar{F}_2(\bar{\gamma}) &= -\frac{4(b-a)}{a}\bar{\gamma}\epsilon_1^{3/2}(\beta_P - \epsilon_2) + \frac{b}{a}\bar{\gamma}^2\epsilon_1^{5/2},\\ \bar{F}_3(\bar{\gamma}) &= 4\bar{\gamma}^{1/2}\epsilon_1(\beta_P - \epsilon_2) + \left(\frac{4b}{a}\beta_P - 1\right)\bar{\gamma}^{3/2}\epsilon_1^2 - \frac{4b}{a}\bar{\gamma}^{3/2}\epsilon_1^2\epsilon_2,\\ \bar{F}_4(\bar{\gamma}) &= \frac{4(b-a)}{a}\bar{\gamma}^{1/2}\epsilon_1^{3/2}(\beta_P - \epsilon_2) - \bar{\gamma}^{3/2}\epsilon_1^{5/2},\end{aligned}$$

$\beta_P = \frac{\mu_P}{H_P^A}$, $\epsilon_1 = \frac{H_C^A}{H_P^A}$, and $\epsilon_2 = \frac{\mu_C}{H_P^A}$.

Note that the eigenvalues have been scaled by the quantity $t_C \equiv a^2/(kH_C^A)$, which is the gel relaxation time [17] for the chondrocyte.

Given that the PCM is much stiffer than the chondrocyte, the characteristic equation (3.8) is analyzed in the regime where $\epsilon_1 \ll 1$. Using (3.8), it is straightforward to show that, to leading-order, the eigenvalues $\bar{\gamma}^{(0)}$ satisfy $\sin(\sqrt{\bar{\gamma}^{(0)}}) - \sqrt{\bar{\gamma}^{(0)}}\cos(\sqrt{\bar{\gamma}^{(0)}}) = 0$. The eigenvalues $\bar{\gamma}_j^{(0)}$ are well separated (e.g., $\bar{\gamma}_1^{(0)} \approx 20.19$, $\bar{\gamma}_2^{(0)} \approx 59.68$, $\bar{\gamma}_3^{(0)} \approx 118.90$) and, as $j \to \infty$, $\bar{\gamma}_j \sim (j + 1/2)^2\pi^2$ with separation $\Delta\bar{\gamma}_j = \bar{\gamma}_j - \bar{\gamma}_{j-1} \sim 2j\pi^2$. As the ratio ϵ_1 increases up to 0.1, the separation of the eigenvalues does not deviate far away from this asymptotic behavior. Hence, the first N roots of the characteristic equation (3.8) are readily obtained in MAPLE using the fsolve routine.

It is noted that the smallest leading-order eigenvalue $\bar{\gamma}_1^{(0)}$ represents a characteristic diffusion time t_P for stress relaxation in the chondron. Comparison to the biphasic gel relaxation time for the cell t_C indicates that

$$t_P = \frac{1}{\gamma_1} \sim \frac{1}{\bar{\gamma}_1}\frac{a^2}{kH_C^A} \approx 0.04953\frac{a^2}{kH_C^A} = (0.04953)t_C. \tag{3.9}$$

Hence, the radial chondron model indicates that the presence of a stiff PCM layer encapsulating the chondrocyte reduces the biphasic gel relaxation time by a factor of 20. This result is consistent with the functional role of the PCM as a mechanical signal transmitter. Components of the transient deformation are dissipated on time scales that are much faster than the typically lengthy relaxation times associated with stress diffusion in the ECM of articular cartilage. Hence, in the context of forced dynamic loading of a cartilage layer, it is likely that the transient-free oscillatory response is the dominant component of the mechanical signal that is transmitted to the chondrocyte.

3.3. Signal transmission model. The solution of (3.1)–(3.2) subject to a time-dependent displacement boundary condition at $\rho = b$ is now considered. At the cell-PCM interface ($\rho = a$) the first and third jump conditions in (2.6) reduce to

(3.10)
$$u(a^+,t) = u(a^-,t), \;\; H_P^A \partial_\rho u(a^+,t) - H_C^A \partial_\rho u(a^-,t) = \frac{2}{a}(\lambda_C - \lambda_P)u(a,t), \;\; t > 0,$$

respectively. The displacement must be zero at $\rho = 0$, and a time-dependent displacement signal is prescribed at the outer cell boundary:

$$u(0,t) = 0, \qquad u(b,t) = I(t), \qquad t > 0. \tag{3.11}$$

The chondron is also assumed to be free of displacement at $t = 0$ so that $u(\rho,0) = 0$ ($0 \le \rho \le b$). The second boundary condition in (3.11) models the arrival of a mechanical signal at the interface between the chondron and ECM. Clearly, the assumption of purely radial deformation is a simplification that facilitates a parametric analysis of signal transmission in the chondron. In reality, the relationship between local mechanics at the cellular scale and external loading of a cartilage layer is highly complex, requiring the solution of the three-dimensional biphasic equations at multiple length scales. As such, the applied displacement signal $I(t)$ is interpreted as a radial displacement averaged over the surface of the interface between the chondron and the extracellular matrix.

3.3.1. Series solution in the case $k_C = k_P$. The homogeneous differential equation (3.1), subject to the time-dependent boundary condition in (3.11), is transformed to a nonhomogeneous equation with a homogeneous boundary condition by writing $u(\rho,t) = v(\rho,t) + w(\rho,t)$, where

$$w(\rho,t) = I(t) \begin{cases} \alpha_3 \rho^2, & 0 \le \rho \le a, \\ \alpha_1 \rho^2 + \alpha_2 \rho, & a < \rho \le b. \end{cases} \tag{3.12}$$

The coefficients $\alpha_1, \alpha_2, \alpha_3$ are chosen such that $w(\rho,t)$ satisfies the interface conditions (3.10) and boundary conditions (3.11). They are

$$\alpha_1 = \frac{H_P^A - 2H_C^A - 2\lambda_C + 2\lambda_P}{b\chi}, \quad \alpha_2 = \frac{2a(H_C^A - H_P^A + \lambda_C - \lambda_P)}{b\chi}, \quad \alpha_3 = \frac{-H_P^A}{b\chi},$$

where $\chi = (b-2a)H_P^A + 2(b-a)(-H_C^A + \lambda_P - \lambda_C)$. The function $v(\rho,t)$ then satisfies the nonhomogeneous equation

$$\partial_t v = \begin{cases} kH_C^A \left(\rho^{-2}\partial_\rho(\rho^2 \partial_\rho v) - 2\rho^{-2} v\right) + h_C(\rho,t), & 0 \le \rho \le a, \\ kH_P^A \left(\rho^{-2}\partial_\rho(\rho^2 \partial_\rho v) - 2\rho^{-2} v\right) + h_P(\rho,t), & a < \rho \le b, \end{cases} \quad t > 0, \tag{3.13}$$

where the functions $h_C(\rho,t), h_P(\rho,t)$ are given by

$$h_C(\rho,t) = 4kH_C^A \alpha_3 I(t) - \alpha_3 \rho^2 I'(t), \qquad h_P(\rho,t) = 4kH_P^A \alpha_1 I(t) - (\alpha_1 \rho^2 + \alpha_2 \rho) I'(t).$$

Equation (3.13) is solved subject to the interface conditions

(3.14)
$$v(a^+,t) = v(a^-,t), \;\; H_P^A \partial_\rho v(a^+,t) - H_C^A \partial_\rho v(a^-,t) = \frac{2}{a}(\lambda_C - \lambda_P)v(a,t), \;\; t > 0,$$

the homogeneous boundary conditions

$$v(0,t)=0, \qquad v(b,t)=0, \qquad t>0, \tag{3.15}$$

and the modified initial condition

$$v(\rho,0)=-w(\rho,0), \qquad 0\le\rho\le b. \tag{3.16}$$

The solution of (3.13)–(3.16) is then obtained via Duhamel's principle [4]. Let

$$v(\rho,t)=\int_0^t \tilde{v}(\rho,t-s;s)ds, \tag{3.17}$$

where the new function $\tilde{v}(\rho,t;s)$ satisfies the homogeneous equation

$$\partial_t\tilde{v}=\begin{cases} kH_C^A\left(\rho^{-2}\partial_\rho(\rho^2\partial_\rho\tilde{v})-2\rho^{-2}\tilde{v}\right), & 0\le\rho\le a,\\ kH_P^A\left(\rho^{-2}\partial_\rho(\rho^2\partial_\rho\tilde{v})-2\rho^{-2}\tilde{v}\right), & a<\rho\le b, \end{cases} \qquad t>0 \tag{3.18}$$

subject to the interface conditions

$$\tilde{v}(a^+,t)=\tilde{v}(a^-,t),\ H_P^A\partial_\rho\tilde{v}(a^+,t)-H_C^A\partial_\rho\tilde{v}(a^-,t)=\frac{2}{a}(\lambda_C-\lambda_P)\tilde{v}(a,t),\ t>0, \tag{3.19}$$

the homogeneous boundary conditions

$$\tilde{v}(0,t)=0, \qquad \tilde{v}(b,t)=0, \qquad t>0, \tag{3.20}$$

and the initial condition

$$\tilde{v}(\rho,0;s)=-w(\rho,0)+\begin{cases} h_C(\rho,s), & 0\le\rho\le a,\\ h_P(\rho,s), & a<\rho\le b. \end{cases} \tag{3.21}$$

Assuming separable solutions, $\tilde{v}(\rho,t;s)=e^{-\gamma t}\phi(\rho)$, the eigenfunctions $\phi_j(j=1,2,\dots)$ and the eigenvalues $\gamma_j(j=1,2,\dots)$ are those of section 3.2. Hence, the solution of the time-dependent mechanical signal transmission model (3.1), (3.11) is given by

$$u(\rho,t)=w(\rho,t)+\sum_{j=1}^{\infty}\left[\int_0^t c_j(s)e^{\gamma_j s}ds\right]e^{-\gamma_j t}\phi_j(\rho), \tag{3.22}$$

where $w(\rho,t)$ is given by (3.12) and

$$c_j(s)=\frac{\langle h(\rho,s),\phi_j(\rho)\rangle}{\langle\phi_j(\rho),\phi_j(\rho)\rangle}, \quad h(\rho,s)=-w(\rho,0)+\begin{cases} h_C(\rho,s), & 0\le\rho\le a,\\ h_P(\rho,s), & a<\rho\le b. \end{cases} \tag{3.23}$$

Recall that the inner product in (3.23) is defined as $\langle f,g\rangle=\int_0^b f(s)g(s)s^2ds$. Once the displacement solution is known, a series solution for the pore pressure $p(\rho,t)$ (see (2.11)) is obtained by substitution of (3.22) into

$$p(\rho,t)=\begin{cases} H_C^A\left[2\rho^{-1}u(\rho,t)+\partial_\rho u(\rho,t)\right]+f_C(t), & 0\le\rho\le a,\\ H_P^A\left[2\rho^{-1}u(\rho,t)+\partial_\rho u(\rho,t)\right]+f_P(t), & a<\rho\le b. \end{cases} \tag{3.24}$$

The function $f_P(t)$ is used to match the pore pressure to a prescribed function $p_0(t)$ at $\rho = b$, so that

$$f_P(t) = p_0(t) - H_P^A \left[2b^{-1} I(t) + \partial_\rho u(b,t)\right]. \tag{3.25}$$

The function $f_C(t)$ is used to satisfy the second condition in (2.6) and is given by

$$f_C(t) = f_P(t) + 2a^{-1}(H_P^A - H_C^A)u(a,t) + H_P^A \partial_\rho u(a^+,t) - H_C^A \partial_\rho u(a^-,t). \tag{3.26}$$

From (3.22) and (3.24), series solutions for the partial stress components σ^s and σ^f in (2.3) and the fluid-phase velocity, $\mathbf{v}$, in (2.5) can also be obtained by substitution.

3.3.2. Finite difference solution in the case $k_C \neq k_P$. The governing displacement equation (3.1) is rewritten as

$$\partial_t u = r\left(\partial_\rho^2 u + 2\rho^{-1}\partial_\rho u - 2\rho^{-2}u\right), \qquad 0 < \rho < b, \quad \text{where } r = \begin{cases} k_C H_C^A & 0 < \rho < a, \\ k_P H_P^A & a < \rho < b. \end{cases} \tag{3.27}$$

Equation (3.27) was solved numerically using a finite difference method on a regular spatial mesh subject to the boundary conditions (3.11) and the interface conditions (3.10). In the discretization, backward and centered finite difference approximations were employed in time and space, respectively, at all interior mesh points except those immediately to the left and right of the interface at $\rho = a$, where the interface conditions (3.10) were enforced. The details of the numerical scheme employed to obtain the displacement solution are presented in Appendix B. Via first-order finite difference approximations of the spatial derivative of displacement, the radial strain $\partial_\rho u$, pore pressure (3.24), and solid stress (2.3) were evaluated once the displacement solution was obtained. The accuracy of the finite difference scheme was verified in the case $k_C = k_P$ by comparison to the analytical solution presented in section 3.3.1.

4. Application: Parametric analysis for a sinusoidal displacement.

4.1. Transient-free displacement solution. Deformation of the chondron was simulated for a prescribed sinusoidal displacement with amplitude u_0 and frequency ω at the outer PCM boundary:

$$u(b,t) = I(t) = u_0 \sin(\omega t), \quad t > 0. \tag{4.1}$$

This boundary condition models the arrival of a mechanical signal at the PCM-ECM interface under conditions of dynamic external loading of a joint.

In the case $k_C = k_P = k$, the function $h(\rho, s)$ in (3.23) reduces to

$$h(\rho, s) = u_0 \left[h_1(\rho)\sin(\omega s) + h_2(\rho)\cos(\omega s)\right],$$

where

$$h_1(\rho) = \begin{cases} 4kH_C^A\alpha_3, & 0 \le \rho \le a, \\ 4kH_P^A\alpha_1, & a < \rho \le b, \end{cases} \qquad h_2(\rho) = \begin{cases} -\alpha_3\omega\rho^2, & 0 \le \rho \le a, \\ -\omega(\alpha_1\rho^2 + \alpha_2\rho), & a < \rho \le b. \end{cases}$$

Hence, the coefficients $c_j(s)$ in (3.22)–(3.23) are given by

$$c_j(s) = \frac{u_0}{A_j}\left[A_j^{(1)}\sin(\omega s) + A_j^{(2)}\cos(\omega s)\right], \tag{4.2}$$

where

$$A_j^{(1)} = \langle h_1(\rho), \phi_j(\rho)\rangle,\ A_j^{(2)} = \langle h_2(\rho), \phi_j(\rho)\rangle,\ A_j = \langle \phi_j(\rho), \phi_j(\rho)\rangle. \tag{4.3}$$

Substituting these relations into (3.22), the following series solution is obtained for the case of a prescribed sinusoidal displacement:

$$\frac{u(\rho,t)}{u_0} = v_{tr}(\rho,t) + v_\infty(\rho,t) + \sin(\omega t)\begin{cases} \alpha_3\rho^2, & 0 \le \rho \le a, \\ \alpha_1\rho^2 + \alpha_2\rho, & a < \rho \le b, \end{cases} \tag{4.4}$$

where

$$v_{tr}(\rho,t) = \sum_{j=1}^{\infty} \frac{\omega A_j^{(1)} - \gamma_j A_j^{(2)}}{A_j(\gamma_j^2+\omega^2)} e^{-\gamma_j t}\phi_j(\rho), \tag{4.5}$$

$$v_\infty(\rho,t) = \sum_{j=1}^{\infty} \left[\frac{\gamma_j A_j^{(1)} + \omega A_j^{(2)}}{A_j(\gamma_j^2+\omega^2)} \sin(\omega t) + \frac{-\omega A_j^{(1)} + \gamma_j A_j^{(2)}}{A_j(\gamma_j^2+\omega^2)} \cos(\omega t) \right] \phi_j(\rho). \tag{4.6}$$

In (4.4), $v_{tr}(\rho,t)$ is the scaled transient displacement that, for a stiff PCM layer, tends to zero rapidly as $t \to \infty$. Consequently, the parametric analysis was based on evaluation of the steady displacement in a range of frequencies (0–3Hz) that is typical of human joint motion.

In the case $k_C \neq k_P$, the finite difference solution of section 3.3.2 was employed to evaluate transient-free deformation in the chondron. The time-marching method was allowed to run until a steady state deformation signal was obtained with amplitude that changed by less than 1%. Via a mesh refinement analysis in the case $k_C = k_P$, it was established that a spatial mesh resolution of $M = 300$ and a time step of $\Delta t = 1/300$ were sufficient for obtaining graphs in the parametric analysis that were indistinguishable in the two cases.

4.2. Parametric analysis. We employed the analytical and finite difference solutions in a parametric analysis of mechanical signal transmission through the PCM layer. Transient-free radial displacement, solid stress, and radial strain were used to assess the functional role of the PCM as a layer that facilitates transmission of mechanical signals to the chondrocyte and, simultaneously, protects the cell from excessive forces. In the parametric analysis, the PCM Young's modulus was varied in the range E_P = 10–100kPa, with particular attention to the mean measured values for normal (E_P = 66.5kPa) and OA (E_P = 41.3kPa) human chondrons obtained in a recent micropipette aspiration study [1]. Values of the Young's modulus for the cell, and the Poisson ratio for the cell and PCM were taken as E_C = 1kPa, $\nu_C = 0.43$ and $\nu_P = 0.04$, respectively [1]. Reference values for the radii of the chondrocyte and chondron were taken as $a = 10\mu$m and $b = 12.5\mu$m, respectively, and the solid volume fraction was $\phi_C = \phi_P = 0.17$ [6]. The amplitude of the prescribed displacement signal was chosen as $u_0 = (b-a)/10$.

To date, only limited estimates of permeability in the chondrocyte and PCM are available. In a micropipette analysis of creep deformation of isolated normal and OA human chondrocytes using a viscoelastic half-space model [24], a rough estimate of chondrocyte permeability on the order of 10^{-15}m^4/N·s was obtained using a characteristic creep time based on the biphasic gel relaxation time [18]. A more direct measurement of cellular permeability for osteoblast-like cells, via cytoindentation, found a much larger permeability on the order of 10^{-10}m^4/N·s [19]. While no direct

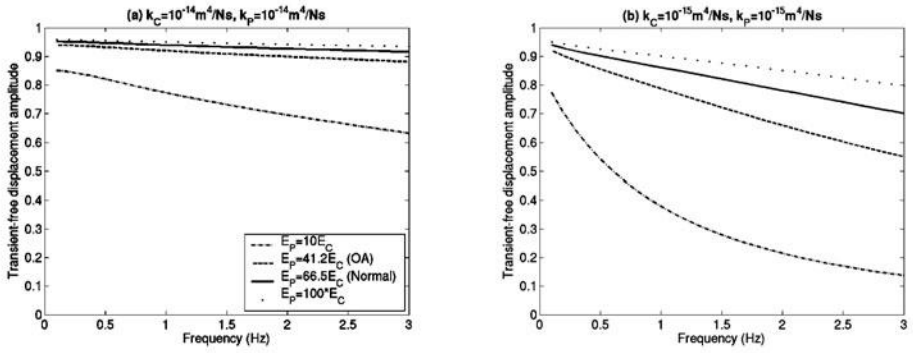


FIG. 4.1. *Scaled amplitude, $A(\omega)$, of the transient-free displacement at the cell-PCM interface* (4.7) *for frequencies in the range* 0–3*Hz in the case* $k_C = k_P$. (a) $k_C = k_P = 10^{-14} m^4/N{\cdot}s$; (b) $k_C = k_P = 10^{-15} m^4/N{\cdot}s$.

measurements of PCM permeability are currently available, given its composition, it is reasonable to assume that k_P should not deviate by more than a few orders of magnitude from the permeability of cartilage ECM (i.e., $\sim 10^{-15}\mathrm{m}^4/\mathrm{N}{\cdot}\mathrm{s}$). Based on these observations, the range of permeability for the parametric analysis was chosen as $(k_C, k_P) = (10^{-15}\text{–}10^{-12}, 10^{-16}\text{–}10^{-14})\mathrm{m}^4/\mathrm{N}{\cdot}\mathrm{s}$.

4.2.1. Transmission of displacement. To assess the role of the PCM as a transmissive layer that encapsulates the chondrocyte, the scaled amplitude of the transient-free displacement at the cell-PCM interface

$$A(\omega) = \frac{\mathrm{Amp}(u_\infty(a,t))}{u_0} \tag{4.7}$$

was evaluated. In the case $k_C = k_P$, $A(\omega)$ is given by the following series obtained from (4.4):

$$A(\omega) \approx \sqrt{\left(\alpha_3 a^2 + \sum_{j=1}^{N} \frac{\gamma_j A_j^{(1)} + \omega A_j^{(2)}}{A_j(\gamma_j^2+\omega^2)}\phi_j(a)\right)^2 + \left(\sum_{j=1}^{N} \frac{-\omega A_j^{(1)} + \gamma_j A_j^{(2)}}{A_j(\gamma_j^2+\omega^2)}\phi_j(a)\right)^2},$$

where the infinite series was truncated at N terms. The decay in series coefficients is governed by the magnitude of the eigenvalues γ_j with increasing j. Given that the eigenfunctions $\phi_j(\rho)$ are $O(\gamma_j^{-1/2})$ as $j \to \infty$, it is straightforward to show that terms in the sums of $A(\omega)$ are $O(\gamma_j^{-1})$ as $j \to \infty$. The scaled amplitude $A(\omega)$ was evaluated in MAPLE for 30 frequencies in the range 0–3Hz, which is typical of human joint motion. Numerical convergence of $A(\omega)$ to a tolerance of 10^{-3} was achieved with $N = 100$ terms of the series. In the case $k_C \neq k_P$, $A(\omega)$ was computed by allowing the finite difference method to proceed until the amplitude of the displacement changed by less than 1% between successive periods of the oscillation.

In the case $k_C = k_P$, plots of scaled amplitude $A(\omega)$ versus forcing frequency $f = \omega/(2\pi)$, for the range of parameters considered, are shown in Figure 4.1. When plotted on the same graph, curves obtained using the numerical solution were indistinguishable from those obtained using the analytical series solution. It is observed that a highly stiff PCM layer preserves transmission of a transient-free time-varying

TABLE 4.1
Intervals of the scaled amplitude $A(\omega)$ for frequencies in the range 0–3Hz with varying permeability in normal and OA chondron models.

[k in m^4/N·s]		$k_C = 10^{-12}$	$k_C = 10^{-13}$	$k_C = 10^{-14}$	$k_C = 10^{-15}$
$k_P = 10^{-14}$	Normal	$(0.95, 0.95)$	$(0.95, 0.95)$	$(0.92, 0.95)$	$(0.82, 0.94)$
	OA	$(0.93, 0.94)$	$(0.93, 0.94)$	$(0.88, 0.94)$	$(0.75, 0.92)$
$k_P = 10^{-15}$	Normal	$(0.81, 0.95)$	$(0.81, 0.95)$	$(0.78, 0.95)$	$(0.70, 0.94)$
	OA	$(0.68, 0.94)$	$(0.67, 0.94)$	$(0.64, 0.94)$	$(0.55, 0.92)$
$k_P = 10^{-16}$	Normal	$(0.12, 0.94)$	$(0.12, 0.94)$	$(0.12, 0.93)$	$(0.11, 0.92)$
	OA	$(0.05, 0.90)$	$(0.05, 0.90)$	$(0.05, 0.90)$	$(0.05, 0.88)$

mechanical signal to the chondrocyte. Transmission is excellent for a permeability of 10^{-14}m^4/N·s in that over 90% signal amplitude is retained in the normal model (Figure 4.1(a)), with a slight decrease in transmission for the OA model and little sensitivity to frequency. As the permeability is decreased by one order of magnitude, transmission is moderately reduced, decreases with increasing frequency, and is more sensitive to the reduction in E_P observed for OA human chondrons (Figure 4.1(b)). An analysis of $A(\omega)$, over the full range of permeabilities under consideration, was conducted and is summarized via range intervals for $A(\omega)$ in Table 4.1. In the normal chondron model, it is observed that $A(\omega)$ is in excess of 75% in the range $(k_C, k_P) = (10^{-14}$–$10^{-12}, 10^{-15}$–$10^{-14})$m^4/N·s, and drops significantly as k_P is reduced to 10^{-16}m^4/N·s. In this regime, ranges of $A(\omega)$ are more sensitive to k_P than to k_C. In the range $(k_C, k_P) = (10^{-13}$–$10^{-12}, 10^{-14})$m^4/N·s, the displacement is almost fully transmitted through the PCM and there are no differences in $A(\omega)$ for the OA model. By contrast, when k_P is reduced to 10^{-15}m^4/N·s, a significant drop in $A(\omega)$ is observed for the OA model. These simulations suggest that a PCM permeability that is 1–10 times the permeability of cartilage ECM is consistent with excellent displacement signal transmission through the PCM for normal human chondrons.

4.2.2. Transmission of solid stress. In regimes where displacement signals are well transmitted through the PCM, it is important to evaluate the magnitude of forces on the chondrocyte. A parametric analysis of stress amplitude can indicate the extent to which the PCM serves as a protective layer for the cell. To this end, the displacement solution and resulting pore pressure (3.24) were substituted into the first relation of (2.3) to evaluate the solid stress at the chondron boundary and at the cell-PCM interface. These stress values were used to compute the transient-free solid stress amplitude ratio

$$S(\omega) = \frac{\mathrm{Amp}(\sigma^s_\infty(a,t))}{\mathrm{Amp}(\sigma^s_\infty(b,t))}. \tag{4.8}$$

The independently prescribed part of the pore pressure at the chondron boundary, which would result from transmission of mechanical signals via the ECM to the chondron, was taken to be zero (i.e., $p_0(t) \equiv 0$) in both the numerator and denominator of (4.8). The ratio $S(\omega)$ provides a measure of the extent to which a stress signal that arrives at the chondron is dissipated as it is transmitted through the PCM. The solid stress was chosen as the measure since it is an indicator of forces that are transmitted to intracellular components (e.g., cytoskeleton, nucleus, organelles), which could be adversely affected by excessive forces.

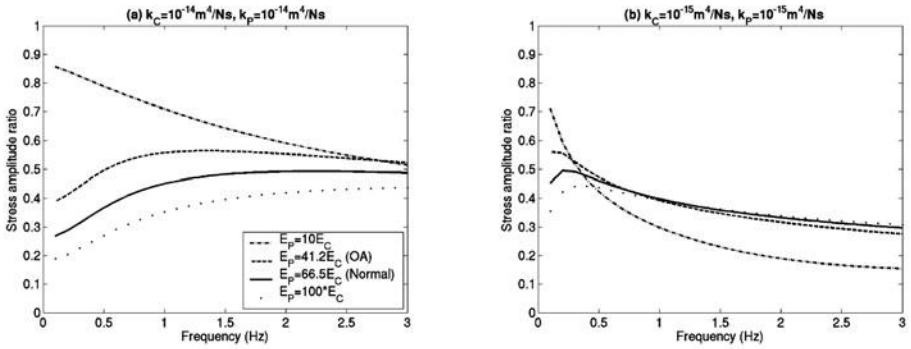


FIG. 4.2. *The solid stress amplitude ratio, $S(\omega)$, for frequencies in the range 0–3Hz in the case $k_C = k_P$:* (a) $k_C = k_P = 10^{-14}, m^4/N{\cdot}s$; (b) $k_C = k_P = 10^{-15} m^4/N{\cdot}s$.

TABLE 4.2
Intervals of the stress amplitude ratio $S(\omega)$ for frequencies in the range 0–3Hz with varying permeability in normal and OA chondron models.

[k in m^4/N·s]		$k_C = 10^{-12}$	$k_C = 10^{-13}$	$k_C = 10^{-14}$	$k_C = 10^{-15}$
$k_P = 10^{-14}$	Normal	(0.23, 0.26)	(0.26, 0.28)	(0.27, 0.49)	(0.47, 0.84)
	OA	(0.25, 0.38)	(0.30, 0.38)	(0.39, 0.56)	(0.64, 0.92)
$k_P = 10^{-15}$	Normal	(0.20, 0.25)	(0.21, 0.25)	(0.23, 0.28)	(0.30, 0.50)
	OA	(0.20, 0.33)	(0.21, 0.33)	(0.22, 0.35)	(0.28, 0.56)
$k_P = 10^{-16}$	Normal	(0.18, 0.21)	(0.18, 0.21)	(0.18, 0.21)	(0.18, 0.26)
	OA	(0.18, 0.21)	(0.18, 0.21)	(0.18, 0.21)	(0.18, 0.26)

In the case $k_C = k_P$, evaluation of $S(\omega)$ involves first-order derivatives of the displacement, and it can be shown that the terms in the sums of $S(\omega)$ are $O(\gamma_j^{-1/2})$ as $j \to \infty$. In contrast to the computation of $A(\omega)$, evaluation of $S(\omega)$ required a larger number of terms ($N = 1000$) to guarantee numerical convergence to a tolerance of 10^{-3}. Plots of the stress amplitude ratio $S(\omega)$ versus forcing frequency, corresponding to those of Figure 4.1, are shown in Figure 4.2. When plotted on the same graph, curves obtained using the analytical and numerical solutions were indistinguishable.

When the chondron permeability is equal to that of cartilage ECM, it is observed that the stress amplitude in the normal chondron model has been reduced to less than 50% of its prescribed value at $\rho = b$ (Figure 4.2(b)). In this case, $S(\omega)$ exhibits only a slight change in the OA chondron model and is rather insensitive to E_P. As the chondron permeability is increased by an order of magnitude, $S(\omega)$ remains less than 0.5 in the normal model but increases significantly in the OA model (Figure 4.2(a)). In this case of higher permeability, further increases in PCM stiffness reduce the stress amplitude ratio. An analysis of $S(\omega)$ corresponding to Table 4.1 was conducted and is summarized in Table 4.2. Our attention is focused on the range $(k_C, k_P) = (10^{-14}$–$10^{-12}, 10^{-15}$–$10^{-14})\mathrm{m}^4/\mathrm{N{\cdot}s}$, where $A(\omega)$ was in excess of 0.75 in the normal chondron model. In this regime, it is observed that, if the case $(k_C, k_P) = (10^{-14}, 10^{-14})\mathrm{m}^4/\mathrm{N{\cdot}s}$ is excluded, then $S(\omega)$ is less than 0.30 for all frequencies, with increases up to 0.38 in the corresponding OA models. As k_P is decreased to $10^{-16}\mathrm{m}^4/\mathrm{N{\cdot}s}$, $S(\omega)$ exhibits

almost no sensitivity to varying k_C. In the range of permeabilities considered, intervals of $S(\omega)$ are more sensitive to k_P than to k_C, though this is less pronounced than was the case for $A(\omega)$.

The transient-free analyses using $A(\omega)$ and $S(\omega)$ indicate that retention of the displacement amplitude at a level of at least 75% occurs in conjunction with transmission of the solid stress amplitude at a level of at most 30% in the ranges of chondron permeability given by $(k_C, k_P) = (10^{-14}\text{–}10^{-12}, 10^{-15})\text{m}^4/\text{N·s}$ and $(k_C, k_P) = (10^{-13}\text{–}10^{-12}, 10^{-14})\text{m}^4/\text{N·s}$. In the range $(k_C, k_P) = (10^{-13}\text{–}10^{-12}, 10^{-14})\text{m}^4/\text{N·s}$, there are only slight differences in the normal and OA models, whereas these differences are more significant in the range $(k_C, k_P) = (10^{-14}\text{–}10^{-12}, 10^{-15})\text{m}^4/\text{N·s}$, particularly for displacement transmission. Overall, the chondron model suggests that the functional role of the PCM as both a transmissive and a protective layer is enhanced in cases where k_C is at least one order of magnitude larger than k_P.

4.2.3. Strain in the chondron. It is known that mechanical loading of a cartilage layer strongly influences the metabolic activity of the chondrocytes [8]. However, little is known regarding the specific components of the local cellular environment that correlate with observed alterations in cell metabolic activity. Based on the hypothesis that strain may play a significant role in this process, the radial strain distribution in the chondron was evaluated. PCM permeability was taken on the order of cartilage ECM permeability (i.e., $k_P = 10^{-15}\text{m}^4/\text{N·s}$) and, based on the analyses of sections 4.2.1 and 4.2.2, two values of chondrocyte permeability, $k_C = 10^{-13}\text{m}^4/\text{N·s}$ and $k_C = 10^{-14}\text{m}^4/\text{N·s}$, were considered.

The amplitude of the transient-free radial strain in the chondron is shown in Figure 4.3 for both the normal and OA models. The case $k_C = k_P = 10^{-15}\text{m}^4/\text{N·s}$ was also included for comparison. In the regime where k_C is at least one order of magnitude greater than k_P (Figures 4.3(a)–4.3(d)), strain amplitudes are consistent with the small strain assumption of linear biphasic theory. As k_C is increased to $10^{-15}\text{m}^4/\text{N·s}$, large strains are present in the cell, near the cell-PCM interface, and the linear biphasic model may be inaccurate. In the normal chondron model, when k_C is at least one order of magnitude larger than k_P (Figures 4.3(a) and 4.3(c)), peak strain amplitudes in the chondrocyte increase with frequency and occur near the interface with the PCM. Strain amplitudes decrease rapidly towards the center of the cell to less than 0.02 and reduce further with increasing frequency. As k_C is increased to $10^{-13}\text{m}^4/\text{N·s}$, strain amplitudes in the entire cell are less than 0.04 and there is less sensitivity to frequency (Figure 4.3(a)), although predictions inside the cell do not accurately reflect the nonhomogeneous nature of the chondrocyte. Strain amplitudes in the PCM are insensitive to k_C, decrease linearly in the direction of the cell, and increase significantly with frequency. In the case $k_C = 10^{-14}\text{m}^4/\text{N·s}$ (Figure 4.3(c)), there is peak-to-peak amplification of the strain amplitude from the PCM boundary to the cell-PCM interface, indicating excellent transmission of radial strain and potential significance as a signalling mechanism. In contrast, for $k_C = 10^{-13}\text{m}^4/\text{N·s}$ (Figure 4.3(a)), the strain amplitude is significantly diminished, particularly at higher frequencies. In the OA chondron model (Figures 4.3(b) and 4.3(d)), where E_P is roughly 38% lower, strain amplitudes are increased in the PCM. In the case $k_C = 10^{-14}\text{m}^4/\text{N·s}$, the peak-to-peak strain amplitude in the OA model drops significantly for higher frequencies, indicating alteration of strain signal transmission with the decreased PCM stiffness associated with OA chondrons (Figure 4.3(d)). As k_C is increased to $10^{-13}\text{m}^4/\text{N·s}$, strain amplitudes in the cell are greatly diminshed and there is little difference between the normal and OA models (Figures 4.3(a) and 4.3(b)).

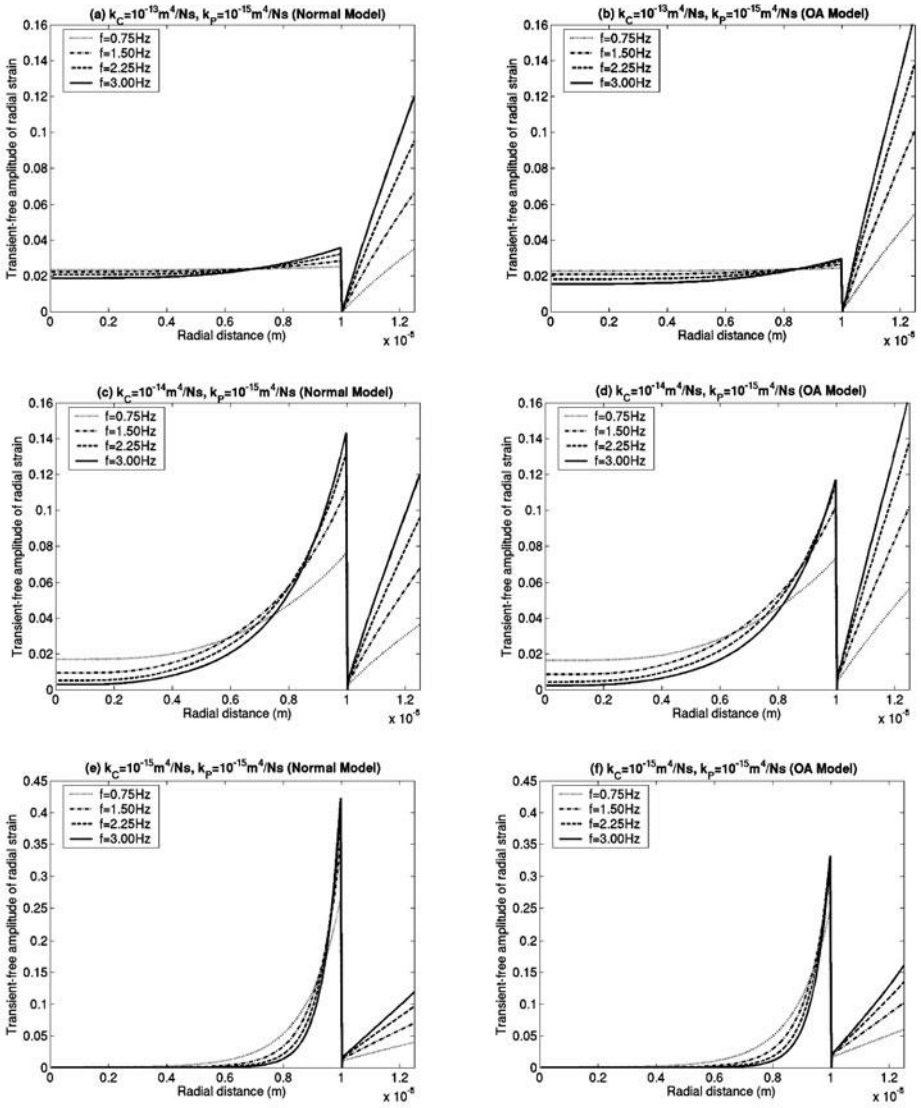


FIG. 4.3. *Amplitude of the transient-free radial strain component in the normal* (a), (c), (e) *and OA* (b), (d), (f) *chondron models. Strain amplitudes, as distributed in the chondron, are shown at four frequencies in the range* 0–3*Hz in the cases* (a) $(k_C, k_P) = (10^{-13}, 10^{-15}) m^4/N{\cdot}s$, $E_P = 66.5kPa$, (b) $(k_C, k_P) = (10^{-13}, 10^{-15}) m^4/N{\cdot}s$, $E_P = 41.2kPa$, (c) $(k_C, k_P) = (10^{-14}, 10^{-15}) m^4/N{\cdot}s$, $E_P = 66.5kPa$, (d) $(k_C, k_P) = (10^{-14}, 10^{-15}) m^4/N{\cdot}s$, $E_P = 41.2kPa$, (e) $(k_C, k_P) = (10^{-15}, 10^{-15}) m^4/N{\cdot}s$, $E_P = 66.5kPa$, (f) $(k_C, k_P) = (10^{-15}, 10^{-15}) m^4/N{\cdot}s$, $E_P = 41.2kPa$. *The cell and chondron radii are* $a = 10\mu m$ *and* $b = 12.5\mu m$, *respectively.*

This analysis of strain amplitude in the chondron suggests that while the PCM serves to significantly reduce transmission of solid stress to the cell, transmission of radial strain is excellent in normal chondrons when k_C is one order of magnitude larger than a value of k_P similar to that of cartilage ECM. For these values of permeability, the analysis also suggests that transmission of radial strain decreases in the presence of the lower PCM stiffness associated with OA, particularly at higher frequencies.

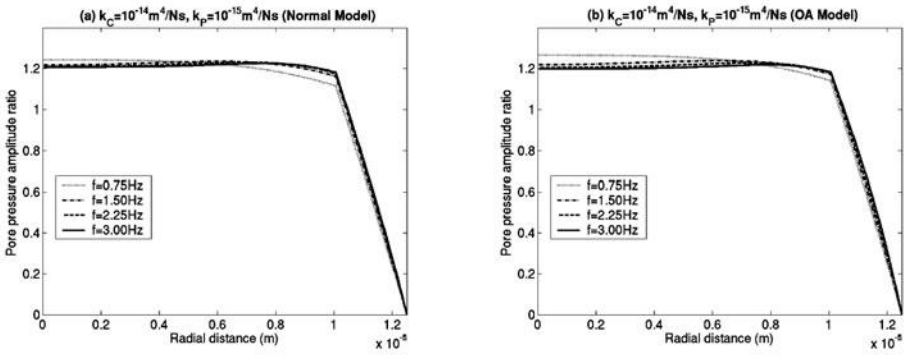


FIG. 4.4. *Pore pressure amplitude ratio for the normal* (a) *and OA* (b) *chondron models in the case* $(k_C, k_P) = (10^{-14}, 10^{-15})m^4/N{\cdot}s$ *for frequencies in the range* $0 - 3Hz$. *The cell and chondron radii are* $a = 10\mu m$ *and* $b = 12.5\mu m$, *respectively.*

Fluid stress in the chondron was evaluated for $(k_C, k_P) = (10^{-14}, 10^{-15})\mathrm{m}^4/\mathrm{N{\cdot}s}$, via the pore pressure amplitude ratio $\mathrm{Amp}(p_\infty(a,t))/\mathrm{Amp}(\sigma^s_\infty(b,t))$ (Figure 4.4). It is observed that this ratio is nearly uniform inside the cell, as was the pore pressure itself (not shown), and does not change significantly in the OA model, or with increasing frequency. The pore pressure amplitude ratio indicates that the transient-free fluid stress $\sigma^f_\infty = -(1-\phi)p$ inside the cell is roughly on the same scale as the solid stress at the outer PCM boundary. The large difference in spatial gradients of the pore pressure amplitude ratio between the cell and PCM has a less pronounced effect on fluid flow in the chondron since k_C is one order of magnitude larger than k_P (see (2.5)). The isolated effects of fluid stress on chondrocyte metabolic activity are not known. However, given that forces and deformation of the intracellular organelles are determined by the solid stress and strain in the biphasic model, these solid phase variables are more likely to serve as predictors of mechanical effects on chondrocyte metabolism.

5. Summary. The presence of a thin, highly stiff and less permeable PCM layer surrounding an articular chondrocyte enhances the transmission of displacement and strain signals from the ECM to the cell while simultaneously protecting the cell from excessive solid stress. The amplitude of these mechanical signals is rather sensitive to decreasing stiffness of the PCM, indicating that mechanical signal transmission may be significantly altered in the presence of OA. The optimal mechanical environment of a healthy chondrocyte is a complex function of the geometric and material properties of the cell, the PCM, and the biomechanical interface between these two regions. A more realistic chondron model would incorporate effects of the lipid bilayer and mechanical connections between the cell and PCM. Such a model would require measurements of material parameters for these subcomponents of the cell and its interface, and once such measurements are available, model predictions could provide a more precise picture of mechanical signal transduction to the cell. A triphasic model of chondron mechanics, incorporating physicochemical effects such as the fixed-charge density in the PCM, can also be considered. Ultimately, chondron mechanics should be coupled to multiphasic macroscopic models for deformation of the ECM. Such models will enable a description of relationships between external dynamic loading of a tissue layer and mechanical signals at the cellular scale. This modeling will be facilitated by further experimental studies that aim to determine geometric and material properties of the chondron, and variations in these properties with site, species, and disease.

Appendix A. Demonstration that L is self-adjoint in the case $k_C = k_P$. Define the inner product $\langle F, G\rangle \equiv \int_0^b F(\rho)\overline{G(\rho)}\rho^2 d\rho$. It is shown that when $k_C = k_P$ and the interface conditions (3.4) are satisfied, then $\langle L\phi, \phi\rangle = \langle \phi, L\phi\rangle$ (i.e., the operator L is self-adjoint).

Consider

$$\langle L\phi, \phi\rangle = \int_0^a \left[k_C H_C^A(\partial_\rho(\rho^2\partial_\rho\phi) - 2\phi)\right]\overline{\phi}d\rho + \int_a^b \left[k_P H_P^A(\partial_\rho(\rho^2\partial_\rho\phi) - 2\phi)\right]\overline{\phi}d\rho.$$

Integrating once by parts,

$$\langle L\phi, \phi\rangle = a^2\overline{\phi}(a)(k_C H_C^A\phi^{'}(a^-) - k_P H_P^A\phi^{'}(a^+)) - k_C H_C^A\int_0^a \rho^2\partial_\rho\phi\partial_\rho\overline{\phi}d\rho$$
$$-k_P H_P^A\int_a^b \rho^2\partial_\rho\phi\partial_\rho\overline{\phi}d\rho - 2k_C H_C^A\int_0^a \phi\overline{\phi}d\rho - 2k_P H_P^A\int_a^b \phi\overline{\phi}d\rho,$$

where the first relation in (3.4) has been employed along with the assumptions that ϕ and $\phi^{'}$ are bounded at $\rho = 0$. A second integration-by-parts yields

$$\langle L\phi, \phi\rangle =$$
$$a^2\overline{\phi}(a)(k_C H_C^A\phi^{'}(a^-) - k_P H_P^A\phi^{'}(a^+)) + a^2\phi(a)(k_P H_P^A\overline{\phi}^{'}(a^+) - k_C H_C^A\overline{\phi}^{'}(a^-))$$
$$+k_C H_C^A\int_0^a \phi(\rho)(\partial_\rho(\rho^2\partial_\rho\overline{\phi}) - 2\overline{\phi})d\rho + k_P H_P^A\int_a^b \phi(\rho)(\partial_\rho(\rho^2\partial_\rho\overline{\phi}) - 2\overline{\phi})d\rho,$$

which simplifies to

$$\langle L\phi, \phi\rangle = \langle \phi, L\phi\rangle + a^2\overline{\phi}(a)(k_C H_C^A\overline{\phi}^{'}(a^-) - k_P H_P^A\phi^{'}(a^+))$$
$$+a^2\phi(a)(k_P H_P^A\overline{\phi}^{'}(a^+) - k_C H_C^A\overline{\phi}^{'}(a^-)).$$

Using the last relation of (3.4), $\phi^{'}(a^+)$ can be eliminated and the expression reduces to

$$\langle L\phi, \phi\rangle = \langle \phi, L\phi\rangle + a^2 H_C^A(k_C - k_P)(\overline{\phi}(a)\phi^{'}(a^-) - \phi(a)\overline{\phi}^{'}(a^-)).$$

Since $k_C = k_P$, $\langle L\phi, \phi\rangle = \langle \phi, L\phi\rangle$ and the operator L is self-adjoint.

Appendix B. Finite difference scheme. Let $u_i^j = u(\rho_i, t_j)$, where $\rho_i = i\Delta\rho$, $t_j = j\Delta t$, and $\Delta\rho = b/M$. Let K denote the index of the mesh point to the immediate left of the cell-PCM interface at $\rho = a$. Equation (3.27) was discretized using a first-order finite difference in time and second-order finite differences in space to obtain the following implicit scheme at time step j. For $i = 1, \ldots, K-1, K+2, \ldots, M-1$,

$$\text{(B.1)}\qquad \frac{u_i^{j+1} - u_i^j}{\Delta t} = r\left(\frac{u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}}{(\Delta\rho)^2} + \frac{2}{\rho_i}\frac{u_{i+1}^{j+1} - u_{i-1}^{j+1}}{2\Delta\rho} - \frac{2}{\rho_i^2}u_i^{j+1}\right).$$

For $i = K$, the first condition in (3.10) was enforced via $u_{K+1}^{j+1} = u_K^{j+1}$. For $i = K+1$, the second condition in (3.10) was discretized using first-order finite difference approximations in space to obtain

$$\text{(B.2)}\qquad H_P^A\left(\frac{u_{K+2}^{j+1} - u_{K+1}^{j+1}}{\Delta\rho}\right) - H_C^A\left(\frac{u_K^{j+1} - u_{K-1}^{j+1}}{\Delta\rho}\right) = \frac{2}{a}\left(\lambda_C u_K^{j+1} - \lambda_P u_{K+1}^{j+1}\right).$$

At $i = 0$ and $i = M$ the boundary conditions at (3.11) were enforced. Marching in time via j, the finite difference equations were assembled at each time step into a linear algebraic system that was solved for each set of parameters using MATLAB 6.5.

Acknowledgments. The author would like to thank F. Guilak and M. Shearer for helpful discussions.

REFERENCES

[1] L. G. Alexopoulos, M. A. Haider, T. P. Vail, and F. Guilak, *Alterations in the mechanical properties of the human chondrocyte pericellular matrix with osteoarthritis*, J. Biomech. Engrg., 125 (2003), pp. 323–333.

[2] K. A. Athanasiou, M. P. Rosenwasser, J. A. Buckwalter, T. I. Malinin, and V. C. Mow, *Interspecies comparisons of in situ intrinsic mechanical properties of distal femoral cartilage*, J. Orthop. Res., 9 (1991), pp. 330–340.

[3] R. M. Bowen, *Incompressible porous media models by use of the theory of mixtures*, Internat. J. Engrg. Sci., 18 (1980), pp. 1129–1148.

[4] J. M. Cooper, *Introduction to Partial Differential Equations with MATLAB*, Birkhäuser Boston, Boston, MA, 1998.

[5] W. L. Dunbar, K. Un, P.S. Donzelli, and R. L. Spilker, *An evaluation of three-dimensional diarthrodial joint contact using penetration data and the finite element method*, J. Biomech. Engrg., 123 (2001), pp. 333–340.

[6] F. Guilak and V. C. Mow, *The mechanical environment of the chondrocyte: A biphasic finite element model of cell-matrix interactions in articular cartilage*, J. Biomech., 33 (2000), pp. 1663–1673.

[7] F. Guilak, A. Ratcliffe, and V. C. Mow, *Chondrocyte deformation and local tissue strain in articular cartilage: A confocal microscopy study*, J. Orthop. Res., 13 (1995), pp. 410–421.

[8] F. Guilak, R. L. Sah, and L. A. Setton, *Physical regulation of cartilage metabolism*, in Basic Orthopaedic Biomechanics, V. C. Mow and W. C. Hayes, eds., Lippincott-Raven, Philadelphia, 1997, pp. 179–207.

[9] M. A. Haider and F. Guilak, *An axisymmetric boundary integral model for assessing elastic cell properties in the micropipette aspiration test*, J. Biomech. Engrg., 124 (2002), pp. 586–595.

[10] J. S. Hou, M. H. Holmes, W. M. Lai, and V. C. Mow, *Boundary conditions at the cartilage-synovial fluid interface for joint lubrication and theoretical verifications*, J. Biomech. Engrg., 111 (1989), pp. 78–87.

[11] J. C. Iatridis, L. A. Setton, R. J. Foster, B. A. Rawlins, M. Weidenbaum, and V. C. Mow, *Degeneration affects the anisotropic and nonlinear behaviors of human annulus fibrosis in compression*, J. Biomech., 31 (1998), pp. 535–544.

[12] W. R. Jones, H. P. Ting-Beall, G. M. Lee, S. S. Kelley, R. M. Hochmuth, and F. Guilak, *Alterations in the Young's modulus and volumetric properties of chondrocytes isolated from normal and osteoarthritic human cartilage*, J. Biomech., 32 (1998), pp. 119–127.

[13] W. M. Lai and V. C. Mow, *Drag-induced compression of articular cartilage during a permeation experiment*, Biorheology, 17 (1980), pp. 111–123.

[14] A. F. T. Mak, D. T. Huang, J. D. Zhang, and P. Tong, *Deformation-induced hierarchial flows and drag forces in bone canaliculi and matrix microporosity*, J. Biomech., 20 (1997), pp. 11–18.

[15] V. C. Mow, N. Bachrach, L. A. Setton, and F. Guilak, *Stress, strain, pressure and flow fields in articular cartilage*, in Cell Mechanics and Cellular Engineering, V. C. Mow et al., eds., Springer-Verlag, New York, 1994, pp. 345–379.

[16] V. C. Mow, M. H. Holmes, and W. M. Lai, *Fluid transport and mechanical properties of articular cartilage: A review*, J. Biomech., 17 (1984), pp. 377–394.

[17] V. C. Mow, J. S. Hou, J. M. Owens, and A. Ratcliffe, *Biphasic and quasilinear viscoelastic theories for hydrated soft tissues*, in Biomechanics of Diarthroidal Joints, V. Mow et al., eds., Springer-Verlag, New York, 1990, pp. 215–260.

[18] V. C. Mow, S. C. Kuei, W. M. Lai, and C. G. Armstrong, *Biphasic creep and stress relaxation of articular cartilage in compression: Theory and experiments*, J. Biomech. Engrg., 102 (1980), pp. 73–84.

[19] D. Shin and K. Athanasiou, *Cytoindentation for obtaining cell biomechanical properties*, J. Orthop. Res., 17 (1999), pp. 880–890.

[20] R. L. Spilker, P. S. Donzelli, and V. C. Mow, *A transversely isotropic biphasic finite-element model of the meniscus*, J. Biomech., 112 (1992), pp. 1027–1045.
[21] R. A. Stockwell, *Biology of Cartilage Cells*, Cambridge University Press, Cambridge, UK, 1979.
[22] R. A. Stockwell, *Structure and function of the chondrocyte under mechanical stress*, in Joint Loading: Biology and Health of Articular Structures, H. J. Helminen et al., eds., Wright and Sons, Bristol, UK, 1987, pp. 126–148.
[23] D. P. Theret, M. J. Levesque, M. Sato, R. M. Nerem, and L. T. Wheeler, *The application of a homogeneous half-space model in the analysis of endothelial cell micropipette measurements*, J. Biomech. Engrg., 110 (1988), pp. 190–199.
[24] W. R. Trickey, G. M. Lee, and F. Guilak, *Viscoelastic properties of chondrocytes from normal and osteoarthritic human cartilage*, J. Orthop. Res., 18 (2000), pp. 891–898.
[25] C. Truesdell and R. Toupin, *The classical field theories*, in Handbuch der Physik, S. Flugge, ed., Springer-Verlag, Berlin, 1960, pp. 226–793.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1609–1635

SUSTAINED SPATIAL PATTERNS OF ACTIVITY IN NEURONAL POPULATIONS WITHOUT RECURRENT EXCITATION*

JONATHAN E. RUBIN† AND WILLIAM C. TROY†

Abstract. Spatial patterns of neuronal activity arise in a variety of experimental studies. Previous theoretical work has demonstrated that a synaptic architecture featuring recurrent excitation and long-range inhibition can support sustained, spatially patterned solutions in integrodifferential equation models for activity in neuronal populations. However, this architecture is absent in some areas of the brain where persistent activity patterns are observed. Here we show that sustained, spatially localized activity patterns, or bumps, can exist and be linearly stable in neuronal population models without recurrent excitation. These models support at most one bump for each background input level, in contrast to the pairs of bumps found with recurrent excitation. We explore the shape of this bump as well as the mechanisms by which this bump is born and destroyed as background input level changes. Further, we introduce spatial inhomogeneity in coupling and show that this induces bump pinning: for a given starting position, bumps can exist only for a small, discrete set of background input levels, each with a unique corresponding bump width.

Key words. neuronal population, spatial pattern, localized activity bump, off-center coupling, bifurcation, spatial inhomogeneity

AMS subject classifications. 34B15, 34C11, 34C23, 34C37, 93C15

DOI. 10.1137/S0036139903425806

1. Introduction.

Evidence suggests that sustained, spatially patterned neuronal activity may play a role in short-term encoding of information. For example, localized persistent activity, or bumps, may provide the basis for a working memory of external stimulus features [14, 7, 18] or a representation of internal states such as head direction (reviewed in [25, 21]). Previous theoretical works have explored the ways in which a network of spiking neurons with short-range recurrent excitation (i.e., positive local coupling) and long-range inhibition can support sustained, spatially organized activity [28, 1, 15, 13, 8, 17, 16, 4, 5]. These studies focus on various forms of rate or activity models, in which a single equation encapsulates the temporal evolution of some measure of the activity level of an entire population of spiking neurons (i.e., neurons firing regularly with some average spike rate). A related result shows that when timescales of synaptic dynamics are taken into account in a conductance-based network model, sustained, localized activity can arise in a two-layer network of *bursting* thalamic cells that lacks recurrent excitation [20]. This leads naturally to the fundamental question of just how crucial the presence of recurrent excitation is for the existence of sustained spatial patterns of activity in rate or activity models of populations of *spiking* neurons. This paper shows that spatially localized activity can be sustained in a neuronal network without recurrent excitation or bursting mechanisms.

We consider a rate model of the form

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

*Received by the editors April 4, 2003; accepted for publication (in revised form) November 5, 2003; published electronically June 22, 2004.
http://www.siam.org/journals/siap/64-5/42580.html

†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (rubin@math.pitt.edu, troy@math.pitt.edu). The first author's research was partially supported by NSF grant DMS-0108857. The first author is a member of the Center for the Neural Basis of Cognition.

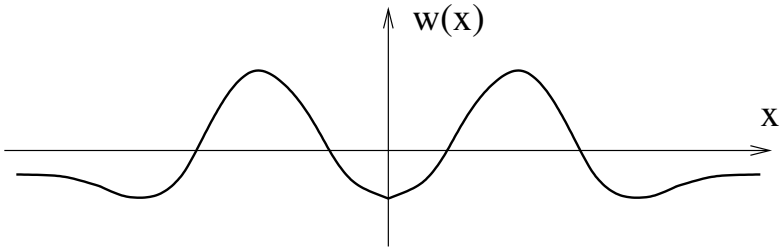


FIG. 1. *Off-center coupling function for an excitatory population. Cells at any position x with $w(x) > 0$ receive excitatory input from cells at position $x = 0$, while cells at x with $w(x) < 0$ receive inhibitory input.*

Equation (1.1) models a single population of spiking neurons. The function $u(x,t)$ encodes the activity level, or average voltage, of a neuronal subgroup at position $x \in (-\infty, \infty)$ and time $t \geq 0$. The connection function $w(x)$ determines the coupling between subgroups, and the nonnegative, nondecreasing function $f(u)$ denotes the neuronal firing rate, or average rate at which spikes are generated, corresponding to an activity level u. Neurons at a point x are said to be active if $f(u(x,t)) > 0$. Finally, the parameter h encodes a constant external stimulus applied uniformly to the entire neural field [1], such as an average background input level received from other areas of the brain, and the parameter σ denotes a positive rate constant; the ratio h/σ represents the baseline level of activity in the population without coupling. Without loss of generality, we set $\sigma = 1$.

In this paper, we take $f(u(x,t)) = H(u(x,t))$, the Heaviside step function, which gives

$$\frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{-\infty}^{\infty} w(x-y)H(u(y,t))\,dy + h. \tag{1.2}$$

For the Heaviside form of firing rate function, the activity level $u = 0$ represents an absolute threshold for synaptic input required to drive spiking activity. This form of (1.1) was considered in [1] and by many subsequent authors. Further, it was shown in [15] that results for (1.2) are crucial in determining solution structure for (1.1) with more general nondecreasing f.

After we detail additional assumptions on the model, we prove that recurrent excitation is not necessary for the existence of stable stationary, spatially localized solutions (i.e., bumps) in populations of spiking cells. The synaptic architecture that we consider, as an alternative to recurrent excitation, takes the form shown in Figure 1. Such an *off-center architecture* may be relevant in several different contexts. For example, consider a network featuring interconnected excitatory (E) and inhibitory (I) populations of cells, in which E-cells are intrinsically capable of spiking and I-cells inhibit both E-cells and other I-cells. In such a network, activity of E-cells leads to activity of corresponding I-cells. This may lead to feedback inhibition onto the active E-cells as well as inhibition of nearby I-cells. This I-I inhibition can in turn disinhibit nearby E-cells, effectively acting as an off-center form of excitation onto E-cells, as portrayed in Figure 1. This form of architecture may arise in interactions of the subthalamic nucleus (E) and external segment of the globus pallidus (I) in the primate basal ganglia [22, 26]. It also may occur in interactions of thalamocortical

relay cells (E) and thalamic reticular cells (I) in the thalamus in awake states, where activity bumps in certain subpopulations of cells can encode head direction [23, 24]. Long-range inhibitory connections have been found in the thalamus [9, 10], but no thalamic recurrent excitatory connections are known to exist. In the first subsection of the appendix, we discuss the derivation of the effective coupling shown in Figure 1 from coupled E and I populations modelled by a pair of equations of the form (1.1), as in [28, 1, 19] but without recurrent excitation, although a rigorous mathematical derivation, or a complete mathematical treatment of the coupled equations, remains for future research. Alternatively, this form of suppression of recurrent excitation by localized inhibitory feedback may be generated by inhibitory interneurons in a variety of cortical areas, or in the CA1 region of the hippocampus, which features at most sparse recurrent excitatory connections [6, 27].

We consider this coupling architecture in section 2. Under certain assumptions, we prove the existence of a bump solution $u(x)$ to (1.2) such that $u(x) > 0$ if and only if $x \in (0, a)$ for a fixed constant a. We also show that this bump solution is linearly stable when it exists. Our proof method for existence generalizes that given by Amari [1] for bumps in (1.2) with lateral inhibition. However, details become much more subtle due to the more complicated synaptic architecture that we consider.

Unlike the case with recurrent excitation, where two bump solutions exist [1], the bump of localized positive activity that we find is unique for each fixed h in some finite interval. We show how the shape of a bump depends on its size, which in turn depends on h, relative to certain features of the coupling function $w(x)$. Further, as h varies, bump solutions (parametrized by h) are created and destroyed by atypical mechanisms that do not involve saddle-node bifurcations (since only a single solution exists for each h) or the entire bump collapsing to 0, and we explain the possible mechanisms and how they are selected.

For consideration of stability, we deviate from [1] to give a rigorous linear stability calculation. A simplified version of the calculation shows that a spatially uniform state can also be stable, such that the system exhibits bistability, consistent with [20, 24].

It has been argued that coupling strengths between neurons should not be purely distance dependent but rather should allow for spatial variation [2, 3]. In section 3, we introduce spatial inhomogeneity in coupling, replacing $w(x-y)$ by $w(x-y)p(y)$ under the integral in equation (1.2). We set up equations relevant to bump existence in this case, which we treat through a combination of analysis and numerics. The presence of spatial inhomogeneity naturally destroys the translation invariance of bumps. In fact, we find that it induces a form of bump pinning, such that for a given starting position, bumps exist for only a small, discrete set of background input levels, each with a unique corresponding bump width. Interestingly, in our primary numerical example, we find that there is a special bump width which is possible for any starting position. We comment on possible functional implications of these results in the discussion in section 4.

2. Spatially homogeneous coupling.

2.1. Assumptions.

In this section, we consider (1.2)

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

with a coupling function $w(x)$ satisfying the hypothesis

($H1$) $w(x)$ is continuous and integrable on $\mathbb{R}$ and is symmetric; i.e., $w(-x) = w(x)$ for all $x \in \mathbb{R}$.

Moreover, we assume that there exist constants $x_* > x_1 > x^* > x_0 > 0$ such that

($H2$) $w(x) < 0$ on $(-x_0, x_0)$ and on (x_1, ∞), with $w(x_0) = w(x_1) = 0$;

($H3$) $w(x)$ is increasing on $(0, x^*)$ and on (x_*, ∞);

($H4$) $w(x) > 0$ on (x_0, x_1);

($H5$) $w(x)$ is decreasing on (x^*, x_*).

Coupling functions that satisfy ($H2$) and ($H4$) are sometimes called off-center coupling functions.

To simplify notation, we will also assume the following symmetry hypothesis in certain cases noted below

($H6$) there exists $\delta > 0$ such that $x_1 = x^* + \delta$ and $x_0 = x^* - \delta$, and $w(x^* + \eta) = w(x^* - \eta)$ for all $\eta \in [0, \delta]$.

We will comment further on the role of this hypothesis in Remark 2.8 after the proof of Theorem 2.5. A coupling function $w(x)$ satisfying ($H1$)–($H6$) appears in Figure 2.

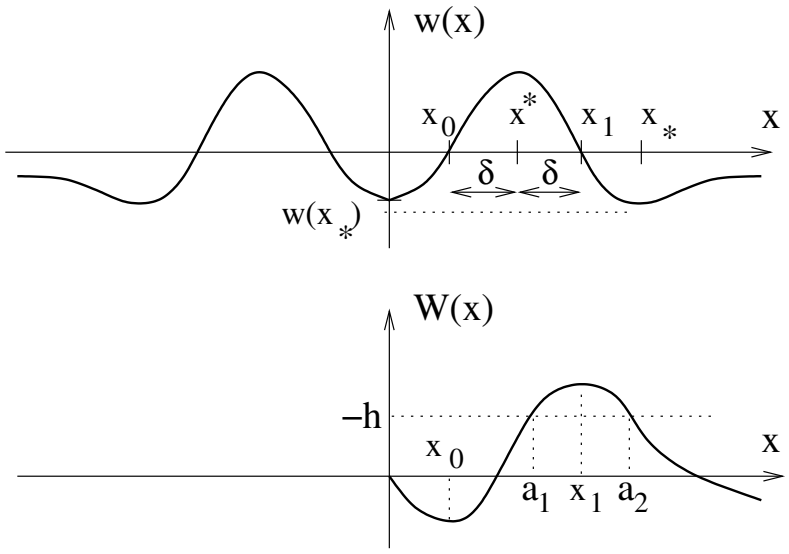


FIG. 2. *Off-center coupling function $w(x)$, together with antiderivative $W(x)$. From the plot of $W(x)$, we can visualize the necessary condition* (2.3)*, $W(a) + h = 0$, for bump existence.*

2.2. Existence of a unique bump. Following Amari [1], we seek stationary bump solutions $u(x)$ to (1.2), for which $u(x) > 0$ if and only if $x \in (0, a)$ for some constant a. Note that this is equivalent to the existence of a bump on any other interval of length a; that is, the system is translation invariant. Such solutions satisfy the condition

$$u(x) = \int_0^a w(x - y)\, dy + h\,. \tag{2.1}$$

Since $\lim_{x\to\infty} w(x) = 0$, the fact that $\lim_{x\to\infty} u(x) \le 0$ for such a solution requires that $h \le 0$. Thus, we impose the existence condition that

($E1$) the constant $h \le 0$.

Let $W(x) = \int_0^x w(t)\,dt$, which is an odd function. Then (2.1) becomes

$$u(x) = W(x) - W(x-a) + h. \tag{2.2}$$

From (2.2), the conditions $u(0) = 0$ and $u(a) = 0$ both give

$$W(a) + h = 0. \tag{2.3}$$

Figure 2 displays a graphical representation of equation (2.3). Note that for fixed $h < 0$, there exist either zero or two solutions of (2.1), unless $-h = W(x_1)$.

We will require a second existence condition, namely, that

($E2$) $W(x) + h > 0$ for some $x \in \mathbb{R}^+$ and $\lim_{x\to\infty} W(x) < -h$.

For fixed h such that conditions ($E1$) and ($E2$) hold, there exist two nonzero solutions of (2.3). We label these solutions as a_1 and a_2, where $a_2 > x_1 > a_1 > 0$. Note that a_2 could fail to exist without ($E2$), if $\lim_{x\to\infty} W(x) > -h > 0$. Let $u_1(x), u_2(x)$ denote the corresponding functions defined by $u_i(x) = W(x) - W(x-a_i) + h$. Note that $u_i(x) = W(x) - W(x-a_i) + h = W(a_i - x) + W(x) + h = u(a_i - x)$ for all x. This yields the following symmetry statement.

PROPOSITION 2.1. *Each solution $u_i(x)$ of* (2.3) *is symmetric about $x = a_i/2$.*

With these definitions, we state one final hypothesis, as an alternative to ($H6$), that will be assumed when noted below:

($H6'$) $W(x_0) - W(a_2) + W(a_2 - a_1) > 0$.

PROPOSITION 2.2. *Assume ($H1$)–($H5$). Fix h such that ($E1$)–($E2$) hold and $a_1 < x_1$ such that $W(a_1) + h = 0$. The function $u_1(x)$ defined by* (2.1) *with $a = a_1$ does not represent a valid bump solution to* (1.2). *In fact, if we assume ($H6$) as well, then $u_1(x) < 0$ on all of $(0, a_1)$.*

Proof. By construction, $u_1(0) = 0$ and $x_0 < a_1 < x_1$. Note that $u_1'(x) = w(x) - w(a_1 - x)$. Since $a_1 < x_1$, it follows that $w(a_1) > w(0)$. Thus, $u_1'(0) = w(0) - w(a_1) < 0$. This establishes that $u_1(x)$ is not a valid bump.

Further, note that $u_1'(x) = 0$ requires $w(x) = w(a_1 - x)$. This occurs at $x = a_1/2$, consistent with the symmetry of $u_1(x)$ about $x = a_1/2$. However, we shall see that when ($H6$) holds, the equation $u_1'(x) = 0$ has no other solutions in $(0, a_1)$, proving the proposition. To see this, note that by ($H6$), if there exists $x \neq a_1/2$ in $(0, a_1)$ such that $u_1'(x) = 0$, then $(a_1 - x) - x^* = x^* - x$, or, equivalently, $a_1 = 2x^*$. But $a_1 < x_1 < x_1 + x_0 = 2x^*$, so this is not possible. □

Remark 2.3. Proposition 2.2 implies that a bump can only possibly exist when there exists $a_2 > x_1$ for which (2.3) holds. Thus, if $\lim_{x\to\infty} W(x) > -h$, in violation of ($E2$), then no bump exists.

Now, define the constant A as the smallest positive x value for which $W(x) = 0$, guaranteed to exist by ($E2$) since $h \leq 0$.

PROPOSITION 2.4. *Assume ($H1$)–($H5$). Fix h such that ($E1$)–($E2$) hold. If $0 < a_2 - a_1 < A$, then the function $u_2(x)$, defined by* (2.1) *with $a = a_2$, does not represent a valid bump solution to* (1.2).

Proof. We compute directly from equation (2.2) that

$$u_2(a_1) = W(a_1) + W(a_2 - a_1) - W(a_1) = W(a_2 - a_1)\,.$$

If $a_2 - a_1 < A$, then $u_2(a_1) < 0$. But $a_1 \in (0, a_2)$, so $u_2(x)$ is not a bump. □

Next we establish some results showing the existence of a valid bump in various cases. Note that if $\lim_{x\to\infty} W(x) = 0$ (see Figure 2), then we can make a_2 arbitrarily large by choosing h sufficiently close to 0. If $a_2/2 > x_1$, then $a_2 - a_1 > 2x_1 - a_1 >$

$a_1 > A$. Thus, when $a_2/2 > x_1$, Proposition 2.4 does not apply, and $u_2(a_1) > 0$. In fact, in this case, we can establish the existence of a bump without hypothesis $(H6)$ or $(H6')$, as addressed below in Theorem 2.5. However, as we shall see in subsection 2.4, a_2 can also become too large for a bump to exist. This motivates a final existence condition,

$(E3)$ $w(a_2 \pm x_0) < w(0)$.

As we make h more negative, such that $-h$ grows toward the peak of W, we will also need to impose $(H6)$ or $(H6')$ to ensure the existence of a bump.

THEOREM 2.5. *Assume that $(H1)$–$(H5)$ hold. Fix h such that $(E1)$–$(E3)$ hold and assume that $a_2/2 > x_1$. Then the function $u_2(x)$ defined by (2.1) with $a = a_2$, such that $-W(a_2) = h$, is a bump solution to (1.2), with $u_2(x) > 0$ if and only if $x \in (0, a_2)$.*

Proof. First, recall that $a_2 > x_1$, since $W(a_2) = -h > 0$ and $W'(a_2) = w(a_2) < 0$ (see Figure 2). Further, since $w(0) < 0$ and $w(x_1) = 0$, the hypotheses of the theorem together with $(H2)$–$(H5)$ imply that there exist exactly two positive values $x'' > x' > x_1$ such that $w(0) = w(x') = w(x'')$, and $a_2 \in (x' + x_0, x'' - x_0)$; see Figure 3. This gives $w(0) > w(a_2)$ and $a_2/2 > x_0$.

We now show that $u_2(x) > 0$ for $x \in (0, a_2)$. First, we consider $x \in (0, x_0]$. Note that

$$u_2'(x) = w(x) - w(a_2 - x) \tag{2.4}$$

from (2.2). In particular, $u_2'(0) = w(0) - w(a_2) > 0$. Since $a_2 > x' + x_0$, we have $a_2 - x > x'$ for all $x \in (0, x_0]$. Thus, $w(a_2 - x) < w(x)$ and $u_2'(x) > 0$ for all $x \in (0, x_0]$.

Next, suppose that $a_2/2 > x_1$. Now, $u_2'(x) = w(x) - w(a_2 - x) > 0$ on $(x_0, x_1]$ as well. This holds because on $(x_0, x_1]$, $w(x) \geq 0$, while $a_2 - x > a_2 - x_1 > x_1$, such that $w(a_2 - x) < 0$ by $(H2)$. It remains to show that $u_2(x) > 0$ for all $x \in (x_1, a_2)$. To do this, it suffices to show that $u_2(x) > 0$ for all $x \in (x_1, a_2/2]$, since symmetry (Proposition 2.1) then gives $u_2(x) > 0$ for all $x \in (0, a_2)$.

Equation (2.2) can be rewritten for $u = u_2(x)$ as

$$u_2(x) = (W(x) - W(a_2)) + W(a_2 - x), \tag{2.5}$$

using equation (2.3) and the fact that $W(x)$ is odd. We will show that $u_2(x)$ in (2.5) is the sum of two positive terms for $x \in (x_1, a_2/2]$. When $x_1 < x \leq a_2/2$, it follows that $a_2 - x_1 > a_2 - x \geq a_2/2 > x_1$. By construction, $W(x) > 0$ on $[x_1, a_2]$. Hence, $W(a_2 - x) > 0$ for all $x \in (x_1, a_2/2]$. Now, consider the first term in $u_2(x)$ in (2.5), namely, $W(x) - W(a_2)$, for $x \in (x_1, a_2/2]$. Since $W'(x) = w(x) < 0$ on (x_1, ∞) by $(H2)$, and $x_1 < x \leq a_2/2 < a_2$, we have $W(x) > W(a_2)$ for all $x \in (x_1, a_2/2]$. Thus, $u_2(x)$ is the sum of two positive terms, and hence is positive, on $(x_1, a_2/2]$, as claimed. This gives $u_2(x) > 0$ on $(0, a_2)$, as desired.

Finally, to complete the proof, we confirm that $u_2(x) < 0$ for all $x \in (a_2, \infty)$. Note that since $W(a_2) = -h > 0$, $u_2(x) < 0$ is equivalent to $W(x) < W(a_2) + W(x - a_2)$ by equation (2.3). Since $W(a_2) > 0$, we have $u_2(x) < 0$ for all x such that $W(x - a_2) \geq W(x)$. Since $W(x)$ decreases on (x_1, ∞), this implies that $u_2(x) < 0$ for $x - a_2 \geq x_1$, namely, $x \in [a_2 + x_1, \infty)$. Thus, it remains to consider $x \in (a_2, a_2 + x_1)$. We will show that $u_2'(x) < 0$ on $(a_2, a_2 + x_1)$, for $u_2'(x)$ given in (2.4), such that $u_2(x)$ remains negative there. Since $w(x - a_2) > 0$ on $(a_2 + x_0, a_2 + x_1)$ by $(H4)$, while $w(x) < 0$ on this interval by $(H2)$, it is obvious from (2.4) that $u_2'(x) < 0$ on $(a_2 + x_0, a_2 + x_1)$. It remains only to consider $x \in (a_2, a_2 + x_0]$. Condition $(E3)$ gives $w(a_2) < w(0)$, so

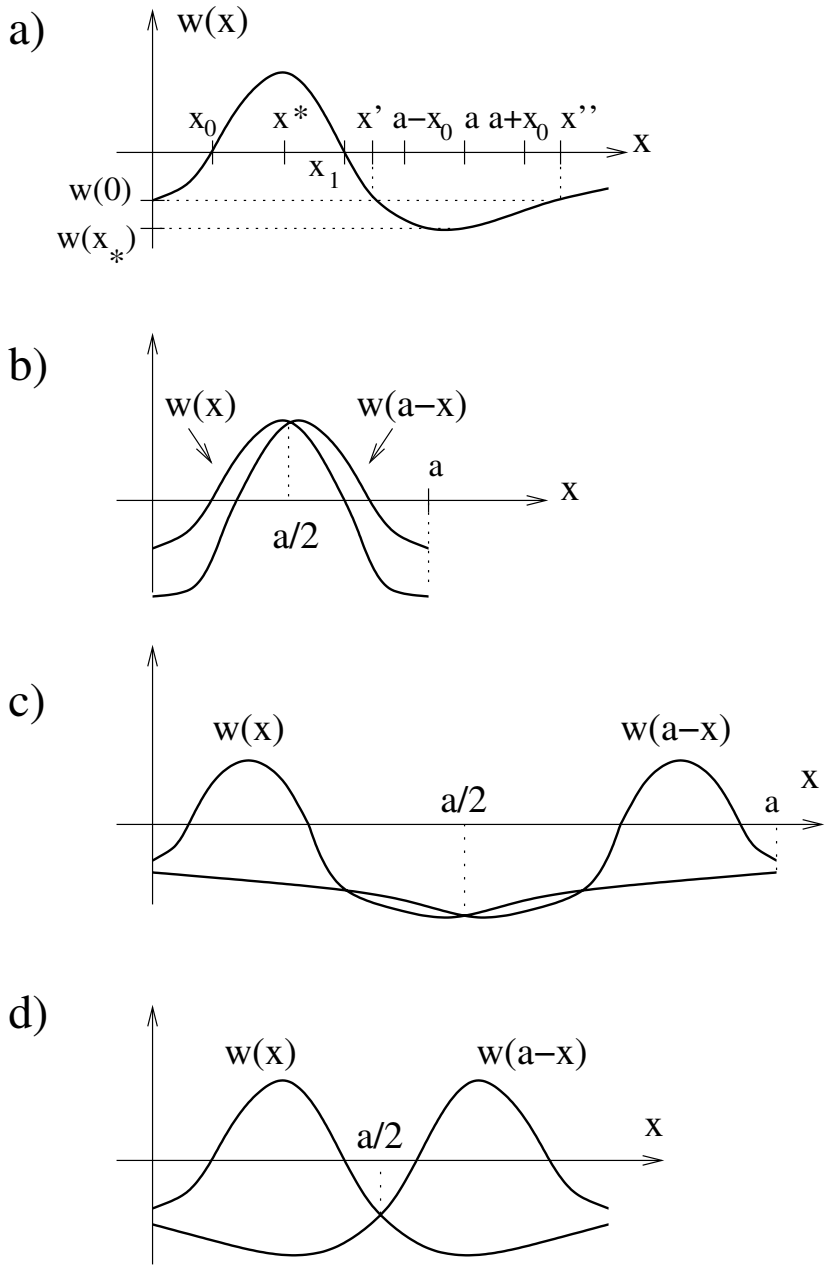


FIG. 3. *Graphical representation of Theorems* 2.5, 2.7, 2.9, *and* 2.13. *Here, a denotes* a_2 *from the theorems.* (a) *Illustration of the hypotheses of the theorems.* (b) *The relation of* $w(x), w(a_2 - x)$ *resulting from the proof of Theorem* 2.9, *if* $a_2/2 \leq x_1$. (c) *The relation of* $w(x), w(a_2 - x)$ *shown in Theorem* 2.13 *if* $a_2/2 \in (x_1, x_*]$. (d) *The relation of* $w(x), w(a_2 - x)$ *shown in Theorem* 2.13 *if* $a_2/2 > x_*$.

$u_2'(a_2) < 0$. For $x \in (a_2, a_2+x_0]$, we have $w(x) < w(0)$ by $(E3)$, while $w(x-a_2) > w(0)$ by $(H3)$. Thus, $u_2'(x)$ remains negative, as desired. ∎

Theorem 2.5 establishes the existence of a bump $u_2(x)$ when $a_2/2 > x_1$. Next, we consider the case of $a_2/2 \leq x_1$, corresponding to more negative choices of h. In this situation, we will prove two different theorems that invoke $(H6')$ and $(H6)$, respectively. Note that it is natural to interpret $(H6')$ as an assumption on the value of h, for fixed w. On the other hand, $(H6)$ is an assumption about the shape of w, independent of h. When we assume $(H6)$ below, we restrict the class of w considered, and we correspondingly restrict the shape of the bump produced (see Theorem 2.9).

Remark 2.6. Note that the proof that $u_2(x) < 0$ for all $x \in (a_2, \infty)$ does not require $a_2/2 > x_1$. Thus, in the proofs of Theorems 2.7 and 2.9 below, we will not repeat the corresponding arguments.

THEOREM 2.7. *Assume that w and h are chosen such that $(H1)$–$(H5)$, $(E1)$–$(E3)$, and $(H6')$ hold and such that $a_2/2 \leq x_1$. Then the function $u_2(x)$ defined by (2.1) with $a = a_2$, such that $-W(a_2) = h$, is a bump solution to (1.2), with $u_2(x) > 0$ if and only if $x \in (0, a_2)$.*

Proof. By symmetry (Proposition 2.1), it suffices to show that $u_2(x) > 0$ on $(0, a_2/2]$. From the proof of Theorem 2.5, we already have $u_2'(x) > 0$ on $(0, x_0]$. We will first consider $x \in (a_1, a_2/2]$ and then $x \in (x_0, a_1]$. Note that by $(H6')$, $W(a_2 - a_1) > W(a_2)$, so $a_2 - a_1 > a_1$, and thus the interval $(a_1, a_2/2]$ is nonempty. For $x \in (a_1, a_2/2]$, we have $a_2 - x \in (a_1, a_2)$, since

$$a_1 < x < a_2/2 \leq a_2 - x < a_2\,.$$

Thus, $W(a_2 - x) > -h > 0$, with a similar inequality for $W(x)$. Equation (2.2), together with the fact that W is odd, therefore gives

$$u(x) = W(x) + W(a_2 - x) + h > W(a_2 - x) > -h > 0$$

for $x \in (a_1, a_2/2]$.

Next, suppose $x \in (x_0, a_1]$. Note that over the range of positive x-values, the minimum value of W occurs at $x = x_0$, so

$$u(x) > W(x_0) + W(a_2 - x) + h =: F(x) \quad \text{for } x \in (x_0, a_1]\,. \tag{2.6}$$

$F(x_0) = u(x_0)$ by comparison of (2.2) and (2.6), and $u'(x) > 0$ on $(0, x_0]$ gives $u(x_0) > 0$, so $F(x_0) > 0$. Next, note that $F'(x) = -w(a_2 - x)$, so since $w(a_2 - x_0) < w(0) < 0$, it follows that $F'(x_0) > 0$.

Suppose that for some $x_f \in (x_0, a_1]$, $F'(x_f) = 0$. Then $w(a_2 - x_f) = 0$, so either $a_2 - x_f = x_0$ or $a_2 - x_f = x_1$ (see Figure 2). In the former case, we would have $x_f = a_2 - x_0 > x_1 > a_1$, however, so this cannot occur, and $a_2 - x_f = x_1$. Thus, $F''(x_f) = w'(a_2 - x_f) = w'(x_1) < 0$. Since $F(a_1) > 0$ by $(H6')$, the maximum principle implies that $F > 0$ on the entire interval $[x_0, a_1]$.

In summary, $u(x) > F(x) > 0$ on $(x_0, a_1]$. Thus, the proof is complete. ∎

Remark 2.8. Note that hypothesis $(H6')$ requires $a_2 - a_1 > A$. That is, if $a_2 - a_1 < A$, then $W(x_0)$, $-W(a_2)$, and $W(a_2 - a_1)$ are all negative terms, and $(H6')$ cannot hold. Thus $(H6')$ ensures that we are in a regime in which Proposition 2.4 does not apply.

To conclude this subsection, we show that under the symmetry hypothesis $(H6)$, there exists a bump $u_2(x)$ that is monotone increasing on $[0, a_2/2)$ and monotone decreasing on $(a_2/2, a_2]$, for $a_2/2 \leq x_1$.

THEOREM 2.9. *Assume that* $(H1)$–$(H6)$ *hold. Fix* h *such that* $(E1)$–$(E3)$ *hold and such that* $a_2/2 \le x_1$. *Then the function* $u_2(x)$ *defined by* (2.1) *with* $a = a_2$, *such that* $-W(a_2) = h$, *is a bump solution to* (1.2), *with* $u_2(x) > 0$ *if and only if* $x \in (0, a_2)$. *Moreover,* $u_2'(x) > 0$ *on* $[0, a_2/2)$ *and* $u_2'(x) < 0$ *on* $(a_2/2, a_2]$.

Proof. Again, from the proof of Theorem 2.5, we have $u_2'(x) > 0$ on $(0, x_0]$. Now, suppose that $a_2/2 \le x_1$. By the assumption of the theorem, $a_2 - x_0 > x' > x_1$. Therefore, $a_2/2 > (x_0 + x_1)/2 = x^*$. Together with $(H6)$, this implies that $w(a_2 - x) < w(x)$ remains true on $(0, x^*]$, that $w(a_2 - x) = w(x)$ precisely at $x = a_2/2 \in (x^*, a_2 - x^*)$, and $w(a_2 - x) > w(x)$ on $(a_2/2, a_2]$ (see Figures 2 and 3(a) and (b)). Thus, $u_2(x) > 0$ for all $x \in (0, a_2)$, with $u_2(a_2) = 0$, if $a_2/2 \le x_1$. ∎

COROLLARY 2.10. *Let* $w = w(x, \mu)$ *be continuous in* $\mu \in \mathbb{R}$ *and set* $W(x, \mu) = \int_0^x w(t, \mu)\, dt$. *Assume that* w *satisfies* $(H1)$–$(H5)$ *for all* μ *in a neighborhood of* $\mu = 0$ *and that* $w(x, 0)$ *satisfies* $(H6)$. *Then there exist* $\mu_1 < 0 < \mu_2$ *and a function* $a(\mu)$, *with* $a(0) = a_2$ *given by* $-W(a_2, 0) = h$, *such that a bump solution* $u_\mu(x)$ *of* (1.2) *exists, with* $u_\mu(x) > 0$ *if and only if* $x \in (0, a(\mu))$, *for all* $\mu \in (\mu_1, \mu_2)$.

Proof. By the implicit function theorem, since $W_x(a_2, 0) \neq 0$, a unique function $a(\mu)$ satisfying $W(a(\mu), \mu) + h = 0$ exists near $\mu = 0$, with $a(0) = a_2$. The existence of the bump solution $u_\mu(x)$ then follows immediately from the proof of Theorem 2.5, for $|\mu|$ sufficiently small. ∎

Remark 2.11. Without hypothesis $(H6)$, or some other restriction on the behavior of $w(x)$, there could be an unlimited variety of zeros of $u_2'(x)$ on $(0, a_2)$, depending on the relative rates of change of w to the left and right of x^*. In fact, without some hypothesis such as $(H6)$ or $(H6')$, $u_2(x)$ could become negative inside $(0, a_2)$, and the bump could fail to exist, as seen in Proposition 2.4. This issue is explored further in subsection 2.4.

Remark 2.12. The condition $(H6)$ as stated may seem to represent somewhat restrictive conditions to be achieved as an architecture of synaptic connections in a biological neuronal network. However, suppose we consider bumps as a form of memory. It is possible that for a given pattern of past experiences, only certain subnetworks within a coupled E-I network should be able to form bumps, corresponding to the particular memories stored in the network. The learning process could consist of the scaling of synaptic connections and their associated weights to develop particular architectural patterns. From this viewpoint, restrictions on synaptic architectures required for the appearance of bumps might be an essential feature of E-I networks, to prevent spurious overactivity. See the discussion in section 4 for consideration of related ideas.

2.3. The shape of the bump without $(H6)$ or $(H6')$. We have already seen that hypothesis $(H6)$ gives certain monotonicity properties for the bump $u_2(x)$ when $a_2/2 \le x_1$. We next characterize more generally how the shape of $u_2(x)$ depends on the position of a_2, without assuming $(H6)$ or $(H6')$, when $a_2/2 > x_1$.

THEOREM 2.13. *If* $a_2/2 \in (x_1, x_*]$, *then* u_2 *has a unique global maximum at* $x = a_2/2$. *If* $a_2/2 > x_*$, *then* $u_2'(x)$ *has at least three zeroes on* $(0, a_2)$, *including a local minimum of* $u_2(x)$ *at* $x = a_2/2$.

Proof. Suppose that $a_2/2 > x_1$. We look for zeroes of $u_2'(x)$, as given in equation (2.4). From the proof of Theorem 2.5, we already know that $u_2'(x) > 0$ on $(0, x_0]$. The condition $a_2/2 > x_1$ implies that

$$w(x_1) = 0 > w(a_2 - x_1), \tag{2.7}$$

so in fact $w(x) \geq 0 > w(a_2 - x)$ for all $x \in (x_0, x_1]$ and $u_2'(x) > 0$ on $(x_0, x_1]$. Suppose that $a_2/2 \leq x_*$, the point where $w(x)$ has its minimum (see Figure 3(a)), and define $\bar{x}$ by $a_2 - \bar{x} = x_*$. Then $\bar{x} = a_2 - x_* \leq x_*$. Thus, $w(a_2 - x) = w(x)$ at exactly one value $x \in (\bar{x}, x_*]$, namely, at $x = a_2/2$, and this is the only zero of $u_2'(x)$ in $(0, a_2)$. As in the previous case, this is a global maximum for $u_2(x)$ (see Figure 3(d)).

We now show that if $a_2/2 > x_*$, then in fact $u_2'(x) = w(a_2 - x) - w(x)$ has at least *three* zeroes (see Figure 3(c)). To see this, first note that inequality (2.7) still holds, since we still have $a_2/2 > x_1$. However, $w(x_*) < w(a_2 - x_*)$, since $a_2 - x_* > x_*$ and $w(x)$ has its minimum at $x = x_*$. Thus, $u_2'(x)$ has at least one zero on (x_1, x_*). For $x > x_*$, $w(x)$ is increasing by $(H3)$, while $w(a_2 - x)$ decreases until $x = a_2 - x_*$, at which point $w(x) > w(a_2 - x)$. Thus, exactly one additional zero of $u_2'(x)$ occurs on $(x_*, a_2 - x_*)$. Since $a_2/2 > x_*$, we have $a_2 - x_* > a_2/2$. As noted earlier, $u_2'(x)$ has a zero at $x = a_2/2$ from the form of (2.4); hence, this second zero must occur at $x = a_2/2$, and $u_2'(x)$ has at least one more zero for $x > a_2 - x_*$, by symmetry.

From (2.4), it follows that $u_2''(x) = w'(x) + w'(a_2 - x)$, such that $u_2''(a_2/2) = 2w'(a_2/2)$. When $a_2/2 > x_*$, we have $w'(a_2/2) > 0$ by $(H3)$, such that $u_2''(a_2/2) > 0$; that is, u_2 has a local minimum at $a_2/2$, completing the proof. ∎

Remark 2.14. Examples of bump solutions $u_2(x)$ of (2.1) with $a_2/2 < x_*$ and $a_2/2 > x_*$, respectively, are shown in Figure 4. These plots were generated by solving (2.1) numerically with a coupling function $w(x)$ defined in a piecewise manner on $[0, \infty)$ as

$$w(x) = \begin{cases} -Kx(x-1) - \epsilon, & x \in [0, 1), \\ -(x - 1 + \epsilon)e^{-b(x-1)}, & x \in [1, \infty) \end{cases} \tag{2.8}$$

for $K, b > 0$ and $0 < \epsilon < \min\{K/4, 1/b\}$, and then extended to be even on $(-\infty, \infty)$. This function satisfies $(H1)$. It also satisfies $(H2)$ and $(H4)$, with $x_0 = \frac{1}{2}[1 - \sqrt{1 - 4\epsilon/K}]$ and $x_1 = \frac{1}{2}[1 + \sqrt{1 - 4\epsilon/K}] < 1$. Assumptions $(H3)$ and $(H5)$ hold for this $w(x)$ as well, with $x^* = 1/2$ and $x_* = 1 + 1/b - \epsilon$. Finally, $w(x)$ is symmetric about $x^* = 1/2$ on (x_0, x_1), satisfying $(H6)$.

From (2.4), no matter what the value of a_2, we have $u_2'(a_2/2) = 0$. Thus, it is of interest to estimate $u_2(a_2/2)$.

PROPOSITION 2.15. *If $a_2/2 > a_1$ and $u_2(x) > 0$ on $(0, a_2)$, then $u_2(a_2/2) > -h$.*

Proof. If $a_2/2 > a_1$, then $a_2/2 \in (a_1, a_2)$, so $W(a_2/2) > -h$ (see Figure 2). From (2.2), this implies

$$u_2(a_2/2) = 2W(a_2/2) + h > W(a_2/2) > -h\,. \qquad \blacksquare$$

Combining Theorem 2.13 and Proposition 2.15 yields the following result.

COROLLARY 2.16. *If a bump $u_2(x)$ exists with $a_2/2 \in (x_1, x_*]$, then the activity level $u_2(x)$ attains a maximum value greater than $-h$. If a bump $u_2(x)$ exists with $a_2/2 > x_*$, then at the local activity minimum $a_2/2$, the activity level is bounded below by $-h$.*

2.4. Birth and death of bumps. We have seen that the function $u_2(x)$ is a bump for some values of a_2, but may fail to be a bump in some cases, such as $a_2 - a_1 < A$, as in Proposition 2.2. In fact, since $\lim_{x\to\infty} w(x) = 0$, we will see below that the interval of a values on which $u_2(x)$ is a valid bump is finite. Thus, a family of bump solutions, parametrized by bump length a, must be born at some finite value of a and must die at some larger finite value of a. In this subsection,

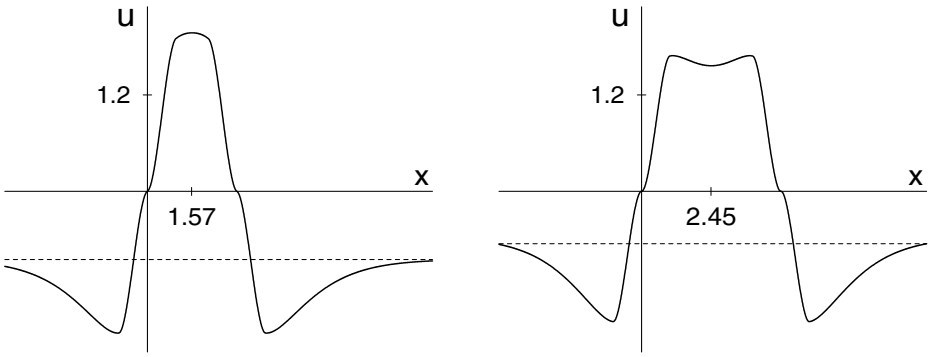


FIG. 4. *Bump solutions of* (2.1) *for the coupling function* $w(x)$ *given in* (2.8) *and extended in an even manner. Parameters used are* $K = 10.0, \epsilon = 0.1, b = 1.0$*, such that* $x_* = 1.9$*. Left: A bump with no local minimum, found with* $h = -0.85$*, such that* $a_2 = 3.15 < 2x_*$*. Right: A bump with a local minimum at* $a_2/2 > x_*$*, found with* $h = -0.57$*, such that* $a_2 = 4.90 > 2x_*$*. Dashed lines show the levels of* h*.*

we discuss possible mechanisms by which bumps may be created or destroyed as a varies (which can be achieved by varying h). We shall see that this does not occur through a "usual" bifurcation mechanism, such as a saddle-node bifurcation, and that when bumps arise, they do so with a finite amplitude. This contrasts with the situation when the coupling function $w(x)$ is derived from recurrent excitation and lateral inhibition, in which case bump amplitudes and widths may go to zero as a parameter varies.

In the following analysis, we will always assume $a > x_1$, corresponding to the possible bump solution $u_2(x)$, since $u_1(x)$ is never a valid bump. We will also assume that $W(a) > 0$, also necessary for a to represent a bump length since $h < 0$ in (2.3).

In general, there are two types of transitions through which a bump $u_2(x)$ may cease to exist as its size a varies, even without interaction with any other solutions. One possibility is that a bump may go negative on its interior; that is, it may develop a dip as in Figure 4, which may continue to drop until the minimum value of $u_2(x)$ on $(0, a)$ passes through 0. We refer to this as the *internal tangency* mechanism, since right at the transitional a value, we have $u_2(x) = u_2'(x) = 0$ for some $x \in (0, a)$. By (2.5),

$$u_2(a/2) = 2W(a/2) - W(a). \tag{2.9}$$

We will use (2.9) to show that $u_2(a/2) = 0$ can occur only at a unique value of $a > x_1$.

PROPOSITION 2.17. *There is at most one value of* $a > x_1$ *for which* $W(a) > 0$ *and* $u_2(a/2) = 0$*.*

Proof. Suppose $u_2(a_0/2) = 0$ and $W(a_0) > 0$ for $a_0 > x_1$. Then $W(a_0) > 0$ implies that $W(a_0/2) > 0$, by (2.9). Thus, $a_0/2 > A$, where A was defined as the smallest positive value for which $W(x) = 0$. Moreover, $W(a_0) = 2W(a_0/2)$ implies that $W(a_0) > W(a_0/2)$, so $a_0/2 < x_1$ (e.g., Figure 2).

Now, let $z(a) = 2W(a/2) - W(a)$. For $a/2 \in (A, x_1)$, $W'(a/2) > 0$, while $a > x_1$ gives $W'(a) < 0$. Thus, $z'(a) > 0$ and z can have at most one zero with $a/2 \in (A, x_1)$. Since $a_0/2$ must lie in this interval whenever $u_2(a_0/2) = 0$, this concludes the proof. ∎

Proposition 2.17 rules out the possibility that a family of bump solutions both arises and dies by passage of $u_2(a/2)$ through 0. However, it does not rule out the possibility that the birth and the death of the family are associated with dips on the interior of $(0, a)$ switching from negative to positive and from positive to negative, respectively. Indeed, it is possible that $u_2(a/2 - \delta) = 0$ for some $\delta \in (0, a/2)$, with $u_2(a/2+\delta) = 0$ as well by Proposition 2.1. Thus, there are infinitely many positions at which interior zeros of u_2 could develop, always occurring in groups of even numbers of dips, placed symmetrically about $a/2$.

An alternative to the internal tangency mechanism for birth and death of bumps is that $u_2'(0)$ (and by symmetry, $u_2'(a)$) may become zero and then negative (positive) as a varies. We refer to this transition as the *boundary tangency* mechanism. Note that $u_2'(0) = w(0) - w(a)$, where a is the length of $u_2(x)$. Thus, $u_2'(0) < 0$ for $a \in (x_1, x')$, and, since $\lim_{x\to\infty} w(x) = 0$, $u_2'(0) < 0$ for all $a > x''$ (see Figure 3). Hence, the boundary tangency mechanism ensures that the family of bumps can exist only on a finite interval of positive a values.

Note further that $u_2''(0) = w'(0) + w'(a) = w'(a)$. If $u_2'(0) = 0$ for some a at which a transition between bump existence and bump nonexistence occurs, then we must have $u_2''(0) > 0$ (see Figure 8 below). This implies that the boundary tangency mechanism can apply only for a values such that $w'(a) > 0$, corresponding to bump death for large a as $-h$ is lowered toward 0 (see Figures 2 and 3). However, as we have noted, $u_2'(0) < 0$ for a sufficiently close to (but above) x_1. Thus, bumps must be born through an internal transition from negativity to positivity (the internal tangency mechanism discussed above) as a increases sufficiently beyond x_1. We summarize the above discussion in the following proposition, stated in terms of loss of existing bumps as a decreases or increases.

PROPOSITION 2.18. *Suppose there exists a_2 such that a bump solution $u_2(x)$ exists, with $u_2(x) > 0$ precisely for $x \in (0, a_2)$. Then as a decreases from a_2, the bump is lost through the internal tangency mechanism. As a increases from a_2, the bump is lost through either the internal tangency mechanism or the boundary tangency mechanism. In the former case, at least one of the internal tangencies must be at $x \neq a/2$.*

Remark 2.19. These birth and death mechanisms suggest that there may exist multibump solutions that satisfy (1.2). The existence and stability of such solutions remain open for future investigation. Note that the existence of multibumps cannot be addressed using (2.1) directly, since multibumps are positive on multiple disjoint regions.

We conclude this subsection by considering a numerical example of bump birth and death. The example coupling function that we introduce here will also be considered in section 3. Define

$$w(x) = (x^2 - c)w_0(x) := (x^2 - c)(De^{-dx^2} - Be^{-bx^2}). \tag{2.10}$$

We will take $c = 0.5, D = 11, d = 0.05, B = 6$, and $b = 0.035$ as our parameter values unless otherwise stated. This satisfies $(H1)$–$(H5)$; the functions $w_0(x)$ and $w(x)$ for these parameters appear in Figure 5. We also show the corresponding function $W(x) = \int_0^x w(t)\,dt$ on a range of positive x values. Note that in this example, $\lim_{x\to\infty} W(x) > 0$.

For this example, we gradually increase a. We plot $u(a/2)$ versus a in Figure 6, where u satisfies equation (2.2) with h such that equation (2.3) holds. A bump solution is formed when $u(a/2)$ reaches 0, at about $a = a_b \approx 7.14$.

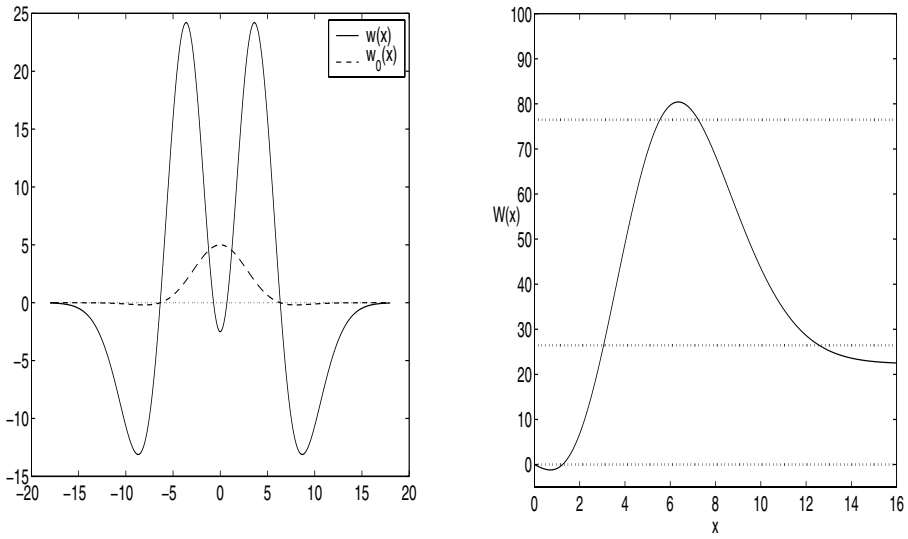


FIG. 5. *Coupling functions from* (2.10). *The left plot shows* $w(x), w_0(x)$ *with parameter values as given in the text following* (2.10). *The right plot shows* $W(x)$ *for these parameters. The dashed lines correspond to* $W(x) = 0$ *and to two special values of* $-h$ *for which bumps exist with spatially inhomogeneous coupling, discussed in section* 3.

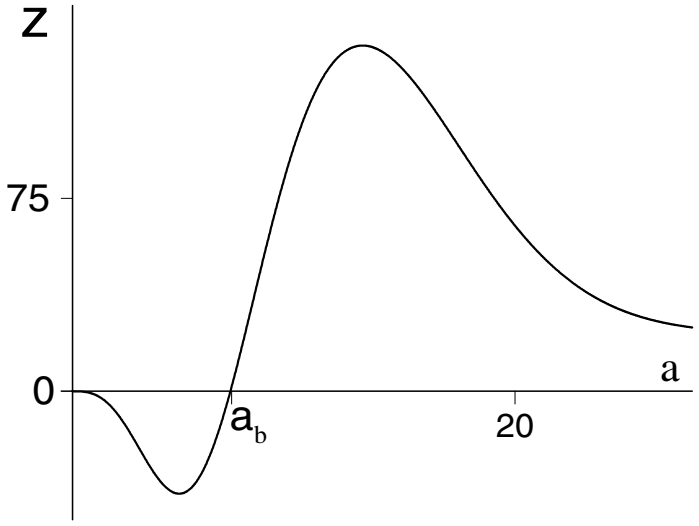


FIG. 6. *The value of* u *at* $a/2$ *plotted versus* a *for* $w(x)$ *from* (2.10). *Note that* $u(a/2) > 0$ *is necessary but not sufficient for bump existence. In this example, a family of bumps is born, as* h *(and thus* a*) is varied, as soon as* a *increases through* a_b*, such that* $u(a/2)$ *becomes positive. The bumps must die by a different mechanism, since* $u(a/2) > 0$ *for all* $a > a_b$.

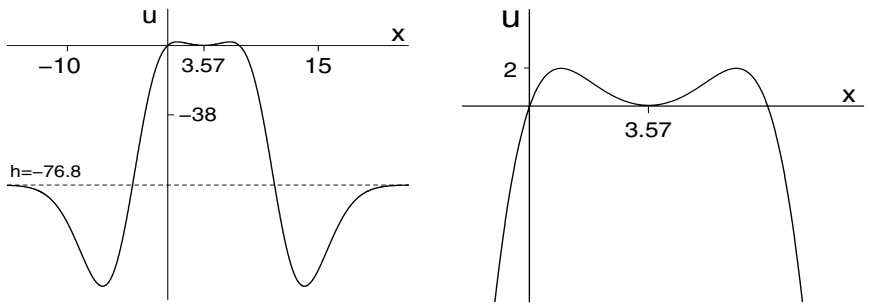


FIG. 7. *Solution $u_2(x)$ at birthpoint $a_b \approx 7.14$ (left), and blowup (right) showing internal tangency of the solution with the x-axis.*

In Figure 7, we plot the solution $u_2(x)$ with $a = a_b$. As a increases from a_b, the bumps that are born persist for an interval of a values; those found for two values of a appear in the top panels of Figure 8. Figure 8 also displays the death of the family of bumps as a continues to increase. At $a = a_d \approx 12.89$, $w(a) = w(0)$ such that $u_2'(0) = u_2'(a) = 0$. For $a > a_d$, bumps cannot exist; careful inspection shows that a representative solution to equations (2.2), (2.3), shown in the lower right plot, goes negative for small $x > 0$ and for x close to, but less than, a. This is not a valid bump solution.

2.5. Linear stability of the bump and bistability. To analyze the linear stability of the bump solution $u_2(x)$, we linearize (1.2) about $u_2(x)$. To compute the correct form of linearized equation, substitute $u = u_2(x) + v(x,t)$ into (1.2). This yields

$$\frac{\partial u}{\partial t} = \frac{\partial v}{\partial t} = -u_2(x) - v(x,t) + \int_{-\infty}^{\infty} w(x-y)H(u_2(y) - v(y,t))\,dy + h.$$

Derivation of the linear equation satisfied to first order by v requires expansion of the Heaviside function H about u_2. The result of this expansion yields [29, 19]

$$\begin{aligned} \frac{\partial v}{\partial t} &= -v + \frac{w(x)[v(0,t) - u_2(0)]}{|u_2'(0)|} + \frac{w(x-a_2)[v(a_2,t) - u_2(a_2)]}{|u_2'(a_2)|} \\ &= -v + \frac{w(x)v(0,t)}{u_2'(0)} - \frac{w(x-a_2)v(a_2,t)}{u_2'(a_2)}, \end{aligned} \tag{2.11}$$

since $u_2(0) = u_2(a_2) = 0$ by construction. Note that if v has its zeros in the same place as those of u, that is, $v(0) = v(a_2) = 0$, then $v' = -v$ and linear stability is immediate.

More generally, for linear stability, we consider perturbations of the form $v(x,t) = e^{\lambda t}v(x)$. Substitution of this expression into (2.11) and cancellation of $e^{\lambda t}$ terms yield the algebraic eigenvalue equation

$$(\lambda + 1)v(x) = w(x)v(0)/u_2'(0) - w(x-a_2)v(a_2)/u_2'(a_2). \tag{2.12}$$

Recall that $u_2'(x) = w(x) - w(x-a_2)$. Thus, (2.12) is equivalent to

$$(\lambda + 1)v(x) = \frac{w(x)v(0) + w(x-a_2)v(a_2)}{w(0) - w(a_2)}. \tag{2.13}$$

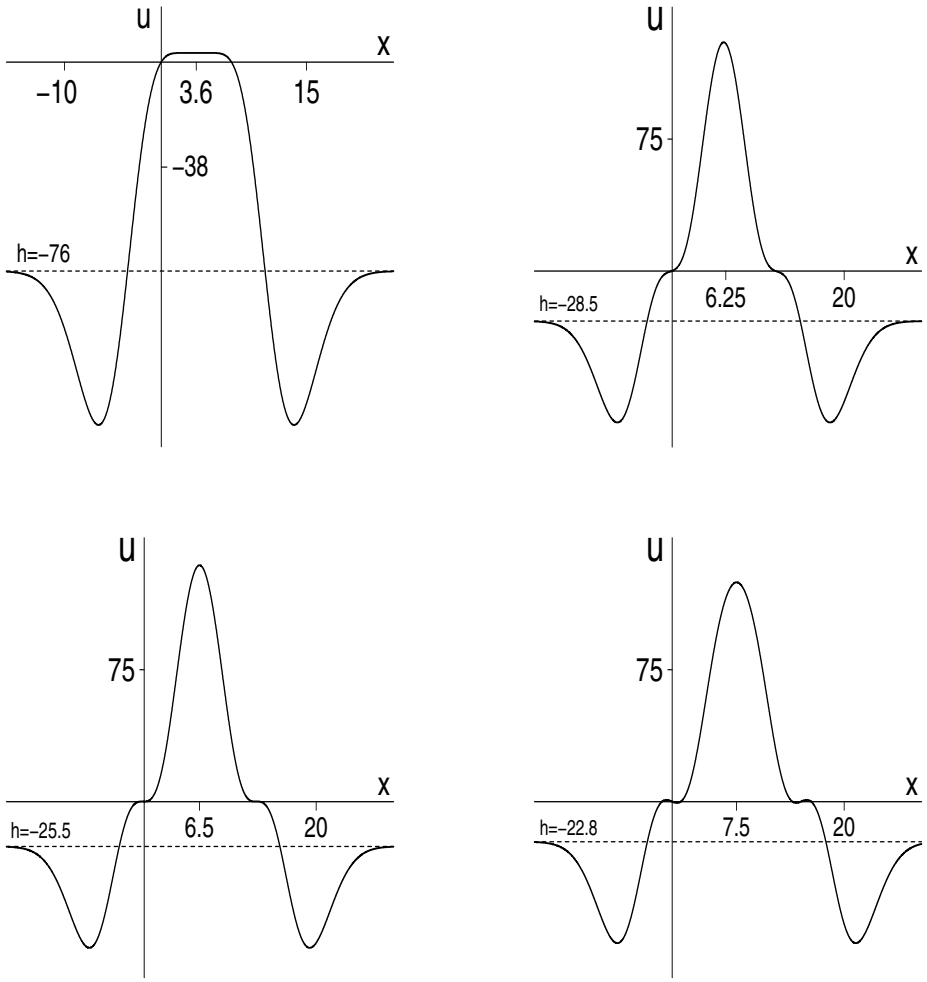


FIG. 8. *Solutions to equations* (2.2), (2.3) *at* $\epsilon = 0$: $a = 7.24$ *(upper left) just after bump birth,* $a = 12$ *(upper right) just before bump death,* $a = 12.89$ *(lower left) where bump death occurs, and* $a = 13$ *(lower right) just after bump death. Note that the solution shown in the lower right is not a valid solution to* (2.1).

Eigenvalues occur at those λ values for which (2.13) has a nontrivial solution $v(x)$. (Note that any such solution decays to 0 asymptotically, since $w(x), w(x - a_2)$ do.) Substitution of $x = 0$ and $x = a_2$ into (2.13) yields a pair of equations in the unknowns $v(0)$ and $v(a_2)$, namely,

$$(\lambda + 1)v(0) = \frac{w(0)v(0) + w(a_2)v(a_2)}{w(0) - w(a_2)}, \tag{2.14}$$

$$(\lambda + 1)v(a_2) = \frac{w(a_2)v(0) + w(0)v(a_2)}{w(0) - w(a_2)}.$$

If $v(a_2) = 0$, then $v(0) = 0$, and only the trivial solution $v \equiv 0$ satisfies (2.13). We

have already observed that perturbations with $v(0) = v(a_2) = 0$ cannot cause an instability, based on (2.11). Thus, assume that $v(a_2) \neq 0$. We use the first equation in (2.14) to write $v(0)$ as a function of $v(a_2)$. Upon substitution of this expression into the second equation in (2.14), cancellation of the nonzero quantity $v(a_2)$ multiplying each term, and algebraic manipulation, we obtain the following quadratic equation in λ:

$$\lambda^2 \left(w(0) - w(a_2)\right)^2 + \lambda \left(\left(w(0) - w(a_2)\right)^2 + w^2(a_2) - w^2(0)\right) = 0. \tag{2.15}$$

The solution $\lambda = 0$ of (2.15) corresponds to translation invariance of the bump. The other solution of (2.15) satisfies

$$\lambda = \frac{2w(a_2)\left(w(0) - w(a_2)\right)}{\left(w(0) - w(a_2)\right)^2}. \tag{2.16}$$

Recall that $w(a_2) < w(0) < 0$. Thus, the unique solution λ of equation (2.16) is real and negative, and the bump solution is linearly stable.

Note from (1.2) that $u = c := h + \int_{-\infty}^{\infty} w(x)\,dx$ is a stationary, spatially uniform solution, provided that $c > 0$. When this solution exists, the same linearization calculation that yields (2.11) yields the linearized stability equation $dv/dt = -v$, since for small perturbations $c - v > 0$, such that $H(c - v) = 1$. Thus, the spatially uniform state is linearly stable, when it exists, which implies that (1.2) features bistability, at least in terms of linear analysis. This is consistent with the findings of [20], in which a network of bursting thalamic cells with an effectively off-center form of coupling displayed bistability between a spatially localized and a spatially uniform state. In [20], however, the spatially uniform state corresponded to a complete absence of activity.

3. Spatially inhomogeneous coupling. It has been argued that the coupling between cells should be spatially inhomogeneous, reflecting local structural variations [2, 3]. In this section, we use analysis and numerics to consider how such a modification affects properties of bump solutions of (1.2). To this end, we consider bump solutions of the equation

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

To allow for concrete calculations and numerics, we mostly consider a spatial inhomogeneity used, for example, in [2], namely,

$$p(x) = 1 + \epsilon(1 + \cos(\rho x + \phi)). \tag{3.2}$$

Without loss of generality, we take $\rho = 1$.

In subsection 3.1, we will consider the special case of $\phi = 0$, restricting our attention to bumps on $(0, a)$. In subsection 3.2, we will address the general bump existence question for $p(x)$ given by (3.2). We shall see that, in contrast to the spatially homogeneous case, the presence of inhomogeneity implies that for fixed $p(x)$, for each bump starting point, there is only a small, discrete set of background input levels for which bumps can occur, each with a unique corresponding size. Based on the mechanisms that we observe with $p(x)$ given by (3.2), we expect qualitatively similar results for nonperiodic $p(x) = 1 + \epsilon p_0(x)$ (see Remark 3.4 at the end of the section). Further, at least for the case of $p(x)$ given by (3.2) and $w(x)$ given by (2.10), we find that among the possible bump sizes, there is a certain invariant size selected independent of ϕ and of bump starting position. Possible functional implications of these results are considered in the discussion in section 4.

3.1. Bumps on $(0, a)$ with no phase shift ($\phi = 0$). Note that spatial inhomogeneity in coupling may destroy spatial translation invariance of bump solutions. For clarity, we first consider the special case of bumps on $(0, a)$ with phase shift $\phi = 0$. We will illustrate the key observation that for a fixed spatial pattern of coupling (fixed ρ and ϕ) and a fixed starting point of an activity bump (here $x = 0$), there is only a small, discrete set of possible bump sizes that can be selected. That is, the spatial inhomogeneity induces a form of bump pinning.

As previously, a bump must satisfy $u(0) = u(a) = 0$ and $u(x) > 0$ if and only if $x \in (0, a)$, for some positive number a. If a bump solution $u(x)$ exists for some a, the Heaviside function H in equation (3.1) implies that $u(x)$ must satisfy

$$u(x) = \int_0^a w(x - \eta)p(\eta)\, d\eta + h \tag{3.3}$$

for that a. Thus, to find a bump solution, we first seek a for which $u(0) = u(a) = 0$, with $u(x)$ specified by (3.3).

The corresponding equations are

$$0 = \int_0^a w(\eta)p(\eta)\, d\eta + h \tag{3.4}$$

and

$$0 = \int_0^a w(a - \eta)p(\eta)\, d\eta + h. \tag{3.5}$$

Subtracting these two equations, i.e., (3.5)−(3.4), yields

$$g(a) := \int_0^a w(a - \eta)p(\eta)\, d\eta - \int_0^a w(\eta)p(\eta)\, d\eta = 0. \tag{3.6}$$

To find candidate values of a, we first seek solutions of $g(a) = 0$, given by

$$\int_0^a w(\eta)p(\eta)\, d\eta = \int_0^a w(a - \eta)p(\eta)\, d\eta. \tag{3.7}$$

Now, from the substitution $y = a - \eta$, note that

$$\int_0^a w(a - \eta)p(\eta)\, d\eta = \int_0^a w(y)p(a - y)\, dy.$$

But $p(a - y) = 1 + \epsilon(1 + \cos(a - y)) = 1 + \epsilon(1 + \cos a \cos y + \sin a \sin y)$. Hence, if $a = 2n\pi$ for any integer n, then $p(a - y) = p(y)$, and we find

$$g(2n\pi) = \int_0^a w(y)p(y)\, dy - \int_0^a w(\eta)p(\eta)\, d\eta = 0.$$

Thus, $a = 2n\pi$ solves $g(a) = 0$ for any integer n (see Figures 9 and 11). However, we also need (3.4), (3.5) to hold such that $u(0) = 0$ and $u(a) = 0$, which occurs only for those special values of n such that $-h = \int_0^{2n\pi} w(\eta)p(\eta)\, d\eta$ $(= \int_0^{2n\pi} w(2n\pi - \eta)p(\eta)\, d\eta$ since $g = 0)$, which may or may not be positive, as required.

Remark 3.1. This does not imply there is a special biological significance to bump sizes that are even integer multiples of π. If $\rho \neq 1$, then other zeros result here. The point is that the nature of the spatial variation $p(x)$ selects possible bump sizes.

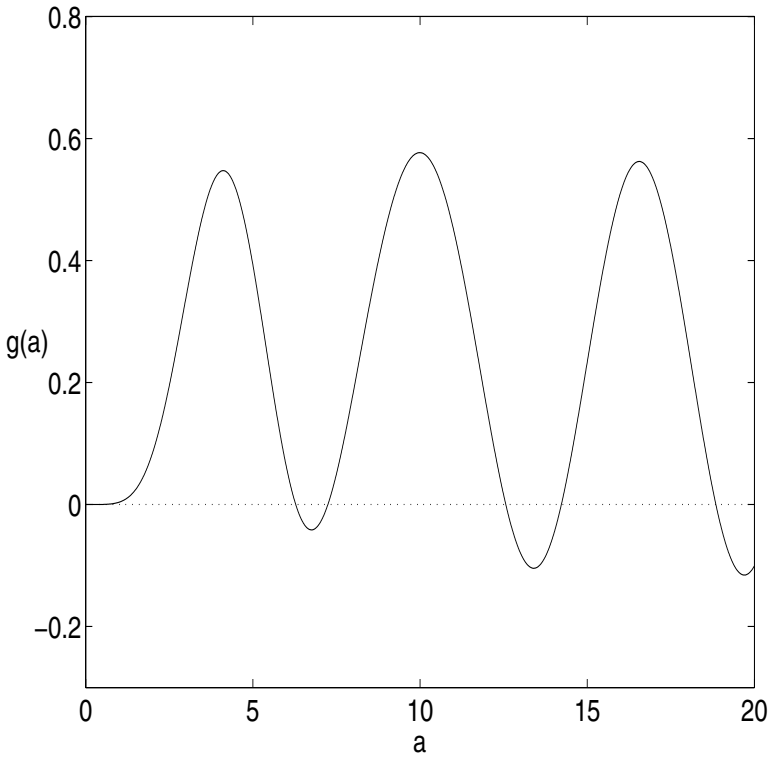


FIG. 9. *The function $g(a)$, defined in* (3.6)*, for $w(x)$ from* (2.10) *with our usual parameter values. Note that $g(0) = g(2\pi) = g(4\pi) = 0$ and that there are other zeros of g that are not even integer multiples of π.*

There also may be other solutions of $g(a) = 0$. We seek these numerically. To do so, we apply Matlab directly, and we also check our results by using XPPAUT [12] to solve an ordinary differential equation for $g(a)$, derived in the second subsection of the appendix.

We consider now the coupling function $w(x)$ defined in equation (2.10) in the previous section and shown in Figure 5, namely,

$$w(x) = (x^2 - c)w_0(x) := (x^2 - c)(De^{-dx^2} - Be^{-bx^2}),$$

with $c = 0.5, D = 11, d = 0.05, B = 6$, and $b = 0.035$ as usual.

The resulting $g(a)$, for $\epsilon = 0.01$, appears in Figure 9. Numerically, the zeros of $g(a)$ in the set of positive a are $\{2\pi, 7.25, 4\pi, 14.23, 6\pi, \ldots\}$, where the zeros that are not integer multiples of π form a single sequence in which the difference between subsequent elements tends to 2π, since $w(x)$ tends to 0 as $x \to \infty$. We find qualitatively similar results, namely, a countable collection of isolated zeros with similar behavior as $a \to \infty$, for a variety of other parameter sets for $w(x)$ with $\epsilon > 0$.

We note that in general, $g'(0) = g''(0) = g'''(0) = 0$ (see subsection 5.2 of the appendix for a proof). Moreover, we find from (5.8) in the appendix that

$$g^{(4)}(a) = -g''(a) - 3w''(a)p'(a) - 2w'(a)p''(a) + w'''(a)(p(0) - p(a)),$$

so $g^{(4)}(0) = 0$, while

$$g^{(5)}(a) = -g'''(a) - 4w'''(a)p'(a) - 5w''(a)p''(a) - 2w'(a)p'''(a) + w^{(4)}(a)(p(0) - p(a)),$$

so $g^{(5)}(0) = -5w''(0)p''(0) = 5\epsilon w''(0) > 0$. This gives a sense of the behavior of g near $a = 0$, which depends on ϵ.

Since $p(x) = 1 + O(\epsilon)$, it is obvious that the zeros of g do not depend on ϵ. More explicitly, the function $g(a)$ defined in (3.6) can be rewritten as

$$g(a) = \int_0^a w(\eta)p(a-\eta)\,d\eta - \int_0^a w(\eta)p(\eta)\,d\eta.$$

Upon substitution of definition (3.2) for p with $\rho = 1$ and $\phi = 0$ and application of a trigonometric identity for $\cos(a - \eta)$, this yields

$$g(a) = \epsilon\left[(\cos a - 1)\int_0^a w(\eta)\cos(\eta)\,d\eta + \sin a \int_0^a w(\eta)\sin(\eta)\,d\eta\right], \tag{3.8}$$

which will also be useful below.

Once we have found the zeros of g for a particular choice of parameters (including ϵ), it remains to check whether these really correspond to a values for which (3.4), (3.5) hold, for some $h < 0$. Only in that case will a bump possibly exist. Note that we restrict further to those a values such that

$$\frac{d}{da}\int_0^a w(\eta)p(\eta)d\eta = w(a)p(a) < 0, \tag{3.9}$$

since only a_2, but not a_1, gives a valid bump in the $\epsilon = 0$ case. In the example shown, the zeros $a \approx 7.25$ and $a = 4\pi$ of g are the only ones which satisfy (3.4), (3.5), and (3.9) for some $h < 0$. The corresponding h values for $\epsilon = 0$ are $h \approx -76.09$ and $h \approx -26.45$, respectively, although these depend on ϵ. The intersections of these values of h with $W(x)$ for $\epsilon = 0$ are displayed in Figure 5. In Figure 10, we plot the corresponding bump solution for $a \approx 7.25$ with $\epsilon = 0.1$. Figure 11 shows the bump solutions for $a = 4\pi$ with $\epsilon = 0, 0.1$, and 0.2, respectively. Note that the bump with $a \approx 7.25$ loses its symmetry for $\epsilon > 0$, while the bump with $a = 4\pi$ is symmetric about $a/2 = 2\pi$ for all ϵ by the 2π-periodicity of $\cos(x)$ and $\sin(x)$. Further, in both cases, the bump widths are independent of ϵ.

3.2. General case: Bumps on (b_1, b_2) with arbitrary ϕ. In this section, we will arrive at the following result: Given a spatial inhomogeneity of coupling of the form (3.2), with ϕ fixed, for any bump starting point b_1, there is a small, discrete set of possible bump sizes. Moreover, there is a subset of these sizes (possibly empty, but nonempty for the main example that we have been considering) which are possible for *all* choices of b_1 and ϕ.

In the general case, the bump existence equations become

$$0 = \int_{b_1}^{b_2} w(b_2 - \eta)p(\eta)\,d\eta + h \tag{3.10}$$

and

$$0 = \int_{b_1}^{b_2} w(b_1 - \eta)p(\eta)\,d\eta + h, \tag{3.11}$$

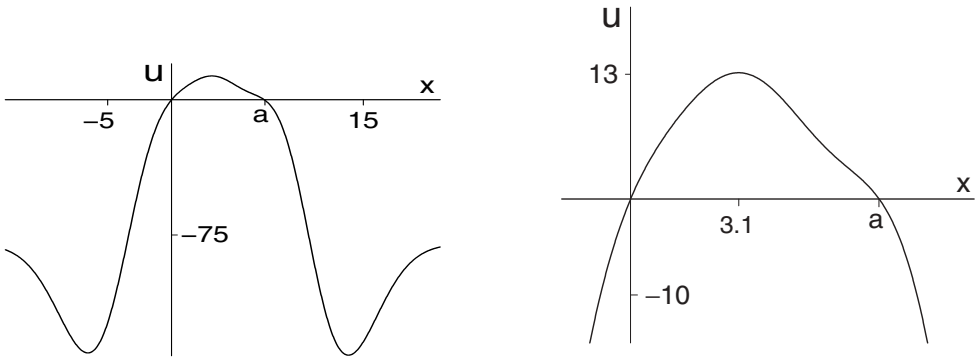


FIG. 10. *Solution at $a \approx 7.25$ and $\epsilon = .1$ (left) and blowup (right). Note that the solution is not symmetric around $x = \frac{a}{2}$ when $\epsilon > 0$ and a is not an even multiple of π.*

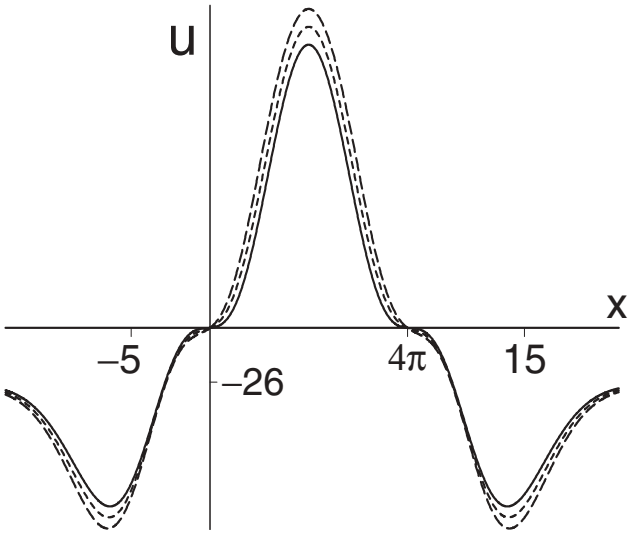


FIG. 11. *Solutions at $a = 4\pi$ and $\epsilon = 0$ (solid curve), $\epsilon = .1$ (dashed curve), and $\epsilon = .2$ (long dashed). Note that solutions are symmetric around $x = \frac{a}{2}$ when $\epsilon > 0$ and a is an even multiple of π.*

where we use (b_1, b_2) to denote the interval on which the bump is positive to avoid confusion with our earlier use of a_1, a_2. Again, we subtract to obtain

$$0 = \int_{b_1}^{b_2} w(b_2 - \eta)p(\eta)\,d\eta - \int_{b_1}^{b_2} w(b_1 - \eta)p(\eta)\,d\eta. \tag{3.12}$$

We seek solutions of (3.12), which are exactly the solutions of the following equation, attained by change of variables and by setting $z_i = b_i - \phi$ for $i = 1, 2$:

$$0 = g(z_1, z_2) := \int_0^{z_2 - z_1} w(y)p(b_2 - y)\,dy + \int_0^{z_2 - z_1} w(y)p(b_1 - y)\,dy. \tag{3.13}$$

It is not apparent by inspection that $g(z_1, z_2)$ as defined in (3.13) is a function of z_1, z_2 only. However, using the definition of p in (3.2) and trigonometric sum and difference

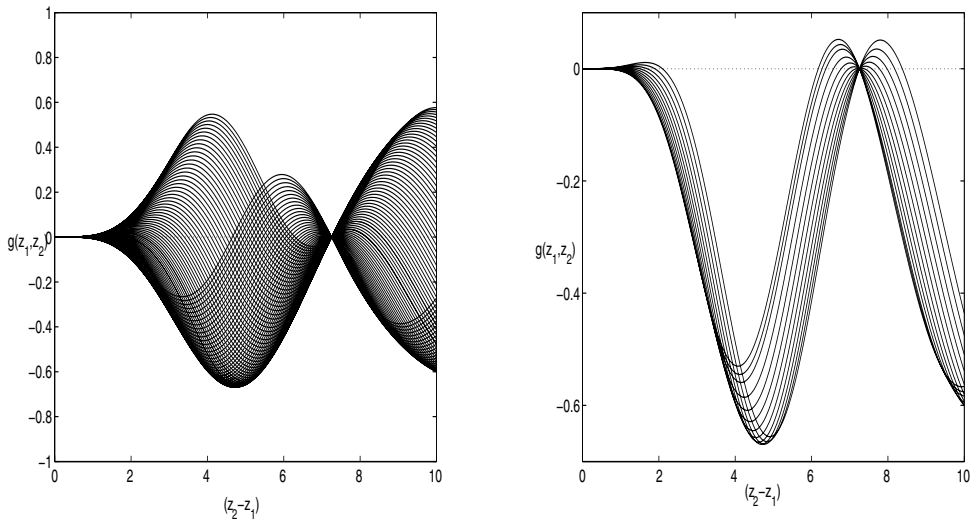


FIG. 12. *The function $g(z_1, z_2)$ given in* (3.14). *Each single curve in each plot shows $g(z_1, z_2)$ versus $z_2 - z_1$ for fixed z_1. Different curves correspond to different z_1 values. Note that $z_2 - z_1 \approx 7.25$ is a zero of $g(z_1, z_2)$ for all z_1. The right plot shows a closer view around $z_2 - z_1 \approx 7.25$, with fewer curves shown than on the left.*

identities, one can calculate that

$$g(z_1, z_2) = \epsilon \int_0^{z_2 - z_1} w(y) \left[\cos y(\cos z_2 - \cos z_1) + \sin y(\sin z_2 + \sin z_1)\right] dy. \tag{3.14}$$

Note that (3.8) corresponds to a special case of (3.14), with $z_1 = 0$. Further, as noted in subsection 3.1, the zeros of $g(z_1, z_2)$ are independent of $\epsilon > 0$.

Again, the realizable bump sizes, determined by (3.10), (3.11) with the restriction $h < 0$, are a subset of the set of the zeros of g. For fixed ϕ, if we start with $b_1 = \phi$ (that is, $z_1 = 0$), then we recover exactly the bump sizes found with $\phi = 0$. As b_1 is varied from ϕ (or, equivalently, z_1 is varied from 0), then we may pick out different bump sizes. *Some of these, however, may be invariant under changes in z_1.* Indeed, Figure 12 shows plots of $g(z_1, z_2)$ for $w(x)$ from (2.10) and $p(x)$ from (3.2). To produce this figure, z_1 was systematically varied (increasing from 0), and for each fixed z_1, z_2 was varied from z_1 up to $z_1 + 10$ to form an individual curve. The figure shows the resulting $g(z_1, z_2)$ values for each fixed z_1 plotted versus $z_2 - z_1$; that is, each curve has been translated so that it begins at $z_2 - z_1 = 0$, with $g = 0$ correspondingly. The value $z_2 - z_1 = z^* \approx 7.25$ gives a zero of g, corresponding to the existence of a bump solution with $h < 0$, for each starting position z_1. The close-up in the right panel of the figure shows how $\partial g(z_1, z_1 + z^*)/\partial z_2$ passes through 0 as z_1 is varied. Note that similar results were obtained with various other choices of parameter values in $w(x)$.

Remark 3.2. Since the bump size $z_2 - z_1 \approx 7.25$ is realized for all z_1, and since $z_i = b_i - \phi$, this size is invariant under changes in ϕ. That is, for any choice of ϕ and starting position b_1, if $b_2 \approx b_1 + 7.25$, then there is a bump solution $u(x)$ of width approximately equal to 7.25 such that $u(x) > 0$ precisely for $x \in (b_1, b_2)$. Although this solution retains its width, it will occur at different levels of h for different choices of b_1, b_2, ϕ.

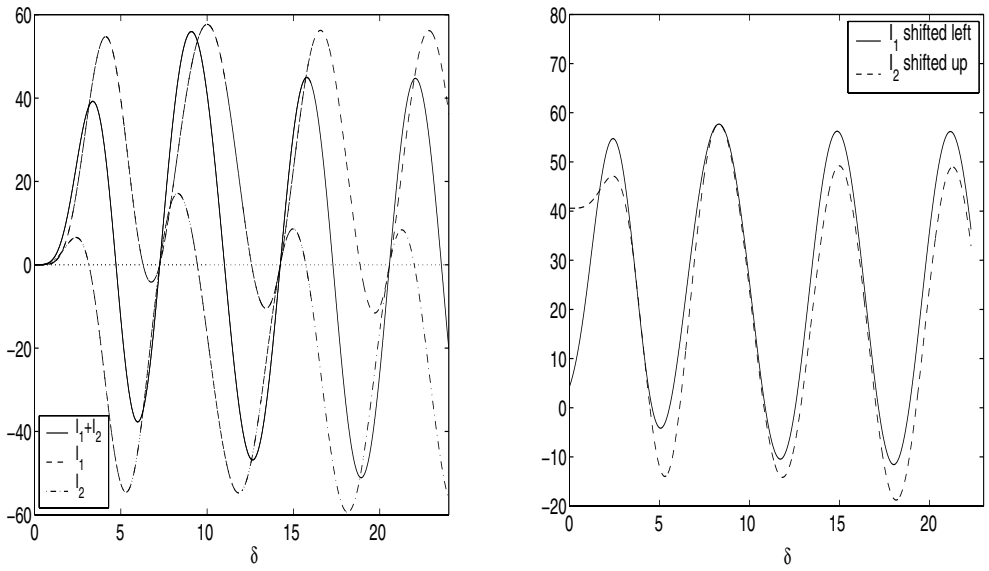


FIG. 13. *Components of* (3.16) *for* $w(x)$ *given in* (2.10) *with usual parameter values. The left plot shows* $I_1(\delta) + I_2(\delta)$ *(middle curve, solid),* $I_1(\delta)$ *(upper curve, dashed), and* $I_2(\delta)$ *(lower curve, dash-dotted), graphed versus* δ*. Note that zeros* δ *of* $I_1(\delta)$ *that are not even multiples of* π*, such as* $\delta \approx 7.25$*, are also zeros of* $I_2(\delta)$*. Since the plot appearance suggests that* $I_1(\delta), I_2(\delta)$ *might be shifted translates of each other, we illustrate graphically in the right plot that this is not the case.*

To understand why there is an invariant bump size, set $\delta = z_2 - z_1$ in (3.14), such that g becomes

$$g(z_1,\delta) = \epsilon \int_0^\delta w(y)\,[\,\cos z_1\,(\cos y(\cos\delta - 1) + \sin y \sin\delta) + \sin z_1\,(\sin y(\cos\delta + 1) - \cos y \sin\delta)]\; dy. \tag{3.15}$$

The function $g(z_1, \delta)$ has some obvious nontrivial zeros, such as $(z_1, \delta) = ((4m + 3)\pi/4, (4n + 1)\pi/2)$ and $(z_1, \delta) = ((4m + 1)\pi/4, (4n + 3)\pi/2)$ for any integers m, n, but these do not give bumps, as they do not solve (3.10), (3.11).

Note that when $z_1 = 0$, the definition of $g(z_1, \delta)$ in (3.15) reduces to (3.8) for $g(a)$ in the $\phi = 0$ case, which for $w(x)$ given by (2.10) with our usual parameter values has a zero at $a = z^* \approx 7.25$. Indeed, we can factor out $\cos z_1$ from the first term on the right-hand side of equation (3.15) and $\sin z_1$ from the second term to write

$$g(z_1,\delta) = \epsilon(\cos z_1 I_1(\delta) + \sin z_1 I_2(\delta)), \tag{3.16}$$

where $I_1(\delta) = 0$ for $\delta = 2\pi, z^*, 4\pi, \ldots$. If there are zeros of $I_2(\delta)$ within this set, then these represent potential bump sizes that are independent of starting position and phase (which were encoded in z_1). Numerical experiments suggest that the zeros of $I_1(\delta)$ that are not even multiples of π are also zeros of $I_2(\delta)$; see Figure 13.

Remark 3.3. Although we have not explored what happens with coupling functions $w(x)$ other than that given in (2.10), the form of (3.14), (3.15) strongly suggests that the phenomena observed here do not depend on the exact form of $w(x)$.

Remark 3.4. For general $p(x) = 1 + \epsilon p_0(x)$ with $p_0(x)$ not necessarily periodic, (3.10), (3.11) still apply, with bumps occurring when both are satisfied. One can also

solve (3.12) to find candidate bump endpoints b_1, b_2, with solutions independent of ϵ as above. Thus, we expect that for each fixed $p_0(x)$ and bump starting position b_1, there will be a small number of possible bump sizes selected. We do not know whether or not there will exist an invariant size, independent of starting position, for general $p_0(x)$, however.

4. Discussion. In this paper, we consider localized, stationary activity bump solutions of the rate model (1.1), which describes the evolution of activity in a neuronal population. Following previous work stemming from [1], we take $f(u(y,t)) = H(u(y,t))$, the Heaviside step function, which is an analytically tractable case that has been shown to organize solution structure for some more general forms of f [15]. A key new feature in this paper is that the coupling function $w(x)$ represents off-center coupling. Off-center coupling models the effective pattern of synaptic inputs to an excitatory population in an excitatory-inhibitory (E-I) network with no recurrent excitation, but rather E to I, I to E, and I to I connections.

In this setting, under certain assumptions, we prove that for nonzero h, (1.1) has exactly one time-independent, localized solution satisfying $u(x) > 0$ if and only if $x \in (0, a)$ for a positive, finite constant a. This shows that coupling need not be locally positive to allow for the existence of such a sustained, localized solution. Earlier results showed that the combination of recurrent excitation and long-range inhibition yields the existence of a pair of bump solutions, a linearly stable wider one and an unstable narrower one, to (1.1) [1, 13, 17, 4, 5]. Here we find that for off-center coupling, the unstable bump does not exist, while the single bump that does exist is linearly stable. The nonlinear stability of these bump solutions remains open for investigation.

We show that the range of activity levels h over which bumps can exist, and correspondingly the range of possible bump widths, is finite. Since there is at most a single bump for each h, this brings up the question of how bumps are born and disappear as h varies. We have discussed two types of mechanisms by which this may occur. One mechanism, which can apply to bump birth or death, is the appearance of a point or points inside $(0, a)$ at which u becomes negative. This fits in well with our results showing that a bump can develop an internal local minimum while remaining a valid bump, with $u > 0$ on $(0, a)$. The second mechanism, which can generate only bump death, not bump birth, is a loss of positivity at the edges $x = 0$ and $x = a$ of u. Numerically, we observe bump birth via the former mechanism and bump death via the latter. These mechanisms will not occur when the coupling function $w(x)$ is not off-center (i.e., when there is recurrent excitation). Further, we do not consider temporally dynamic solutions. It is possible that there may be interactions of time-dependent solutions with stationary bumps, which remain to be explored.

We also do not consider temporal details of synaptic dynamics. Our results require sufficiently strong long-range inhibition for bumps to exist with off-center coupling. Thus, our analysis supports the idea that when long-range inhibition is weak [10], slow synaptic dynamics may be necessary to allow for localized activity [20]. Even richer forms of pattern formation can be expected when models incorporating such additional features are considered in future work.

In section 4, we allow for spatial variations in coupling strength, which may correspond to regional structural variations in the brain [2]. Numerically, we observe that this induces bump pinning, such that for each fixed starting position, bumps exist for only a small, discrete set of background input levels h, each with a single corresponding width. Moreover, a unique invariant width is selected, which is possible

at all starting positions. The form of the relevant equations suggests that these results do not depend on the exact form of the coupling function $w(x)$ or on the fact that it is off-center, although this remains to be thoroughly explored. We provide mathematical insight into this size invariance (e.g., Figure 13), but we do not provide analytical proof that this size invariance must occur.

Is it physiologically plausible that spatial inhomogeneities in coupling strength could so severely limit the possible background input levels needed for bumps? We can only speculate on this issue. Since it is believed that attention has significant effects on neuronal activity across wide areas (for example, [11]), it seems possible that the background input level to a brain region could be related to attention. We know from experience that attention is needed to allow for effective working memory or navigation, for example; one needs to first pay attention to a stimulus if one wishes to remember it, and one needs to maintain focus on the memory of this stimulus to keep it "in mind" until it is internalized. Perhaps attention is the process of bringing overall network activity in an appropriate brain region to a level at which a bump can form and subsequently maintaining this level to sustain the localized bump. Since bump sizes are selected by integral conditions relating the spatially homogeneous and inhomogeneous parts of the coupling pattern, perhaps some part of cognitive decline with aging or disease could be associated with a loss of effectiveness of a subset of synaptic connections, which could compromise the "orthogonality" of the system.

Similarly, while a severe limitation on the number of possible bump sizes might initially seem computationally restrictive, there would be advantages to this limitation. In particular, suppose that only a unique bump size were realizable in a certain brain area and that bumps were always symmetric about their centers. If an activity level $u > 0$ were observed from one cell in that area (e.g., by a neuron postsynaptic to it from another area), this would immediately indicate the exact distance of the presynaptic cell from the center of any bump to which it belonged, and activity levels of two cells would suffice to indicate exactly which other cells were in the bump and with which activity levels. This allows for highly efficient decoding by the postsynaptic cell. Note that we observe the development of asymmetric bumps when the coupling is spatially inhomogeneous and the bump length is not an even integer multiple of 2π. Even without symmetry, inputs from a small number of cells in a bump would effectively convey information about the entire bump. Of course, this requires that the postsynaptic cell somehow "knows" that a bump exists in the presynaptic area, and is highly speculative, but nonetheless it suggests that there might be some computational relevance to the bump pinning phenomenon that we have observed.

5. Appendix.

5.1. Coupling profile. The activity levels $u_E(x,t)$ and $u_I(x,t)$ of coupled excitatory and inhibitory populations satisfy the model equations [28, 1, 19]

$$\begin{aligned} \tfrac{\partial u_E}{\partial t} &= -u_E + w_{EE} * f_E(u_E) - w_{IE} * f_I(u_I) + h_E, \\ \tau \tfrac{\partial u_I}{\partial t} &= -u_I + w_{EI} * f_E(u_E) - w_{II} * f_I(u_I) + h_I, \end{aligned} \tag{5.1}$$

where $w * f(u)$ denotes the convolution $\int_{-\infty}^{\infty} w(x-y) f(u(y,t))\, dy$, $f_i(u)$ is the firing rate function for population i, and w_{ij} denotes the synaptic connection function from population i to population j, which we take here to be nonnegative for all i, j. We consider (1.1) to represent a reduction of (5.1), with $w_{EE} \equiv 0$, to a single equation for the activity level of the excitatory population. The connection function $w(x)$ that we

consider in (1.1), as shown in Figure 1, corresponds to the time-independent input to the excitatory population that results when precisely those excitatory cells at $x = 0$ are active.

To derive the function shown in Figure 1, we therefore set the time derivatives in system (5.1) to zero. This gives $u_E = h_E - w_{IE} * f_I(u_I)$. We assume $h_E > 0$ and aim for an activity profile of u_E which has the form of $w(x)$ shown in Figure 1. This will imply that the activity of cells at $x = 0$ has the desired effect on the activity of the other cells in the excitatory population. For simplicity, assume that $w_{IE}(x)$ has a simple profile; for example, suppose that each inhibitory cell inhibits only those excitatory cells that share its x-coordinate. Then we seek an activity profile of u_I which has the qualitative form of $-w(x)$, for $w(x)$ shown in Figure 1.

Time-independence implies that $u_I = h_I + w_{EI} * f_E(u_E) - w_{II} * f_I(u_I)$. Further, the assumption that only those cells at $x = 0$ are active gives

$$u_I(x) = h_I + w_{EI}(x) - w_{II} * f_I(u_I), \tag{5.2}$$

although other positive coefficients of w_{EI} may result from non-Heaviside choices of f_E. Thus, the mathematical justification of off-center coupling for (1.1), as in Figure 1, may be achieved by finding a consistent solution of (5.2) having the qualitative form of $-w(x)$, for an appropriate firing rate function f_I. Note that (5.2) has a form very similar to that of the steady state equation (2.1) analyzed in this paper, but with a spatially varying input function, as studied, for example, in [1]. The desired solution would be positive on $(-\infty, -b) \cup (-a, a) \cup (b, \infty)$ for some $b > a > 0$. The proof of the existence of such a solution remains open.

5.2. Derivation of ODE. Recall that for $p(x) = 1 + \epsilon(1 + \cos x)$, we define

$$g(a) = \int_0^a w(a - \eta)p(\eta)\, d\eta - \int_0^a w(\eta)p(\eta)\, d\eta\,. \tag{5.3}$$

Thus, using integration by parts, the fact that $w(x)$ is even, and the fact that $w'(0) = 0$, we have

$$\begin{aligned} g'(a) &= p(a)(w(0) - w(a)) + \textstyle\int_0^a w'(a - \eta)p(\eta)\, d\eta \\ &= w(a)(p(0) - p(a)) + \textstyle\int_0^a w(a - \eta)p'(\eta)\, d\eta. \end{aligned} \tag{5.4}$$

Similarly,

$$g''(a) = -w(a)p'(a) + w'(a)(p(0) - p(a)) + \int_0^a w(\eta - a)p''(\eta)\, d\eta \tag{5.5}$$

and

$$\begin{aligned} g'''(a) = &- 2w'(a)p'(a) - w(a)p''(a) + w''(a)(p(0) - p(a)) + w(a)p''(0) \\ &+ \int_0^a w(\eta - a)p'''(\eta)\, d\eta\,. \end{aligned} \tag{5.6}$$

But since

$$p'''(\eta) = -p'(\eta),$$

(5.6) and (5.4) can be combined to give

$$g''' + g' = w(a)[p(0) - p(a) + p''(0) - p''(a)] - 2w'(a)p'(a) + w''(a)(p(0) - p(a))\,. \tag{5.7}$$

Finally, the fact that

$$p''(x) = -\epsilon \cos x$$

yields $p(x) + p''(x) = 1 + \epsilon$ for any x, such that $p(0) + p''(0) - (p(a) + p''(a)) = 0$. Thus, the ODE (5.7) simplifies to

$$g''' + g' = -2w'(a)p'(a) + w''(a)(p(0) - p(a)) = \epsilon(2w'(a)\sin a + w''(a)(1 - \cos a)). \tag{5.8}$$

Note that from (5.3), (5.4), (5.5), (5.6), it follows that

$$g(0) = g'(0) = g''(0) = g'''(0) = 0\,.$$

REFERENCES

[1] S. Amari, *Dynamics of pattern formation in lateral-inhibition type neural fields*, Biol. Cybernet., 27 (1977), pp. 77–87.

[2] P. C. Bressloff, *Travelling fronts and wave propagation failure in an inhomogeneous neural network*, Phys. D, 155 (2001), pp. 83–100.

[3] P. C. Bressloff, *Bloch waves, periodic feature maps, and cortical pattern formation*, Phys. D, 89, (2002), 088101.

[4] C. Chow and Y. Guo, *Existence and stability of standing pulses in neural networks* I*: Existence*, SIAM J. Applied Dynamical Systems, submitted.

[5] C. Chow and Y. Guo, *Existence and stability of standing pulses in neural networks* II*: Stability*, SIAM J. Applied Dynamical Systems, submitted.

[6] E. P. Christian and F. E. Dudek, *Electrophysiological evidence from glutamate microapplications for local excitatory circuits in the CA*1 *area of the rat hippocampus*, J. Neurophysiol., 59 (1988), pp. 110–123.

[7] C. L. Colby, J. R. Duhamel, and M. E. Goldberg, *Oculocentric spatial representation in parietal cortex*, Cereb. Cortex, 5 (1995), pp. 470–481.

[8] A. Compte, N. Brunel, P. Goldman-Rakic, and X.-J. Wang, *Synaptic mechanisms and network dynamics underlying spatial working memory*, Cereb. Cortex, 10 (2000), pp. 910–923.

[9] C. L. Cox, J. R. Huguenard, and D. A. Prince, *Heterogeneous axonal arborizations of rat thalamic reticular neurons in the ventrobasal nucleus*, J. Comp. Neurol., 366 (1996), pp. 416–430.

[10] C. L. Cox, J. R. Huguenard, and D. A. Prince, *Nucleus reticularis neurons mediate diverse inhibitory effects in thalamus*, Proc. Nat. Acad. Sci. USA, 94 (1997), pp. 8854–8859.

[11] A. K. Engel, P. Fries, and W. Singer, *Dynamic predictions: Oscillations and synchrony in top-down processing*, Nat. Rev. Neurosci., 2 (2001), pp. 704–716.

[12] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, SIAM, Philadelphia, 2002.

[13] G. B. Ermentrout, *Neural nets as spatio-temporal pattern forming systems*, Rep. Prog. Phys., 61 (1998), pp. 353–430.

[14] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic, *Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex*, J. Neurophysiol., 61 (1989), pp. 331–349.

[15] K. Kishimoto and S. Amari, *Existence and stability of local excitations in homogeneous neural fields*, J. Math. Biol., 7 (1979), pp. 303–318.

[16] C. R. Laing, W. C. Troy, B. Gutkin, and G. B. Ermentrout, *Multiple bumps in a neuronal model of working memory*, SIAM J. Appl. Math., 63 (2002), pp. 62–97.

[17] C. R. Laing and C. C. Chow, *Stationary bumps in networks of spiking neurons*, Neural Comp., 13 (2001), pp. 1473–1494.

[18] E. K. Miller, C. A. Erickson, and R. Desimone, *Neural mechanisms of visual working memory in prefrontal cortex of the macaque*, J. Neurosci., 16 (1996), pp. 5154–5167.

[19] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* II*. Lateral inhibition and standing pulses*, SIAM J. Appl. Math, 62 (2001), pp. 226–243.

[20] J. Rubin, D. Terman, and C. Chow, *Localized bumps of activity sustained by inhibition in a two-layer thalamic network*, J. Comp. Neurosci., 10 (2001), pp. 313–331.

[21] P. E. Sharp, H. T. Blair, and J. Cho, *The anatomical and computational basis of the rat head-direction cell signal*, Trends in Neurosci., 24 (2001), pp. 289–294.

[22] E. Shink, M. D. Bevan, J. P. Bolam, and Y. Smith, *The subthalamic nucleus and the external pallidum: Two tightly interconnected structures that control the output of the basal ganglia in the monkey*, Neuroscience, 73 (1996), pp. 335–357.

[23] M. Steriade, E. G. Jones, and R. R. Llinás, *Thalamic Oscillations and Signaling*, Wiley, New York, 1990.

[24] J. S. Taube, *Head direction cells recorded in the anterior thalamic nuclei of freely moving rats*, J. Neurosci., 15 (1995), pp. 70–86.

[25] J. S. Taube, *Head direction cells and the neurophysiological basis for a sense of direction*, Prog. Neurobiol., 55 (1998), pp. 225–256.

[26] D. Terman, J. E. Rubin, A. C. Yew, and C. J. Wilson, *Activity patterns in a model for the subthalamopallidal network of the basal ganglia*, J. Neurosci., 22 (2002), pp. 2963–2976.

[27] R. Traub and R. Miles, *Neuronal Networks of the Hippocampus*, Cambridge University Press, Cambridge, UK, 1991.

[28] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetic, 13 (1973), pp. 55–80.

[29] L. Zhang, *Existence and Asymptotic Stability of Traveling Wave Solutions of Neuronal Network Equations*, Ph.D. thesis, The Ohio State University, Columbus, OH, 1999.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1636–1667

A MATHEMATICAL MODEL FOR THE SULPHUR DIOXIDE AGGRESSION TO CALCIUM CARBONATE STONES: NUMERICAL APPROXIMATION AND ASYMPTOTIC ANALYSIS*

D. AREGBA-DRIOLLET†, F. DIELE‡, AND R. NATALINI§

Abstract. We introduce a degenerate nonlinear parabolic system that describes the chemical aggression of calcium carbonate stones under the attack of sulphur dioxide. For this system, we present some finite element and finite difference schemes to approximate its solutions. Numerical stability is given under suitable CFL conditions. Finally, by means of a formal scaling, the qualitative behavior of the solutions for large times is investigated, and a numerical verification of this asymptotics is given. Our results are in qualitative agreement with the experimental behavior observed in the chemical literature.

Key words. chemical aggression, porous media, nonlinear parabolic equations, fast reaction limit, free boundary problems, finite element, finite differences

AMS subject classifications. Primary, 65M06; Secondary, 76M20, 76R, 82C40

DOI. 10.1137/S003613990342829X

1. Introduction.

In this paper we introduce and investigate a differential model to describe the evolution of the chemical action of SO_2 (sulphur dioxide) in $CaCO_3$ (calcium carbonate) stones. Let Ω be a given region in $\mathbb{R}^n$, with $n = 1, 2, 3$, namely, our stone specimen. Our basic equations read, in their adimensional form, as

$$\begin{cases} \partial_t(\varphi(c)s) - \nabla \cdot (\varphi(c)\nabla s) = -\varphi(c)sc, \\ \\ \partial_t c = -\varphi(c)sc, \end{cases} \tag{1.1}$$

for $x \in \Omega$ and $t \in \mathbb{R}$. Here c and s are both nonnegative, since c stands for the local density of $CaCO_3$ and s for the porous concentration of SO_2, namely, the concentration taken with respect to the volume of the pores; here the porosity φ is a linear function of the density c. For this problem we also have to specify the initial and boundary conditions, according to the problem under examination.

There is an extensive chemical literature about the deterioration mechanisms of natural building stones [22, 23, 14, 26, 18, 8] in connection with problems concerning both modern and historical buildings. Acidity in the air is essentially caused by pollutants, such as sulphur and nitrogen oxides, which are emitted into the atmosphere by sources related to industry, transportation, and heating. These species are transformed, through complex reaction pathways, into gaseous nitric and nitrous acids and into acidic sulphates as suspended particles. Although in recent years the levels of

*Received by the editors May 20, 2003; accepted for publication (in revised form) December 19, 2003; published electronically June 22, 2004. The research activity reported in this paper was partially conducted within the European Union RTN project FRONTS-SINGULARITIES: HPRN-CT-2002-00274 and the European Union RTN HYKE project HPRN-CT-2002-00282.

http://www.siam.org/journals/siap/64-5/42829.html

†Mathématiques Appliquées de Bordeaux, Université Bordeaux 1, 351 cours de la Libération, F-33405 Talence, France (aregba@math.u-bordeaux.fr).

‡Istituto per le Applicazioni del Calcolo "M. Picone," Consiglio Nazionale delle Ricerche, Sez. Bari, Via Amendola 122/I, I-70126 Bari, Italy (f.diele@iac.cnr.it).

§Istituto per le Applicazioni del Calcolo "M. Picone," Consiglio Nazionale delle Ricerche, Viale del Policlinico 137, I-00161 Rome, Italy (r.natalini@iac.cnr.it).

pollution in the urban areas of Europe have decreased, levels of HNO_3 and other aggressive pollutants such as sulphur dioxide and ozone have remained consistent. As is well known, SO_2 and NO_3 react with calcium carbonate stones to form sulphates and nitrates, which, due to their solubility in water, may be drained away or, if protected from the rain, may form crusts, which eventually exfoliate; see [8, 5]. Observe that the chemical deterioration is mainly expected to occur when the surface is wet. There is in fact a strong experimental relationship between deterioration and time of wetness [14].

Effective simulation tools seem to be crucial in considering the fine-scale evolution of reaction pathways, possibly in complex geometries, as requested by an improved policy of prevention and monitoring of chemical damage on historical monuments. For instance, it should be important to assist stakeholders to assign a degree of priority for an optimal scheduling of cleaning operations, also taking into account the local geometry and the exposure of the concerned stones. Actually, the standard methods used for studying the evolution of this kind of damage have been the development of models of atmospheric corrosion; they are based on the statistic determination on the ratio of dose and response of the materials. For instance, the Lipfert formula [19] was applied, using an extended database containing values taken in the field (i.e., meteorological value and pollution of the air). If this procedure could be meaningful for the determination of corrosion for civil uses, this approach is clearly insufficient for artistic and historical artwork.

In this paper we introduce a different approach in the framework of hydrodynamic models by using some basic physical relations, the balance laws of the chemical reactions, and the Fick law, and by neglecting the permeability of the medium. As a particular feature, the model takes into account the effects of sulphation on carbonate rocks, by assuming a direct (linear) dependence of porosity and diffusivity on the density of calcium carbonate. The main issue of our model will be a proper determination of the thickness of the gypsum crust ($CaSO_4 \cdot 2H_2O$) formed as a product of the reaction of SO_2 with calcium carbonate stones. There are two main advantages in this approach: it is possible to solve numerically the equations by finite element or finite difference methods, and also in several space dimensions and for geometrically complex domains; time asymptotic analysis in one space dimension yields a precise characterization of the behavior of the limit solutions, which are expressed in terms of a simple free boundary problem.

Global existence of smooth solutions for this system is considered in a separate work [12]. A more general model, which includes convective effects due to the pressure gradient, for stones with a greater permeability, will be introduced and studied in [1].

The paper is organized as follows. In section 2, we introduce in some detail a basic model starting from the main ideas of macroscopic modeling of filtration in porous media [3, 21]. Next, we propose some different numerical approximations, to both finite element and finite difference schemes, for the one-dimensional case, by proving some rigorous nonlinear stability and positivity results, under some proper CFL conditions. Numerical tests are given for comparison of the accuracy of different schemes. Finally, we investigate the asymptotic behavior of solutions and assess the agreement of the model with experimental tests. By scaling arguments, it is possible to show that the limit profile of solutions, for the case of the half-line $x > 0$ and with Dirichlet boundary conditions for s at $x = 0$, is given by the solution of a simple one-phase Stefan problem [20]. This approach is inspired by a related paper [15], where a similar model was considered, but with constant porosity and diffusivity.

The analytical verification of this asymptotic behavior will be considered in a future work. Here we just give a numerical verification, which actually yields quite precise information on the propagation of the main front and on the rate of convergence of the solution toward its limit profile.

In particular, it is worth mentioning that the limit profile has a free boundary $\zeta(t)$ which gives the right limit of the gypsum crust and evolves according to a diffusive law:

$$\zeta(t) = C\sqrt{t}. \tag{1.2}$$

This behavior is in good qualitative agreement with experimental data (see [22, 23, 18]) and in particular with the results of the new laboratory tests performed in [10], and the possibility of a successful calibration of the model against laboratory and in situ tests is shown.

2. Derivation of the model. In [18, 8] the authors conducted experimental studies involving the exposure of different types of marble to 10 and 300 ppm SO_2 atmospheres. It was then possible to estimate practically the extent of damage to marble due to an industrial environment. Analogous experiments were conducted with dolomite rocks [26]. Other laboratory tests and in situ measurements can be found in [22, 23, 6, 5].

Considering the mathematical description of the time evolution of the sulphation process, some models were proposed in [18] to give some measurements of the main physicochemical parameters. The different regions and time regimes were described by different parameters and then matched to fit the experimental behavior of the reaction. Here we develop a single mathematical model that in principle can take into account the full behavior of the solutions.

The path of reaction of SO_2 with calcite is revealed by the X-ray diffraction counts, which suggest that the reaction occurs in the following manner [2, 18]:

$$CaCO_3 + SO_2 + \frac{1}{2}H_2O \to CaSO_3 \cdot \frac{1}{2}H_2O + CO_2, \tag{2.1}$$

$$CaSO_3 \cdot \frac{1}{2}H_2O + \frac{1}{2}O_2 + \frac{3}{2}H_2O \to CaSO_4 \cdot 2H_2O \tag{2.2}$$

(see also [6, 9, 7, 26, 25]). The $CaSO_3$ (calcium sulphite) reaches equilibrium soon after the initial reaction, and then the amount of gypsum continues to increase with the progress of the reaction. Therefore we can assume a simplified one-step reaction:

$$CaCO_3 + SO_2 + \frac{1}{2}O_2 + 2H_2O \to CaSO_4 \cdot 2H_2O + CO_2. \tag{2.3}$$

We neglect all heat effects and assume the air contains enough water to give rise to the reaction. Moreover, we assume that the change in concentrations of oxygen (O_2), water (H_2O), and carbon dioxide (CO_2) does not affect the reaction.

Let Ω be the domain occupied by the specimen of calcite under consideration and set ρ_s for the concentration of SO_2; c for the density of $CaCO_3$; and g for the density of $CaSO_4 \cdot 2H_2O$. All quantities are defined with respect to the whole volume, incorporating both solid and gaseous material, and depend on the position $x \in \Omega$ and on time t. Since we are including the bulk volume in the definition of c, g, ρ_s, we will call them total concentrations or densities. We will define later the porous concentrations for SO_2.

Following [3], we assume that the total concentrations ρ_s and c satisfy the balance laws,

$$\partial_t \rho_s + \nabla \cdot (\rho_s \mathbf{V}_s) = \dot{r}_s, \tag{2.4}$$

$$\partial_t c = \dot{r}_c, \tag{2.5}$$

where $\mathbf{V}_s$ is the sulphur dioxide "fluid" velocity and $\dot{r}_s$, $\dot{r}_c$ are rates of production (or consumption) of sulphur dioxide and calcite.

Following the usual model for the rate of production, we have

$$\dot{r}_s = -m_s \omega, \qquad \dot{r}_c = -m_c \omega. \tag{2.6}$$

Here, m_s, m_c are the masses of single molecules of sulphur dioxide and calcite, and, to complete the notations, let us set m_g for the molecular mass of gypsum. The quantity $\omega > 0$ measures the rate of reaction and according to [18] is given by

$$\omega = A \left(\frac{\rho_s}{m_s} \right) \left(\frac{c}{m_c} \right). \tag{2.7}$$

In general, the constant A depends on the temperature and on the activation energy. In the present paper we shall neglect this dependence.

Assuming initial densities c_0 and g_0, for calcite and gypsum, we have the relation

$$c + \frac{m_c}{m_g} g = c_0 + \frac{m_c}{m_g} g_0, \tag{2.8}$$

which expresses the density of gypsum as a function of calcite.

Next, we introduce the porosity of the calcite specimen φ, which cannot be assumed constant, since the transformation of calcite in gypsum alters the volume of void (occupied by air and sulphur dioxide). Therefore, following [24], it is reasonable to regard it as a function of the amount of gypsum or, equivalently, as a function of the amount of calcite, that is, $\varphi = \varphi(c)$.

Let φ_0 be the porosity of the pure calcite specimen, i.e., for $c = c_0,\ g = 0$, and $\varphi_{\tilde{g}}$ the porosity of the final sulphate product, when all the calcium carbonate has been converted in gypsum, namely, when $c = 0$, $\tilde{g} = \frac{m_g}{m_c} c_0$. Then, according to the rigorous derivation in [1], we can express the porosity of the specimen during the reaction as a linear combination of these porosities:

$$\varphi(c) = \varphi_{\tilde{g}} + (\varphi_0 - \varphi_{\tilde{g}}) \frac{c}{c_0}. \tag{2.9}$$

We denote by s the porous concentration of SO_2, which is defined as the concentration taken with respect to the volume of the pores, which is related to the total concentration by

$$\rho_s = \varphi(c) s. \tag{2.10}$$

The seepage velocity $\mathbf{v}_s$ is related to the fluid velocity $\mathbf{V}_s$ by the classical Dupuit–Forchheimer relation

$$\mathbf{v}_s = \varphi(c) \mathbf{V}_s. \tag{2.11}$$

The balance laws for ρ_s and c become

$$\partial_t(\varphi(c)s) + \nabla \cdot (s\mathbf{v}_s) = -\left(\frac{A}{m_c}\right)\varphi(c)sc, \tag{2.12}$$

$$\partial_t c = -\left(\frac{A}{m_s}\right)\varphi(c)sc. \tag{2.13}$$

To close the system (2.12)–(2.13), we need an expression for the seepage velocity $\mathbf{v}_s$. In the following we make the main assumption of our model, namely, that all the contributions given by the pressure gradient to the seepage velocity can be neglected. As shown in [1], this corresponds to a zero permeability limit, which is a realistic assumption for many species of marble.

Therefore, we shall express $\mathbf{v}_s$ by the classical Fick law [21],

$$s\mathbf{v}_s = -D(c)\nabla s, \tag{2.14}$$

where $D(c) = d\varphi(c)$, and d is the (scalar) effective molecular diffusive coefficient. This yields

$$\partial_t(\varphi(c)s) = -\left(\frac{A}{m_c}\right)\varphi(c)sc + d\nabla \cdot (\varphi(c)\nabla s), \tag{2.15}$$

$$\partial_t c = -\left(\frac{A}{m_s}\right)\varphi(c)sc. \tag{2.16}$$

System (2.15)–(2.16) forms a closed set of nonlinear degenerate parabolic differential equations which have to be supplemented by initial conditions at time $t = 0$ for s and c, and by Dirichlet or Neumann boundary conditions for s. Let us also notice that in general we do not expect to give any boundary condition for c. Clearly, it is also possible to consider system (2.15)–(2.16) as a one parabolic equation coupled with an ordinary differential equation. Unfortunately, it is difficult to use this remark, since we have to take into account the strong coupling between s and c into the divergence term.

It is easy to see that we can recover the scaled model (1.1) just by taking the new variables

$$y = \sqrt{\frac{A}{dm_c}}x, \quad \tau = \frac{A}{m_c}t, \quad \tilde{s} = \frac{m_c}{m_s}s. \tag{2.17}$$

To improve the physical accuracy of our model, it should be possible to consider three main modifications. The first is to assume the dependence of the reaction rate A on the internal temperature and degree of wetness by introducing two supplementary equations related to the evolution of these quantities for some given initial and boundary conditions.

Another important modification arises if, according to [13], we consider a more general nonlinear Fick law,

$$s\mathbf{v}_s(Bs\,|\mathbf{v}_s| + 1) = -D\nabla s, \tag{2.18}$$

where B is the high concentration coefficient. The form (2.18) of the Fick law is best suited when the gradient of the concentration is very high.

Finally, let us mention that a more accurate model, which takes into account both the pressure gradient effects due to the Darcy law and the diffusivity given by the Fick law, will be considered in [1].

3. Numerical approximation in one space dimension. In this section we construct numerical schemes for the one-dimensional version of model (1.1),

$$\begin{cases} \partial_t \rho_s - \partial_x(\varphi(c)\partial_x s) = -\rho_s c, \\ \partial_t c = -\rho_s c \end{cases} \tag{3.1}$$

for $x \in [0,1]$, $t \geq 0$ and where we recall that $\rho_s = \varphi(c)s$ and $\varphi(c)$ is given by (2.9). In the following, we shall assume that the initial calcite density c_0 is a positive constant. Then, setting $\alpha = \frac{1}{c_0}(\varphi_0 - \varphi_{\tilde{g}})$ and $\beta = \varphi_{\tilde{g}}$, we can rewrite the function $\varphi(c)$ in the following form:

$$\varphi(c) = \alpha c + \beta. \tag{3.2}$$

In what follows, we shall assume that

$$\varphi_0 > \varphi_{\tilde{g}}, \tag{3.3}$$

with $\alpha, \beta > 0$ and $0 < \beta \leq \varphi(c) \leq \alpha c_0 + \beta < 1$. The case $\varphi_0 < \varphi_{\tilde{g}}$ is similar and can be considered by using the same arguments.

As initial conditions we have

$$\begin{cases} \rho_s(x,0) = 0, \\ c(x,0) = c_0, \end{cases} \tag{3.4}$$

where c_0 is a positive constant, and we impose the boundary conditions,

$$\begin{cases} \rho_s(0,t) = \rho_{s0}, \\ \dfrac{\partial \rho_s}{\partial x}(1,t) = 0. \end{cases} \tag{3.5}$$

Here ρ_{s0} is a positive constant. The case where ρ_{s0} is a bounded measurable positive function can be treated in the same way.

It is easy to see that s satisfies the same initial condition as ρ_s, and the boundary conditions read as

$$\begin{cases} s(0,t) = \dfrac{\rho_{s0}}{\varphi(c(0,t))}, \\ \dfrac{\partial s}{\partial x}(1,t) = 0 \end{cases} \tag{3.6}$$

with

$$c(0,t) = c_0 \mathrm{e}^{-\rho_{s0} t}. \tag{3.7}$$

Two methods of solving this problem are under consideration. First, we use a finite element method, looking for s as a continuous piecewise linear function and c as a piecewise constant function. The second method is a finite difference scheme, where the main unknowns are ρ_s and c. For some problems in one space dimension, finite element methods have an interpretation in terms of finite differences via mass lumping and staggered variables. However, the philosophies are quite different, and our aim is to study afterward two- or three-dimensional systems, with possibly complicated geometries. On the other hand, for the methods presented in this paper it is easier to modify the time discretization for the finite difference schemes. We compare these approaches and analyze the effect of time discretization modifications.

3.1. A finite element method. Following [12], we look for a smooth solution (s, c) of (3.1), (3.4), (3.6).

3.1.1. The scheme. To deal with the nonhomogeneous Dirichlet condition, we introduce the unknown $\sigma(x,t) = s(x,t) - s(0,t)$, and we denote $H = \{u \in H^1(]0,1[), u(0) = 0\}$. The first equation of (3.1) can be written as

$$\partial_t(\varphi\sigma) - \partial_x(\varphi(c)\partial_x\sigma) = F(\sigma, c, t),$$

and a variational formulation of the problem is as follows: find $(\sigma, c) \in C^1([0, +\infty[, H \times L^2(]0,1[))$ such that for all $(p, q) \in H \times L^2(]0,1[)$,

$$\left\{ \begin{array}{l} \partial_t \displaystyle\int_{]0,1[} \varphi\sigma p dx + \int_{]0,1[} \varphi\partial_x\sigma\partial_x p dx = \int_{]0,1[} pF dx, \\ \\ \partial_t \displaystyle\int_{]0,1[} cq dx = -\int_{]0,1[} \varphi(\sigma + s(0,.))cq dx. \end{array} \right. \tag{3.8}$$

Let us define a regular mesh

$$[0,1] = \cup_{1\le i\le N}[x_i, x_{i+1}], \quad x_i = (i-1)\Delta x, \quad \Delta x = 1/N.$$

We denote by $\{p_i, i = 1, \dots, N+1\}$ the classical P1 basis functions:

$$p_i(x) = \left\{ \begin{array}{ll} \frac{x - x_{i-1}}{\Delta x} & \text{if} \quad x \in [x_{i-1}, x_i], \\ \\ \frac{x_{i+1} - x}{\Delta x} & \text{if} \quad x \in [x_i, x_{i+1}], \\ \\ 0 & \text{else.} \end{array} \right.$$

For $i = 1, \dots, N$, we denote by q_i the characteristic function of $[x_i, x_{i+1}[$.

The solution (s, c) is approximated by

$$s_h(x,t) = \sum_{i=1}^{N+1} \xi_i(t)p_i(x), \quad c_h(x,t) = \sum_{k=1}^{N} \eta_k(t)q_k(x).$$

Moreover, we define $\rho_{s,h}(x,t) = \varphi(c_h(x,t))s_h(x,t)$.

The unknowns for s_h are located on the nodes, while the ones for c_h can be considered as approximations of the mean value of c on each cell $[x_i, x_{i+1}]$. The Dirichlet boundary condition at $x = 0$ is taken into account by putting $\xi_1(t) = \rho_{s0}/\varphi(\eta_1(t))$, according to (3.6), so that $\sigma_h = s_h - \xi_1 p_1$ and $(\sigma_h, c_h) \in C^1([0, +\infty[, V \times W)$, where $V = \text{lin}\{p_i, i = 2, \dots, N+1\}$ and $W = \text{lin}\{q_i, i = 1, \dots, N\}$.

In practice, as usual for this kind of problem we write a (false) variational approach formulation with $N+1$ basis functions $\{p_1, \dots, p_{N+1}\}$, and we put the Dirichlet condition a posteriori: for all $i = 1, \dots, N+1$, for all $k = 1, \dots, N$,

$$\left\{ \begin{array}{l} \partial_t \displaystyle\int_{]0,1[} \varphi_h s_h p_i dx + \int_{]0,1[} \varphi_h\partial_x s_h\partial_x p_i dx = -\int_{]0,1[} \varphi_h s_h c_h p_i dx, \\ \\ \partial_t \displaystyle\int_{]0,1[} c_h q_k dx = -\int_{]0,1[} \varphi_h s_h c_h q_k dx. \end{array} \right. \tag{3.9}$$

In this formula, φ_h is defined by

$$\varphi_h = \varphi(c_h).$$

The first set of equations in (3.9) is a differential system which is linear with respect to ξ:

$$\partial_t \sum_{j=1}^{N+1} \xi_j \int_{]0,1[} \varphi_h p_j p_i dx + \sum_{j=1}^{N+1} \xi_j \int_{]0,1[} \varphi_h \partial_x p_j \partial_x p_i dx$$

$$= - \sum_{j=1}^{N+1} \xi_j \int_{]0,1[} \varphi_h c_h p_j p_i dx, \quad i = 1, \dots, N+1.$$

As $\varphi_h = \sum_{k=1}^{N} \varphi(\eta_k) q_k$, it can be summarized as

$$\partial_t (M(\eta)\xi) + K(\eta)\xi = 0. \tag{3.10}$$

Denoting $\varphi(\eta_k) = \varphi_k$, the matrix $M(\eta)$ is defined by

$$M(\eta) = \frac{\Delta x}{6} \begin{pmatrix} 2\varphi_1 & \varphi_1 & 0 & \dots & 0 \\ \varphi_1 & 2(\varphi_1+\varphi_2) & \varphi_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \varphi_{N-1} & 2(\varphi_{N-1}+\varphi_N) & \varphi_N \\ 0 & \dots & 0 & \varphi_N & 2\varphi_N \end{pmatrix},$$

and the matrix $K(\eta)$ is defined by

$$K(\eta) = \tfrac{1}{\Delta x} \begin{pmatrix} \varphi_1 & -\varphi_1 & 0 & \dots & 0 \\ -\varphi_1 & \varphi_1+\varphi_2 & -\varphi_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & -\varphi_{N-1} & \varphi_{N-1}+\varphi_N & -\varphi_N \\ 0 & \dots & 0 & -\varphi_N & \varphi_N \end{pmatrix}$$

$$+ \tfrac{\Delta x}{6} \begin{pmatrix} 2\varphi_1\eta_1 & \varphi_1\eta_1 & 0 & \dots & 0 \\ \varphi_1\eta_1 & 2(\varphi_1\eta_1+\varphi_2\eta_2) & \varphi_2\eta_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \varphi_{N-1}\eta_{N-1} & 2(\varphi_{N-1}\eta_{N-1}+\varphi_N\eta_N) & \varphi_N\eta_N \\ 0 & \dots & 0 & \varphi_N\eta_N & 2\varphi_N\eta_N \end{pmatrix}.$$

The second equation of problem (3.9) can be written as

$$\Delta x\, \partial_t \eta_k = -\eta_k(\alpha\eta_k + \beta) \sum_{j=1}^{N+1} \xi_j(t) \int_{]x_k, x_{k+1}[} p_j(x) dx, \quad k = 1, \dots, N$$

or, equivalently,

$$\partial_t \eta_k = -\frac{(\xi_k + \xi_{k+1})}{2} \eta_k(\alpha\eta_k + \beta) = -\gamma_k \eta_k(\alpha\eta_k + \beta), \quad k = 1, \dots, N. \tag{3.11}$$

We now discretize in time. The approximated quantities at time $t_n = n\Delta t$ are denoted with a superscript n, and for $t \in [t_n, t_{n+1}[$ we put $s_h(x,t) = s_h^n(x)$, $c_h(x,t) = c_h^n(x)$.

In view of initial and boundary data we take

$$\begin{cases} \xi_1^0 = \dfrac{\rho_{s0}}{\varphi(c_0)}, \quad \xi_j^0 = 0 \quad \text{for } j = 2, \dots, N+1, \\ \\ \eta_k^0 = c_0 \qquad\qquad\qquad\quad \text{for } k = 1, \dots, N. \end{cases}$$

We first discretize (3.11) by fixing $\xi = \xi^n$ and solving exactly the equation. As α and β are positive, if the ξ_j^n and η_k^n are nonnegative, without any time step restriction, we obtain the intermediate value $\eta_k^{n+1/2}$:

$$\eta_k^{n+1/2} = \beta \eta_k^n \frac{\mathrm{e}^{-\gamma_k \beta \Delta t}}{\alpha \eta_k^n + \beta - \alpha \eta_k^n \mathrm{e}^{-\gamma_k \beta \Delta t}}. \tag{3.12}$$

Then, we solve the differential system (3.10) by the θ method ($\theta \in [0,1]$):

$$\frac{M(\eta^{n+1/2})\xi^{n+1} - M(\eta^n)\xi^n}{\Delta t} + (1-\theta)K(\eta^n)\xi^n + \theta K(\eta^{n+1/2})\xi^{n+1} = 0. \tag{3.13}$$

This is a linear tridiagonal system $UX = G$.

The boundary condition is taken into account by replacing g_1 by $\rho_{s0}/\varphi(\eta_1^{n+1/2})$, g_2 by $g_2 - u_{21}\rho_{s0}/\varphi(\eta_1^{n+1/2})$, u_{1j} by δ_{1j}, and u_{i1} by δ_{i1}. This modified linear system is symmetric and positive and can be easily solved, for example, by a Choleski method.

This method is first order in time, even if $\theta = 1/2$, because the matrices M and K depend on $\eta(t)$. When ξ^{n+1} is computed we can set

$$\eta_k^{n+1} = \eta_k^{n+1/2}$$

or improve the approximation with the second order Heun method: we solve exactly (3.11) with $\xi = \xi^{n+1}$ and initial value $\eta_k^{n+1/2}$. We obtain a second intermediate value $\eta_k^{n+1/2,1}$, and, finally, we set

$$\eta_k^{n+1} = \frac{\eta_k^n + \eta_k^{n+1/2,1}}{2}. \tag{3.14}$$

In the following, we call this finite element method FE-θ.

3.1.2. Stability. It is well known that the discrete maximum principle does not always hold in the finite element method. We have the following stability results.

PROPOSITION 3.1. *Suppose that* $\theta \in]1/3, 1]$ *and* $\Delta x^2 < \frac{3(3\theta-1)}{c_0}$. *If the time step satisfies the condition*

$$\frac{\Delta x^2}{\theta(6 - c_0 \Delta x^2)} < \Delta t \leq \frac{\Delta x^2}{(1-\theta)(3 + c_0 \Delta x^2)}, \tag{3.15}$$

then for all $x \in [0,1]$, $c_h(x,.)$ *is a nonincreasing function of* t *and for all* $t \geq 0$,

$$\rho_{s,h}(x,t) \geq 0, \quad c_h(x,t) \in]0, c_0].$$

Moreover, the condition (3.15) *is not empty.*

Proof. The result is true for $t \in [0, t_1[$. Suppose that for all $j = 1, \dots, N+1$ and all $k = 1, \dots, N$ we have $\xi_j^n \geq 0$ and $\eta_k^n \in]0, c_0]$. In view of (3.11) and (3.12), $0 < \eta_k^{n+1/2} \leq \eta_k^n$, and thus $0 < \varphi_k^{n+1/2} \leq \varphi_k^n$.

After the modifications due to boundary conditions, the system (3.13) has an $N\times N$ symmetric irreducibly diagonally dominant matrix, still denoted $U = M^{n+1/2}+ \Delta t\,\theta\, K^{n+1/2}$. Hence, U^{-1} is positive if $u_{ij} < 0$ for $j = i \pm 1$ and $u_{ii} > 0$ for all i [27].

In the right-hand side, we have for $i \geq 2$,

$$g_i = \sum_{j=i-1}^{i+1} \left(m_{ij}^n - (1-\theta)\Delta t k_{ij}^n\right) \xi_j^n, \quad \xi_1^n = \rho_{s0}/\varphi_1^n,$$

and

$$\begin{aligned} g_{2,\text{modified}} &= g_2 - \frac{u_{21}\rho_{s0}}{\varphi_1^{n+1/2}} \\ &= \sum_{j=2}^{3} \left(m_{2j}^n - (1-\theta)\Delta t k_{2j}^n\right) \xi_j^n \\ &\quad + \frac{\Delta t}{\Delta x}\left[1 - \frac{\Delta x^2}{6}\eta_1^n + \theta\frac{\Delta x^2}{6}(\eta_1^n - \eta_1^{n+1/2})\right]\rho_{s0}. \end{aligned} \tag{3.16}$$

Due to the condition on Δx and the fact that η is nonincreasing, the third line is nonnegative. Consequently, the positivity is preserved if the following requirements are fulfilled:

$$\begin{aligned} &m_{ij}^{n+1/2} + \theta\Delta t k_{ij}^{n+1/2} < 0 \qquad \text{for } j = i \pm 1, \;\; i \geq 2, \\ &m_{ii}^{n+1/2} + \theta\Delta t k_{ii}^{n+1/2} > 0 \qquad \text{for all } i \geq 2, \\ &m_{ij}^{n} - (1-\theta)\Delta t k_{ij}^{n} \geq 0 \qquad \text{for all } i \geq 2,\ j. \end{aligned} \tag{3.17}$$

For $j \neq i$, we have to consider $j = i - 1$:

$$k_{ij} = -\frac{\varphi_{i-1}}{\Delta x} + \frac{\Delta x}{6}\varphi_{i-1}\eta_{i-1} \leq 0.$$

Moreover, k_{ii} is positive. Hence, the second requirement of (3.17) is fulfilled. So is the third for $i \neq j$.

Consequently, we have to satisfy for all i and $j = i \pm 1$,

$$-\frac{m_{ij}^{n+1/2}}{\theta k_{ij}^{n+1/2}} < \Delta t \leq \frac{m_{ii}^n}{(1-\theta)k_{ii}^n}\,. \tag{3.18}$$

We have for $2 \leq i \leq N$,

$$\begin{aligned} \frac{m_{ii}}{(1-\theta)k_{ii}} &= \frac{\Delta x(\varphi_{i-1}+\varphi_i)}{3(1-\theta)\left(\dfrac{\varphi_{i-1}+\varphi_i}{\Delta x} + \dfrac{\Delta x(\varphi_{i-1}\eta_{i-1}+\varphi_i\eta_i)}{3}\right)} \\ &\geq \frac{\Delta x^2}{(1-\theta)\,(3 + c_0\Delta x^2)}\,. \end{aligned} \tag{3.19}$$

For $i = N + 1$ the same lower estimate holds. For $j = i - 1$,

$$\begin{aligned} -\frac{m_{ij}}{\theta k_{ij}} &= \frac{\Delta x\,\varphi_{i-1}}{6\theta\left(\dfrac{\varphi_{i-1}}{\Delta x} - \dfrac{\Delta x\varphi_{i-1}\eta_{i-1}}{6}\right)} \\ &\leq \frac{\Delta x^2}{\theta\,(6 - c_0\Delta x^2)}. \end{aligned} \tag{3.20}$$

The fact that condition (3.15) is not empty, i.e.,

$$\frac{\Delta x^2}{\theta(6-c_0\Delta x^2)} < \frac{\Delta x^2}{(1-\theta)(3+c_0\Delta x^2)},$$

is ensured by the conditions $\Delta x^2 < \frac{3(3\theta-1)}{c_0}$ and $\theta > 1/3$.

As it is enough to consider the case $\eta^{n+1} = \eta^{n+1/2}$, the proof is complete.

Remark 3.1. As usual, there is no upper limitation on Δt in the fully implicit case $\theta = 1$.

As far as we are concerned by the lower limitation on Δt, let us point out that, as is well known, if the matrix M is lumped, this limitation disappears.

If the data are small enough, we can prove a uniform bound for $\rho_{s,h}$. We fix the initial condition c_0, and we determine for which values of ρ_{s0} the bound exists.

In what follows, we denote $\varphi_0 = \varphi(c_0)$.

PROPOSITION 3.2. *We make the same assumptions as in Proposition* 3.1, *and we take* Δt *satisfying condition* (3.15). *If, moreover, the two inequalities*

$$\text{(i)} \qquad \Delta t < \frac{\beta^2 - \alpha\varphi_0\rho_{s0}}{\beta^2\rho_{s0}},$$

$$\text{(ii)} \qquad \rho_{s0} < \frac{\beta^2}{\alpha\varphi_0},$$

are true, then the numerical solution has the additional bound

$$0 \le \rho_{s,h}(x,t) \le \frac{\varphi_0}{\beta}\rho_{s0}. \tag{3.21}$$

Let us point out that condition (i) *on* Δt *does not restrict condition* (3.15) *if the space step* Δx *is small enough.*

Proof. By Proposition 3.1 positivity is preserved, and for all $n \ge 0$ and $i \in \{1,\dots,N\}$ we have

$$0 < \eta_i^{n+1/2} \le \eta_i^n \le c_0. \tag{3.22}$$

Let us denote $X_n = \max\{\xi_j^n, 1 \le j \le N+1\}$. For all $(x,t) \in [x_i, x_{i+1}[\times[t_n, t_{n+1}[$

$$\rho_{s,h}(x,t) = \varphi(\eta_i^n)\left[\xi_i^n p_i(x) + \xi_{i+1}^n p_{i+1}(x)\right].$$

Therefore,

$$0 \le \rho_{s,h}(x,t) \le \varphi_0 X_n,$$

and we have to show that

$$X_n \le \frac{\rho_{s0}}{\beta}. \tag{3.23}$$

This inequality is true for $n = 0$:

$$X_0 = \xi_1^0 = \frac{\rho_{s0}}{\varphi_0} \le \frac{\rho_{s0}}{\beta}.$$

Consider $n \ge 0$ and suppose that X_n satisfies (3.23). If $X_{n+1} = \xi_1^{n+1}$, then

$$X_{n+1} = \frac{\rho_{s0}}{\varphi_1^{n+1/2}} \le \frac{\rho_{s0}}{\beta}.$$

Suppose now that $X_{n+1} = \xi_i^{n+1}$ with $2 \le i \le N$. Our goal is to show that $X_{n+1} \le X_n$. By (3.17), we have

$$\sum_{j=i-1}^{i+1} \left(m_{ij}^{n+1/2} + \Delta t \theta k_{ij}^{n+1/2} \right) X_{n+1} \le \sum_{j=i-1}^{i+1} \left(m_{ij}^{n} - \Delta t (1-\theta) k_{ij}^{n} \right) X_n.$$

This inequality reads as

$$\begin{aligned} X_{n+1} \left[(\varphi_{i-1}^{n+1/2} + \varphi_i^{n+1/2}) + \Delta t \theta \left(\varphi_{i-1}^{n+1/2} \eta_{i-1}^{n+1/2} + \varphi_i^{n+1/2} \eta_i^{n+1/2} \right) \right] \\ \le X_n \left[(\varphi_{i-1}^{n} + \varphi_i^{n}) - \Delta t (1-\theta) \left(\varphi_{i-1}^{n} \eta_{i-1}^{n} + \varphi_i^{n} \eta_i^{n} \right) \right]. \end{aligned} \tag{3.24}$$

Using (3.22) and denoting

$$B = \frac{2}{\Delta x} \sum_{j=i-1}^{i+1} \left(m_{ij}^{n+1/2} + \Delta t \theta k_{ij}^{n+1/2} \right), \quad e_i = \eta_i^n - \eta_i^{n+1/2},$$

we obtain

$$B X_{n+1} \le X_n \left[B + \alpha (e_{i-1} + e_i) - \Delta t \left(\varphi_{i-1}^{n+1/2} \eta_{i-1}^{n+1/2} + \varphi_i^{n+1/2} \eta_i^{n+1/2} \right) \right].$$

Hence, $X_{n+1} \le X_n$ as soon as for all $i = 1, \dots, N$,

$$\alpha e_i - \Delta t \varphi_i^{n+1/2} \eta_i^{n+1/2} \le 0. \tag{3.25}$$

By (3.12) we have

$$e_i = \eta_i^n \frac{\varphi_i^n (1 - \mathrm{e}^{-\gamma_i \beta \Delta t})}{\alpha \eta_i^n + \beta - \alpha \eta_i^n \mathrm{e}^{-\gamma_i \beta \Delta t}},$$

and (3.25) can be written as

$$\alpha \varphi_i^n (1 - \mathrm{e}^{-\gamma_i \beta \Delta t}) - \Delta t \varphi_i^{n+1/2} \beta \mathrm{e}^{-\gamma_i \beta \Delta t} \le 0.$$

We are led to prove that $g(\Delta t) \le 0$ with $g(\tau) = \alpha \varphi_0 (1 - \mathrm{e}^{-\gamma_i \beta \tau}) - \tau \beta^2 \mathrm{e}^{-\gamma_i \beta \tau}$.

We have $g(0) = 0$, and recalling that $\gamma_i = \frac{1}{2}(\xi_i^n + \xi_{i+1}^n)$, it is easy to see that $g'(\tau) < 0$ as soon as conditions (i) and (ii) are satisfied.

The case where $X_{n+1} = \xi_{N+1}^{n+1}$ is identical.

As it is enough to consider the case $\eta^{n+1} = \eta^{n+1/2}$, the proof is complete.

3.2. Finite difference schemes. Here the main variables are $u = (\rho_s, c)$. Denoting $S(u) = -\rho_s c$ we can write

$$\begin{cases} \partial_t \rho_s - \partial_x \left(\varphi(c) \partial_x \dfrac{\rho_s}{\varphi(c)} \right) = S(u), \\ \\ \partial_t c = S(u). \end{cases} \tag{3.26}$$

3.2.1. The schemes. We again mesh $[0,1]$ with a step $\Delta x = 1/N$, and we denote

$$\lambda = \frac{\Delta t}{\Delta x}, \mu = \frac{\Delta t}{\Delta x^2}, x_{m-1/2} = m\Delta x, x_m = (m-0.5)\Delta x, X = (x_m)_{1\le m\le N}.$$

We look for u_m^n, an approximation of $\frac{1}{\Delta x}\int_{x_{m-1/2}}^{x_{m+1/2}} u(x,t_n)dx$, and we set $\mathbf{u}(t) = (\mathbf{u}_1(t),\dots,\mathbf{u}_m(t))^T$, a smooth function such that $\mathbf{u}_m(t_n) = u_m^n$. We also denote by $u^n = (u_m^n)_{1\le m\le N}$ the approximation of u at the time t_n.

It is clear that c_m^n plays the same role as η_k^n in the finite element method, while there is a staggering from vertices to cell centers between s and ρ_s.

The most simple, consistent approximation of $\partial_x\left(a(x)\partial_x r\right)$ by means of Taylor expansions is the following first order one:

$$\Delta_m(a,r) := \frac{(a_m + a_{m+1})(r_{m+1} - r_m) - (a_{m-1} + a_m)(r_m - r_{m-1})}{2\Delta x^2}. \tag{3.27}$$

Hence, a scheme which is comparable to the finite element method FE-θ is the following: $\rho_{s,m}^n$ being fixed, solve exactly the second equation of (3.26) and obtain

$$c_m^{n+1} = c_m^n \mathrm{e}^{-\Delta t \rho_{s,m}^n}.$$

Then discretize the first equation of (3.26) as follows:

$$\frac{\rho_{s,m}^{n+1} - \rho_{s,m}^n}{\Delta t} - \Delta_m\left(\varphi^n, \frac{\rho_s^n}{\varphi^n}\right) = S((1-\theta)\rho_{s,m}^n + \theta\rho_{s,m}^{n+1}, c_m^{n+1}) \tag{3.28}$$

with $\theta \in [0,1]$. In the following, we call this scheme FD1. We should have put a combination of explicit and implicit terms in the derivatives as well, but this leads to solving a linear system to find ρ_s, and we want to avoid that because we lose the simplicity of the approach. Let us remark, moreover, that considering the scheme

$$\begin{aligned}&\frac{\rho_{s,m}^{n+1} - \rho_{s,m}^n}{\Delta t} - (1-\theta)\Delta_m\left(\varphi^n, \frac{\rho_s^n}{\varphi^n}\right) - \theta\Delta_m\left(\varphi^{n+1}, \frac{\rho_s^{n+1}}{\varphi^{n+1}}\right)\\ &\quad = S((1-\theta)\rho_{s,m}^n + \theta\rho_{s,m}^{n+1}, c_m^{n+1})\end{aligned}$$

leads us to solve a nonsymmetric linear system, while the finite element matrices are symmetric.

A second semi-implicit approximation, which we call FD2, is the following:

$$\begin{cases} \dfrac{\rho_{s,m}^{n+1} - \rho_{s,m}^n}{\Delta t} - \Delta_m\left(\varphi^n, \dfrac{\rho_s^n}{\varphi^n}\right) = \dfrac{1}{2}\left[S(\rho_{s,m}^n, c_m^{n+1}) + S(\rho_{s,m}^{n+1}, c_m^n)\right], \\ \dfrac{c_m^{n+1} - c_m^n}{\Delta t} = \dfrac{1}{2}\left[S(\rho_{s,m}^n, c_m^{n+1}) + S(\rho_{s,m}^{n+1}, c_m^n)\right]. \end{cases} \tag{3.29}$$

Because the source term is quadratic, this scheme has an explicit representation:

$$\begin{cases} \dfrac{\rho_{s,m}^{n+1} - \rho_{s,m}^n}{\Delta t} = \dfrac{S(\rho_{s,m}^n, c_m^n)}{\delta} + \left(1 - \Delta t\dfrac{c_m^n}{2\delta}\right)\Delta_m\left(\varphi^n, \dfrac{\rho_s^n}{\varphi^n}\right), \\ \dfrac{c_m^{n+1} - c_m^n}{\Delta t} = \dfrac{S(\rho_{s,m}^n, c_m^n)}{\delta} - \Delta t\dfrac{c_m^n}{2\delta}\Delta_m\left(\varphi^n, \dfrac{\rho_s^n}{\varphi^n}\right) \end{cases} \tag{3.30}$$

with $\delta = 1 + \dfrac{\Delta t}{2}\left(c_m^n + \rho_{s,m}^n\right)$.

This discretization takes the interaction into account in a symmetric way, without computational cost. Notice also that for the single equation $y' = y^2$, the scheme $y^{n+1} = y^n + \Delta t y^{n+1} y^n$ is exact:

$$y^{n+1} = \frac{y^n}{1 - \Delta t y^n} \ .$$

Finally, we construct another scheme by remarking that

$$\Delta_m \left(\varphi, \frac{\rho_s}{\varphi} \right) = \Delta_m \left(1, \rho_s \right)$$

$$- \frac{1}{2\Delta x} \left[\left(\frac{\rho_{s,i}}{\varphi_i} + \frac{\rho_{s,i+1}}{\varphi_{i+1}} \right) \frac{\varphi_{i+1} - \varphi_i}{\Delta x} - \left(\frac{\rho_{s,i-1}}{\varphi_{i-1}} + \frac{\rho_{s,i}}{\varphi_i} \right) \frac{\varphi_i - \varphi_{i-1}}{\Delta x} \right].$$

This is a consistent centered approximation of $\partial_x(\varphi \partial_x \frac{\rho_s}{\varphi}) = \partial_{xx}\rho_s - \partial_x(\frac{\rho_s}{\varphi}\partial_x\varphi)$.

More generally, we can put system (3.26) under the semiconservative form,

$$\left\{ \begin{array}{l} \partial_t \rho_s + f(\rho_s, c, x, t)_x = B(\rho_s, c)_{xx} + S(\rho_s, c), \\ \\ \partial_t c = S(c, \rho_s), \end{array} \right. \tag{3.31}$$

where

$$B(\rho_s, c) = \rho_s, \quad f(\rho_s, c, x, t) = \rho_s \frac{\varphi'(c)}{\varphi(c)} c_x.$$

In the expression of f, c_x is considered as a known function of (x, t). In practice, a difference formula is used to compute it. Hence in what follows we no longer mention the (x, t) dependence in f. Let us denote $\mathbf{f} = (f, 0)$, $\mathbf{B} = (B, 0)$, $\mathbf{S} = (S, S)$. System (3.31) can be written as

$$\partial_t u + \mathbf{f}(u)_x = \mathbf{B}(u)_{xx} + \mathbf{S}(u). \tag{3.32}$$

Now the convective part may be approximated by any method for conservation law. The particular form of our system allows the flux vector splitting,

$$\mathbf{f}(u) = \mathbf{f}_+(u) - \mathbf{f}_-(u),$$

where $sp(\partial_u \mathbf{f}_\pm(u)) \subset [0, +\infty[$ and $\partial_u \mathbf{f}(u)$, $\partial_u \mathbf{f}_+(u)$ and $\partial_u \mathbf{f}_-(u)$ have a common basis of eigenvectors. In fact we have

$$\partial_u \mathbf{f} = \begin{pmatrix} \partial_{\rho_s} f & \partial_c f \\ 0 & 0 \end{pmatrix},$$

and it is sufficient to take

$$\mathbf{f}_+(u) = \begin{pmatrix} f_+(u) \\ 0 \end{pmatrix}, \quad \mathbf{f}_-(u) = \begin{pmatrix} f_-(u) \\ 0 \end{pmatrix},$$

where

$$f_+(u) = \left\{ \begin{array}{ll} f(u) & \text{if } \partial_{\rho_s} f(u) > 0, \\ 0 & \text{else} \end{array} \right. \quad \text{and} \quad f_-(u) = \left\{ \begin{array}{ll} -f(u) & \text{if } \partial_{\rho_s} f(u) < 0, \\ 0 & \text{else.} \end{array} \right.$$

Now, we can take the source term into account either like in FD1 or like in FD2. For example, let us take the same method as for FD1. The numerical scheme FD3 is the following:

(3.33)
$$\begin{cases} \rho_{s,m}^{n+1} = \rho_{s,m}^n - \lambda\left[-f_-(u_{m+1}^n) + f_+(u_m^n) + f_-(u_m^n) - f_+(u_{m-1}^n)\right] \\ \qquad\qquad +\mu\left[\rho_{s,m-1}^n - 2\rho_{s,m}^n + \rho_{s,m+1}^n\right] + \Delta t \;\; S((1-\theta)\rho_{s,m}^n + \theta\rho_{s,m}^{n+1}, c_m^{n+1}), \\ \\ c_m^{n+1} = c_m^n \mathrm{e}^{-\Delta t \rho_{s,m}^n}. \end{cases}$$

To take into account the initial and boundary conditions in each of these numerical schemes, we put

(3.34)
$$\begin{cases} \rho_{s,m}^0 = 0, \quad c_m^0 = c_0 & \text{for } 1 \le m \le N, \\ \rho_{s,0}^n = \rho_{s0}, \quad \rho_{s,N+1}^n = \rho_{s,N}^n & \text{for } n \ge 0. \end{cases}$$

In the following, if $\rho_{s,m} \ge 0$, we take

(3.35)
$$f_\pm(u_m) = \rho_{s,m}\left[\pm\frac{\varphi_{m+1} - \varphi_{m-1}}{2\varphi_m \Delta x}\right]_+ .$$

3.2.2. Higher order in time. To reach higher order in time for these three schemes, we remark that they all can be written in the conservative form:

$$\frac{u_m^{n+1} - u_m^n}{\Delta t} = G_m(u^{n+1}, u^n, \Delta t).$$

Moreover, when Δt tends to zero and u^n and Δx are fixed, u^{n+1} tends to u^n, and $G_m(u^{n+1}, u^n, \Delta t)$ has a limit $F_m(u^n)$. We obtain the semidiscretized scheme:

$$\mathbf{u}_m'(t_n) = F_m(\mathbf{u}(t_n)), \quad m = 1, \dots, N.$$

For FD1 and FD2, F_m is defined by

$$F_m(\mathbf{u}) = \left(\Delta_m\left(\varphi(\mathbf{c}), \frac{\rho_{\mathbf{s}}}{\varphi(\mathbf{c})}\right) + \mathbf{S}(\mathbf{u}), \mathbf{S}(\mathbf{u})\right).$$

For FD3, F_m is defined by

$$F_m(\mathbf{u}) = \left(\overline{\Delta}_m + \mathbf{S}(\mathbf{u}), \mathbf{S}(\mathbf{u})\right)$$

with

$$\overline{\Delta}_m = \frac{-f_-(u_{m+1}) + f_+(u_m) + f_-(u_m) - f_+(u_{m-1})}{\Delta x} + \frac{\rho_{s,m-1} - 2\rho_{s,m} + \rho_{s,m+1}}{\Delta x^2}.$$

This is a new starting point at which to apply a temporal scheme, solving on $[t_n, t_{n+1}]$ the ordinary differential system:

$$\mathbf{u}'(t) = F(\mathbf{u}(t)).$$

As a particular case, we may use an implicit method in order to have a large time step, and this involves the resolution of a nonlinear system. As we already have an implicit scheme with the finite element method, we prefer here to approximate u by

means of the second order Heun method or by the optimal third order strong-stability preserving (SSP) Runge–Kutta (RK) method given by

$$\begin{aligned} v^{(1)} &= u^n + \Delta t F(u^n), \\ v^{(2)} &= \frac{3}{4}u^n + \frac{1}{4}v^{(1)} + \frac{1}{4}\Delta t F(v^{(1)}), \\ u^{n+1} &= \frac{1}{3}u^n + \frac{2}{3}v^{(2)} + \frac{2}{3}\Delta t F(v^{(2)}) \end{aligned} \tag{3.36}$$

introduced in [11].

3.2.3. Stability. The finite difference schemes FD1, FD3 preserve positivity under a suitable time step restriction.

The choice of the third order RK scheme (3.36) was dictated by further stability consideration. In fact, as shown by Gottlieb, Shu, and Tadmor in [11], it is possible to relate the CFL condition for the temporal first order Euler scheme (3.33) to a strong stability property verified by the temporal third order scheme (3.36), at least for conservation laws. The rigorous extension of this property to the present case is beyond the aims of this paper, but we can expect that this scheme works for the same Δt used in the Euler time discretization, thanks to the dissipation induced by the diffusion and by the source terms.

PROPOSITION 3.3 (scheme FD1). *For all $n \geq 0$, all $m = 1, \ldots, N$,*

$$\rho_{s,m}^n \geq 0, \quad c_m^n \in [0, c_0] \tag{3.37}$$

under the time step restriction

$$\Delta t \leq \frac{\beta \Delta x^2}{\varphi_0 + \beta(1 + \Delta x^2 c_0(1-\theta))}. \tag{3.38}$$

Proof. It is clear that we have to check only the first order in time. From the expression of c_m^{n+1}, it is clear that if $\rho_{s,m}^n$ is nonnegative, then $c_m^{n+1} \in [0, c_m^n]$. This is true for $n = 0$. Let us suppose it is true for all $k \leq n$. Then $c_m^{n+1} \in [0, c_0]$ and

$$\beta \leq \varphi_m^n \leq \varphi_0. \tag{3.39}$$

Now we can write

$$\begin{aligned} \rho_{s,m}^{n+1}(1 + \Delta t \theta c_m^{n+1}) = \rho_{s,m}^n &\left[1 - \mu\left(1 + \frac{\varphi_{m+1}^n + \varphi_{m-1}^n}{2\varphi_m^n}\right) - \Delta t(1-\theta)c_m^{n+1}\right] \\ &+ \mu \rho_{s,m-1}^n \frac{\varphi_m^n + \varphi_{m-1}^n}{2\varphi_{m-1}^n} + \mu \rho_{s,m+1}^n \frac{\varphi_m^n + \varphi_{m+1}^n}{2\varphi_{m+1}^n}. \end{aligned} \tag{3.40}$$

Thus, positivity is ensured as soon as condition (3.38) is satisfied.

PROPOSITION 3.4 (scheme FD3 with (3.35)). *For all $n \geq 0$, all $m = 1, \ldots, N$,*

$$\rho_{s,m}^n \geq 0, \quad c_m^n \in [0, c_0] \tag{3.41}$$

under the time step restriction

$$\Delta t \leq \frac{2\beta \Delta x^2}{\varphi_0 + \beta(3 + 2\Delta x^2(1-\theta))}. \tag{3.42}$$

Proof. In view of (3.35) we can write

$$\begin{aligned}\rho_{s,m}^{n+1}(1+\Delta t\,\theta c_m^{n+1}) = \rho_{s,m}^n\left[1-\mu\left(\left|\frac{\varphi_{m+1}-\varphi_{m-1}}{2\varphi_m}\right|+2+\Delta x^2(1-\theta)c_m^{n+1}\right)\right]\\ +\rho_{s,m-1}^n\left(\lambda[\partial_{\rho_s}f(u_{m-1})]_+ +\mu\right)\\ +\rho_{s,m+1}^n\left(\lambda[-\partial_{\rho_s}f(u_{m+1})]_+ +\mu\right),\end{aligned} \tag{3.43}$$

and condition (3.42) follows by a straightforward computation.

Remark 3.2. In practice, we update Δt at each time step, and we use a less restrictive condition. More precisely, we require that in (3.40) or (3.43), for all $m = 1,\dots,N$, the coefficient of $\rho_{s,m}^n$ is nonnegative. By bounding c_m^{n+1} by c_m^n we use the numerical condition

$$\Delta t_n \le \frac{\Delta x^2}{\max_{1\le m\le N}\left[1+\frac{\varphi_{m+1}^n+\varphi_{m-1}^n}{2\varphi_m^n}+\Delta x^2(1-\theta)c_m^n\right]}$$

instead of condition (3.38). Similarly, instead of condition (3.42) we impose that

$$\Delta t_n \le \frac{\Delta x^2}{\max_{1\le m\le N}\left[\left|\frac{\varphi_{m+1}^n-\varphi_{m-1}^n}{2\varphi_m^n}\right|+2+\Delta x^2(1-\theta)c_m^n\right]}\,.$$

4. Numerical experiments. This section is devoted to some numerical experiments. It is more significant to perform our tests on the original unscaled model (2.15)–(2.16), even if we observe that with the change of variable (2.17) the shape of the solutions is not dependent on the given parameters. The data are fixed as follows:

$$\rho_{s0}=1,\ \ c_0=10,\ \ \alpha=0.01,\ \ \beta=0.1,\ \ m_c=100.09,\ \ m_s=64.06,\ \ d=1. \tag{4.1}$$

As shown in section 5, letting A go to infinity is equivalent to making the time go to infinity. First we take $A=1$ and the final time $T_{\max}=0.1$; the last paragraph of this section and section 5 are devoted to time asymptotics.

For $\theta=1$ and $\theta=0.5$, let us analyze the numerical order of accuracy γ defined by

$$\gamma=\log_2\left(\frac{||\rho(h)-\rho(h/2)||_1}{||\rho(h/2)-\rho(h/4)||_1}\right)$$

for each component of the solution at the final time $T_{\max}$. Here, $h=\Delta x$. We recall that schemes FD1 and FD3 are obtained by solving exactly the equation for c, and FD2 is obtained by a semi-implicit method. We call FD4 the scheme obtained by the same spatial discretization as for FD3, while the temporal resolution is the same semi-implicit one as FD2.

We first put $\theta=1$ and for FE, $\eta^{n+1}=\eta^{n+1/2}$. As there is no upper bound on Δt for the finite element method in this case, we have chosen $\Delta t=\Delta x$. The scheme is first order in time. We do not experiment with FD2 and FD4 here, because they correspond to $\theta=1/2$. We obtain Tables 1–3.

Scheme FD1, that is, when the nonlinear diffusion is approximated by formula (3.27), converges better than scheme FD3. For the finite element case, the choice of a large time step prevents one from a fast convergence. Nevertheless, the convergence

TABLE 1
The L^1 errors for FD1, $\theta = 1$.

h	γ_s	$\|\|\rho_s(h) - \rho_s(h/2)\|\|_1$	γ_c	$\|\|c(h) - c(h/2)\|\|_1$
0.2	0.312985829	0.00834144243	1.87384348	0.000525266
0.1	2.40471333	0.00671464464	1.3467364	0.0001433165
0.05	1.33397634	0.00126803776	1.49975333	5.634925E-05
0.025	2.2831098	0.000502996885	3.1590019	1.9925875E-05
0.0125	1.67597534	0.000103342928	0.83185905	2.2308125E-06
0.00625	3.29717645	3.23416303E-05	4.7235384	1.25328125E-06
0.003125	——	3.29012875E-06	——	4.74375001E-08

TABLE 2
The L^1 errors for FD3, $\theta = 1$.

h	γ_s	$\|\|\rho_s(h) - \rho_s(h/2)\|\|_1$	γ_c	$\|\|c(h) - c(h/2)\|\|_1$
0.2	1.58822799	0.045138069	2.35937874	0.000867266
0.1	0.215780607	0.0150120053	-0.396735465	0.0001690085
0.05	0.981610986	0.0129265394	1.01033173	0.000222504
0.025	1.13931055	0.00654617968	1.29691797	0.000110458125
0.0125	0.908544914	0.00297181158	0.833852008	4.4955875E-05
0.00625	1.02937252	0.00158314976	1.06622513	2.52215625E-05
0.003125	——	0.000775621775	——	1.20449844E-05

TABLE 3
The L^1 errors for the finite element method, $\theta = 1$.

h	γ_s	$\|\|\rho_s(h) - \rho_s(h/2)\|\|_1$	γ_c	$\|\|c(h) - c(h/2)\|\|_1$
0.2	-2.25922932	0.00487514386	——	0.
0.1	-0.71418536	0.023339053	3.10173338	0.001743642
0.05	-0.991921129	0.0382890373	-3.52219246	0.00020311525
0.025	1.33297029	0.0761504476	1.61661945	0.00233360913
0.0125	2.08378342	0.0302279307	1.38231411	0.000760986875
0.00625	-0.343775269	0.00713061687	-0.138516672	0.000291916844
0.003125	——	0.00904927466	——	0.000321334109

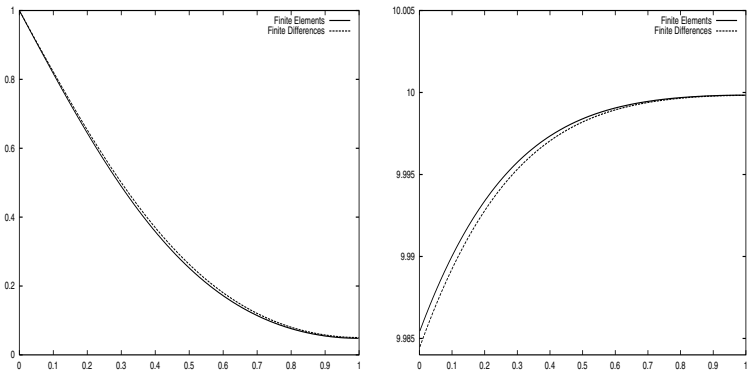


FIG. 1. *Comparison among finite element methods and finite difference methods, with $\triangle x = 0.003125$ and $\theta = 1$. Left: SO_2 concentration; right: calcite density. Time $t = 0.1$.*

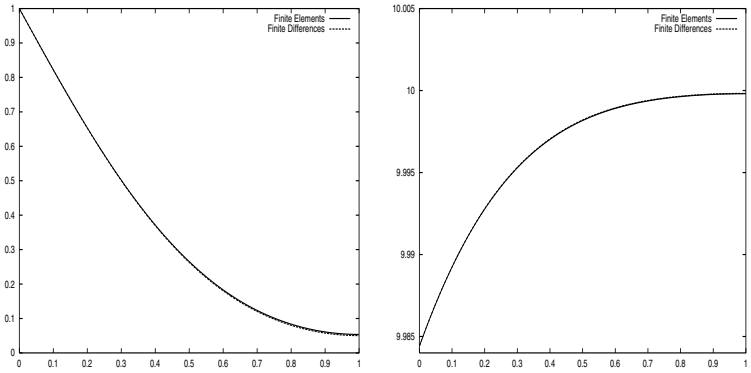


FIG. 2. *Comparison among finite element methods and finite difference methods, with* $\triangle x = 0.0015625$ *and* $\theta = 1$. *Left:* SO_2 *concentration; right: calcite density. Time* $t = 0.1$.

TABLE 4
The L^1 *errors for FD*1, $\theta = 1/2$.

h	γ_s	$\|\|\rho_s(h) - \rho_s(h/2)\|\|_1$	γ_c	$\|\|c(h) - c(h/2)\|\|_1$
0.2	1.17222647	0.0084354273	3.36048996	0.000529403
0.1	2.00077673	0.00374310282	2.001092	5.1544E-05
0.05	2.0139494	0.000935272028	2.0079489	1.287625E-05
0.025	2.0292682	0.000231568118	2.01466401	3.201375E-06
0.0125	2.05711582	5.67293969E-05	2.03945881	7.9225E-07
0.00625	2.09897113	1.36318438E-05	2.04442335	1.9271875E-07
0.003125	——	3.18200844E-06	——	4.67187501E-08

TABLE 5
The L^1 *errors for FD*2, $\theta = 1/2$.

h	γ_s	$\|\|\rho_s(h) - \rho_s(h/2)\|\|_1$	γ_c	$\|\|c(h) - c(h/2)\|\|_1$
0.2	1.17179262	0.00843434345	1.12743968	0.000263196
0.1	2.0006973	0.00374374754	2.01027543	0.000120472
0.05	2.01389756	0.00093548462	2.00610367	2.990425E-05
0.025	2.02923763	0.000231629077	2.00869842	7.4445E-06
0.0125	2.05709045	5.67455331E-05	2.01474625	1.8499375E-06
0.00625	2.09895461	1.36359609E-05	2.03115964	4.5778125E-07
0.003125	——	3.18300594E-06	——	1.12E-07

holds, and as shown in Figures 1 and 2, the solution is close to the one computed by FD1. Let us point out that a computation for $\Delta x = 0.0015625$ and $\Delta t = \Delta x$ is much faster than a computation for $\Delta x = 0.00625$ and $\Delta t = C\Delta x^2$.

Let us now make the same experiment with $\theta = 1/2$. In that case, for the finite element method, we choose the Heun method for η^{n+1} with formula (3.14). Tables 4–8 show that numerically FD1, FD2, and FE-1/2 are second order accurate, while FD3 and FD4 remain less than first order accurate. Let us recall that unless a is a constant, the formula (3.27) used in FD1 and FD2 is only first order in space, and that for FD1 the choice of θ affects only the source term and not the diffusion term. Therefore, second order was not expected. This result can be explained by the fact that here the variations of c and $a = \varphi(c)$ are small in space and in time; see Figure 1, for example. On the contrary, the semiconservative formulation (3.31) used to construct schemes FD3 and FD4 does not allow such a gain.

TABLE 6
The L^1 errors for FD3, $\theta = 1/2$.

h	γ_s	$\|\|\rho_s(h) - \rho_s(h/2)\|\|_1$	γ_c	$\|\|c(h) - c(h/2)\|\|_1$
0.2	1.33149717	0.0452259789	1.7650704	0.000870125
0.1	0.743923424	0.0179707991	0.686674157	0.0002560015
0.05	0.884689522	0.0107306029	0.850347453	0.00015904975
0.025	0.945292297	0.00581173816	0.926766823	8.8217125E-05
0.0125	0.973358008	0.00301817698	0.963707356	4.6405375E-05
0.00625	0.986854034	0.00153721547	0.98201168	2.37937813E-05
0.003125	——	0.00077564336	——	1.20461563E-05

TABLE 7
The L^1 errors for FD4, $\theta = 1/2$.

h	γ_s	$\|\|\rho_s(h) - \rho_s(h/2)\|\|_1$	γ_c	$\|\|c(h) - c(h/2)\|\|_1$
0.2	1.33151869	0.0452259099	1.79104267	0.000677384
0.1	0.743919588	0.0179705035	0.453029589	0.000195739
0.05	0.884682254	0.0107304549	0.766586208	0.00014298875
0.025	0.945286757	0.00581168729	0.890333851	8.404975E-05
0.0125	0.973354679	0.00301816215	0.946713033	4.53439375E-05
0.00625	0.986852184	0.00153721146	0.973711741	2.35250312E-05
0.003125	——	0.000775642332	——	1.19788125E-05

TABLE 8
The L^1 errors for finite element, $\theta = 1/2$.

h	γ_s	$\|\|\rho_s(h) - \rho_s(h/2)\|\|_1$	γ_c	$\|\|c(h) - c(h/2)\|\|_1$
0.2	3.50066927	0.0545620474	6.44459105	0.001924897
0.1	1.99293298	0.00482041247	1.99127884	2.21E-05
0.05	2.00834819	0.00121102077	2.01460837	5.5585E-06
0.025	2.02053637	0.000301008351	2.03328576	1.375625E-06
0.0125	2.03989415	7.41884825E-05	2.07065766	3.360625E-07
0.00625	2.0685759	1.80412722E-05	2.14077583	8.E-08
0.003125	——	4.30094375E-06	——	1.8140625E-08

TABLE 9
The L^1 errors for FD1, $\theta = 1/2$, order 3 in time.

h	γ_s	$\|\|\rho_s(h) - \rho_s(h/2)\|\|_1$	γ_c	$\|\|c(h) - c(h/2)\|\|_1$
0.2	2.34357642	0.0132414991	2.19330999	0.000390955
0.1	1.99675824	0.00260885747	2.01397425	8.5482E-05
0.05	2.01399821	0.000653681548	2.00827172	2.11645E-05
0.025	2.03298887	0.000161842415	2.0117367	5.260875E-06
0.0125	2.0516014	3.95459231E-05	2.02235917	1.3045625E-06
0.00625	2.05713482	9.53911656E-06	2.0462204	3.21125E-07
0.003125	——	2.29218078E-06	——	7.775E-08

TABLE 10
The L^1 errors for FD2, $\theta = 1/2$, order 3 in time.

h	γ_s	$\|\|\rho_s(h) - \rho_s(h/2)\|\|_1$	γ_c	$\|\|c(h) - c(h/2)\|\|_1$
0.2	2.3431196	0.0132405316	-0.16126272	0.000139208
0.1	1.99663095	0.00260949301	2.01341207	0.0001556715
0.05	2.01393858	0.000653898483	2.00595212	3.855775E-05
0.025	2.03295263	0.000161902817	2.00708072	9.59975E-06
0.0125	2.05161953	3.95616763E-05	2.01087687	2.3881875E-06
0.00625	2.05719473	9.54279656E-06	2.02687804	5.925625E-07
0.003125	——	2.29296984E-06	——	1.4540625E-07

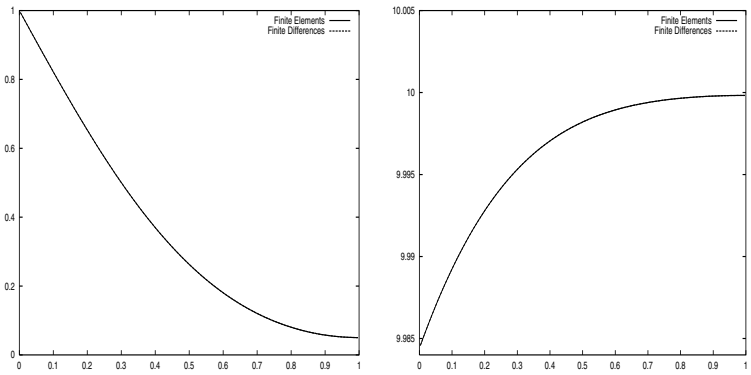


FIG. 3. *Comparison among finite element methods and finite difference methods, with* $\triangle x = 0.00625$ *and* $\theta = 0.5$. *Left:* SO_2 *concentration; right: calcite density. Time* $t = 0.1$.

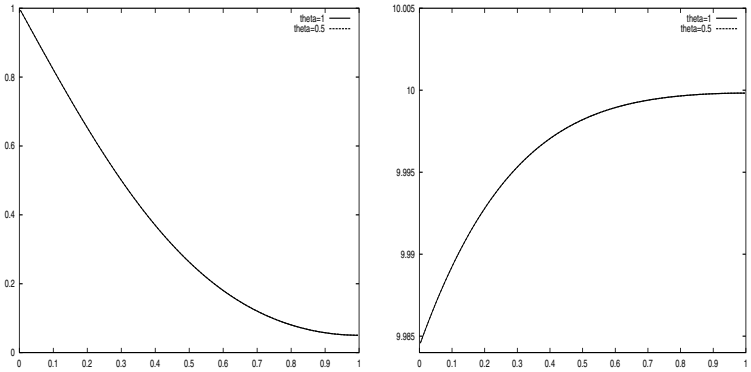


FIG. 4. *Finite differences: comparison among* $\theta = 0.5$ *and* $\theta = 1$, *with* $\triangle x = 0.00625$. *Left:* SO_2 *concentration; right: calcite density. Time* $t = 0.1$.

To end this part of the tests, in Tables 9 and 10 we give the numerical order for the third order time discretization applied to FD1 and FD2. The numerical order of accuracy is not sensitive to time order increasing, but the error $\|u(h) - u(h/2)\|_1$ is smaller. This is not surprising, since the spatial discretization remains unchanged. We do not present graphical results for high order in time discretizations because in all our experiments they coincide with what happens for first order. Taking into account that high time order makes the computation longer, we conclude that one should prefer first order schemes.

To complete these tests, we present some graphical results. In Figure 1, we compare the results for finite element and finite difference FD1 methods, for $\theta = 1$ and $\Delta x = 0.003125$. Then in Figure 2 we make the same comparison for $\Delta x = 0.0015625$, where both results coincide. In Figure 3 we show that for $\theta = 0.5$ finite element and finite difference FD1 methods give very similar results for $\Delta x = 0.00625$. This is because in this case both time steps are comparable, while for $\theta = 1$ we take $\Delta t = \Delta x$ in the finite element method. Hence, the figures confirm the observations made in the tables. We also remark that for finite difference method FD1, the results do not depend on the value of θ; see Figure 4. All these results are computed at time $t = 0.1$.

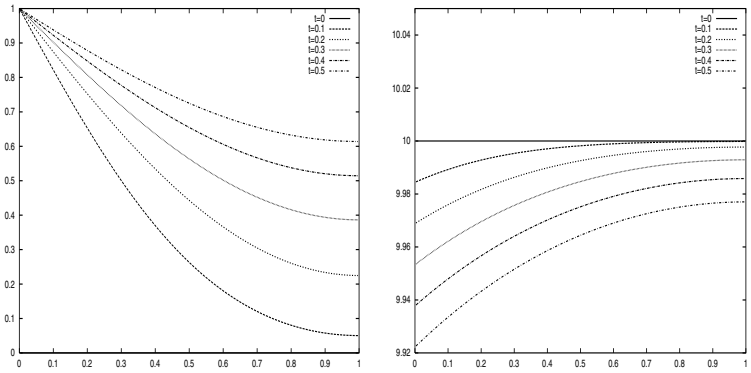


FIG. 5. *Solution from $t = 0$ to $t = 0.5$ with $A = 1$. Left: SO_2 concentration; right: calcite density.*

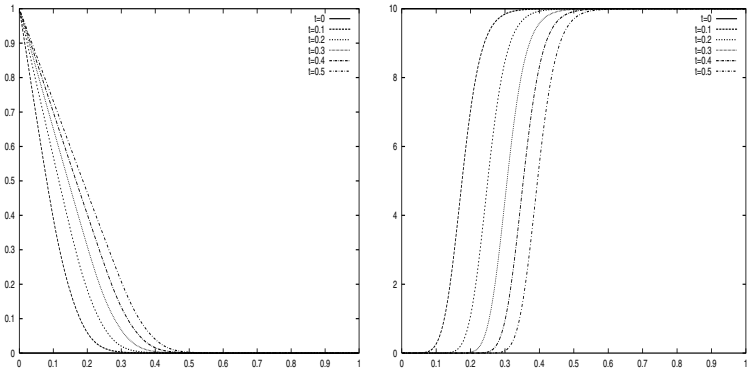


FIG. 6. *Solution from $t = 0$ to $t = 0.5$ with $A = 10000$. Left: SO_2 concentration; right: calcite density.*

From now on, we take $\theta = 0.5$, and we use the finite difference method FD1. Figure 5 shows the solution for different times, from $t = 0$ to $t = 0.5$, with $A = 1$.

Finally, we put $A = 10000$, and we plot the solution from $t = 0$ to $t = 0.5$ (Figure 6). The solution has a very different qualitative aspect: the transition zone is smaller as the interaction coefficient A increases, so that the calcite deterioration is more important for the boundary of the sample, while the interior is not touched by SO_2. The next section is devoted to the study of such solutions.

5. Qualitative behavior of the solutions. In this section we discuss, by means of a formal scaling, the qualitative behavior of the solutions for large times, and we give a numerical verification of this asymptotics. These results will be very useful in calibrating our model against experimental tests.

5.1. A scaling argument. Let us rewrite the system in the one-dimensional case and in the scaled form

$$\begin{cases} \partial_t(\varphi(c)s) - \partial_x(\varphi(c)\partial_x s) = -\varphi(c)sc, \\ \\ \partial_t c = -\varphi(c)sc. \end{cases} \tag{5.1}$$

For simplicity we assume that the domain is the half-line $x > 0$. Therefore we have to give the initial and boundary conditions. Here we consider the simple case of invariant data

$$s(x,0) = 0, \quad c(x,0) = c_0, \tag{5.2}$$

and

$$s(0,t) = \hat{s} \tag{5.3}$$

for two positive constant values $c_0,\ \hat{s}$.

Following [15], we make the scaling (kx, k^2t) in the unknowns, which yields

$$s^k(x,t) = s(kx,k^2t), \quad c^k(x,t) = c(kx,k^2t). \tag{5.4}$$

Clearly, the new unknowns satisfy the scaled problem

$$\begin{cases} \partial_t(\varphi(c)s) - \partial_x(\varphi(c)\partial_x s) = -k^2\varphi(c)sc, \\ \\ \partial_t c = -k^2\varphi(c)sc, \end{cases} \tag{5.5}$$

with the same initial boundary conditions.

Assume now that there exists the limit of the sequence (s^k, c^k) for $k \to \infty$. Namely, there exist (S, C), such that

$$(s^k, c^k) \to_{k\to\infty} (S, C), \tag{5.6}$$

in some suitable (strong) topology. Using the scaling properties of the sequence (s^k, c^k), we have that (S, C) is a self-similar weak solution to the problem

$$\begin{cases} \partial_t(\varphi(C)S - C) - \partial_x(\varphi(C)\partial_x S) = 0, \\ \\ CS = 0, \end{cases} \tag{5.7}$$

with the initial boundary conditions (5.2)–(5.3). Self-similarity is just a consequence of the definition of the limit under the scaling (kx, k^2t). Therefore we have that we can write

$$S(x,t) = \Sigma\left(\frac{x}{\sqrt{t}}\right), \quad C(x,t) = \Gamma\left(\frac{x}{\sqrt{t}}\right), \tag{5.8}$$

where $\Sigma,\ \Gamma$ are one-dimensional functions such that

$$\begin{cases} \frac{1}{2}\xi\left(\varphi(\Gamma)\Sigma - \Gamma\right)' + \left(\varphi(\Gamma)\Sigma'\right)' = 0, \\ \\ \Gamma\Sigma = 0, \end{cases} \tag{5.9}$$

and

$$\Sigma(0) = \hat{s}, \quad \lim_{\xi\to\infty}\Sigma(\xi) = 0, \quad \lim_{\xi\to\infty}\Gamma(\xi) = c_0. \tag{5.10}$$

Let us now give one explicit solution to problem (5.9)–(5.10). Since $\hat{s} > 0$, at least for small values of ξ we have that $\Sigma > 0$ and $\Gamma = 0$. Let $\xi_0 > 0$ be the supremum value of the set $\{\xi > 0 | \Sigma(\xi) > 0\}$. In the interval $(0, \xi_0)$, the function Σ satisfies

$$\frac{1}{2}\xi\Sigma' + \Sigma'' = 0,$$

which implies that, from the condition at infinity,

$$\Sigma(\xi) = \hat{s} - \alpha \int_0^{\xi} e^{-\frac{1}{4}\eta^2} d\eta \tag{5.11}$$

for some $\alpha > 0$. In particular, let ξ_0 be a finite value, i.e., such that $\Sigma(\xi_0) = 0$. Then, clearly,

$$\alpha = \frac{\hat{s}}{\int_0^{\xi_0} e^{-\frac{1}{4}\eta^2} d\eta}. \tag{5.12}$$

Let us continue our solution by setting $\Sigma \equiv 0$ for $\xi \geq \xi_0$. Considering the unknowns S and C, let us denote by $\zeta(t)$ the curve where $S = 0$, which is now given by the equation

$$\zeta(t) = \xi_0 \sqrt{t},$$

which gives

$$\zeta'(t) = \frac{\xi_0}{2\sqrt{t}}.$$

On the other side, we have to satisfy the Rankine–Hugoniot condition for the conservation law (5.9)–(5.10), which yields

$$\zeta'(t) = -\frac{(\partial_x s)(\zeta(t)-)}{c_0}. \tag{5.13}$$

Now, by equating the right-hand side in the equations we obtain

$$\alpha = \frac{\xi_0 c_0}{2} e^{\frac{1}{4}\xi_0^2}. \tag{5.14}$$

Therefore, by using (5.12), we obtain a relation for ξ_0:

$$F(\xi_0) := \xi_0 \int_0^{\xi_0} e^{\frac{1}{4}(\xi_0^2 - \eta^2)} d\eta = \frac{2\hat{s}}{c_0}. \tag{5.15}$$

Since $F(0) = 0$ and $\lim_{\xi \to \infty} F(\xi) = \infty$ and $F' > 0$, there exists $G = F^{-1}$, the inverse function of F, and we take as ξ_0 the unique value $\xi_0 = G(\frac{2\hat{s}}{c_0})$.

Let us resume our situation. There exists a self-similar weak solution of problem (5.9)–(5.10), which is given by

$$(S(x,t),\ C(x,t)) = \begin{cases} S = \hat{s} - \dfrac{\xi_0 c_0}{2} \displaystyle\int_0^{\frac{x}{\sqrt{t}}} e^{\frac{1}{4}(\xi_0^2 - \eta^2)} d\eta,\ \ C = 0,\ x \in (0, \xi_0\sqrt{t}), \\ \\ S = 0,\ C = c_0,\ x > \xi_0\sqrt{t}, \end{cases} \tag{5.16}$$

where ξ_0 is the unique solution of (5.15). If we restrict our attention to the unknown S, we find that it is just a weak solution to the one-phase Stefan problem,

$$\begin{cases} \partial_t S - \partial_{xx} S = 0 \ \ \text{for } x \in (0, \zeta(t)), \\ \\ S(x,0) = 0, \\ \\ S(\zeta(t), 0) = 0, \\ \\ \zeta'(t) = -\frac{(\partial_x s)(\zeta(t)-)}{c_0}. \end{cases} \tag{5.17}$$

As is well known [20], this problem has a unique explicit solution. Therefore we can find that the limit problem (5.8) also has a unique self-similar weak solution.

Let us now investigate the relation between this scaling and the asymptotic behavior of the solution (s, c) to (5.1), (5.2), (5.3). Assuming the limit (5.6), we have that, fixing $t = 1$,

$$(s(kx, k^2), c(kx, k^2)) = (s^k(x, 1), c^k(x, 1)) \to (S(x, 1), C(x, 1)) \quad \text{as } k \to \infty.$$

Now, setting $y = kx$ and $\tau = k^2$, we find that

$$\left(s(y,\tau) - S\left(\frac{y}{\sqrt{\tau}}, 1\right), c(y,\tau) - C\left(\frac{y}{\sqrt{\tau}}, 1\right)\right) \to (0,0) \quad \text{as } \tau \to \infty. \tag{5.18}$$

The rigorous proof of this result is beyond the aims of this paper and will be considered in a future work. In the following we present a consistent numerical verification of this asymptotic behavior.

5.2. A numerical study. Let us analyze numerically the asymptotic behavior of the solution. The following questions are under consideration:

- Does the approximate solution have the correct asymptotic limit?
- How does the front appear?
- What is the convergence rate to the limit?

We consider the unscaled model (2.15)–(2.16) with the previous parameters (4.1) and $A = 100$. Let us denote $g(\xi) = \int_0^\xi \mathrm{e}^{-x^2/4} dx$. The above results read

$$\lim_{t\to+\infty} s(x,t) - \sigma\left(\frac{x}{\sqrt{t}}\right) = 0, \quad \lim_{t\to+\infty} c(x,t) - \gamma\left(\frac{x}{\sqrt{t}}\right) = 0$$

with

$$\begin{cases} \sigma(\xi) = \dfrac{\rho_{s0}}{\varphi(0)}\left[1 - \dfrac{g(\xi)}{g(\xi_0)}\right] & \text{if } \xi < \xi_0, \quad 0 \text{ otherwise,} \\ \gamma(\xi) = c_0 & \text{if } \xi > \xi_0, \quad 0 \text{ otherwise.} \end{cases}$$

Here ξ_0 is the unique solution of

$$\xi \mathrm{e}^{\xi^2/4} g(\xi) = \frac{2\rho_{s0} m_c}{c_0 m_s}.$$

By strict convexity, Newton's method converges to solve this equation, and we find $\xi_0 = 0.545$ approximatively.

As mentioned in section 4, the fully implicit finite element method ($\theta = 1$) is the cheaper in terms of computation times, so we choose this method here. We compute the solution on a rather large space interval—$[0, 10]$—to avoid the influence of the right boundary condition. As the front position is $x(t) = \xi_0\sqrt{t}$, this allows us to compute the solution for $t < 300$. We take $\Delta x = 0.01$ and $\frac{\Delta t}{\Delta x} = 0.5$. To make sure that this choice is correct, we first compared the solution at $t = 10$ with the one obtained by explicit finite differences with $\frac{\Delta t}{\Delta x^2} = C$. Both results coincide; see Figure 7.

Figure 8 represents early times and the formation of the front. We have $c(0, t) = c_0 \mathrm{e}^{-At/m_s}$. This gives $c(0, 5) = 0.004$, and $10^{-7} < c(0, 10) < 10^{-6}$. In fact, after the time $t = 5$, we already observe the asymptotic profile on the calcite density.

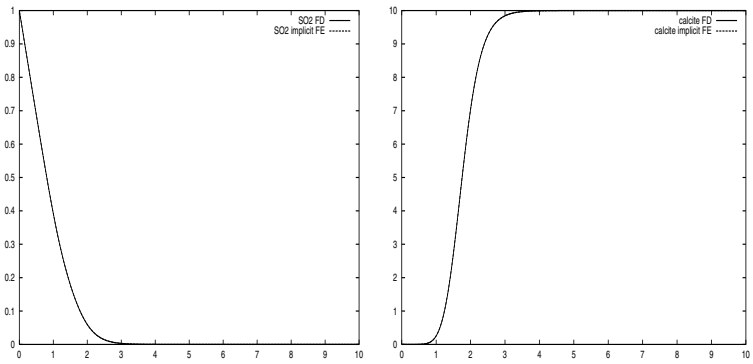


FIG. 7. *Comparison among implicit finite element methods and explicit finite difference methods. Left: SO_2 concentration; right: calcite density. Time $t = 10$.*

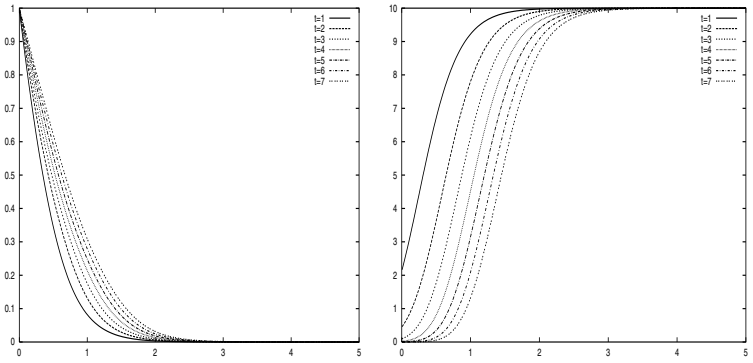


FIG. 8. *Front formation. Left: SO_2 concentration; right: calcite density.*

If we now compute the solution for large times and compare it to the theoretical limit, we obtain qualitatively a good agreement with our prediction, as shown in Figure 9, where we represent the solution with respect to an $x/\sqrt{t}$-scale, in order to observe the convergence to the asymptotic state. Let us study the convergence more carefully. As suggested by theoretical results [15], which were obtained for the case with constant porosity and constant diffusion, we expect that s converges on $\mathbb{R}_x^+$ while c converges out of the front. Let us denote

$$e_s(x,t) = \left| s(x,t) - \sigma\left(\frac{x}{\sqrt{t}}\right)\right|, \quad e_c(x,t) = \left| c(x,t) - \gamma\left(\frac{x}{\sqrt{t}}\right)\right|$$

and for $\delta \geq 0$, $t \geq 0$,

$$X(\delta,t) = \left\{ x \in \mathbb{R}^+, \ \left| \frac{x}{\sqrt{t}} - \xi_0 \right| \geq \delta \right\}.$$

For $p \in [1, +\infty]$, we study

$$E_s(A,p,\delta,t) = \|e_s(.,t)\|_{L^p(X(\delta,t))}, \quad E_c(A,p,\delta,t) = \|e_c(.,t)\|_{L^p(X(\delta,t))}.$$

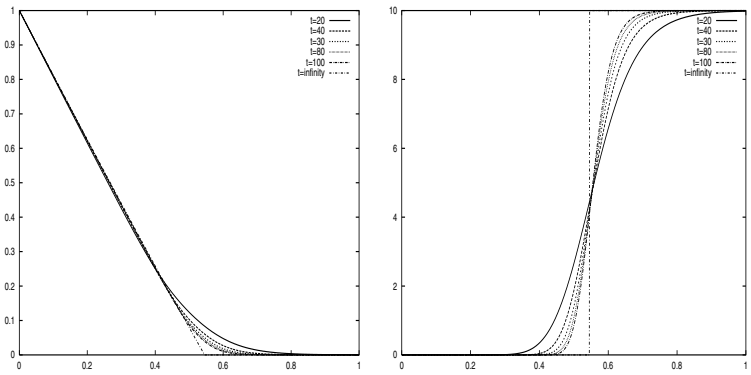


FIG. 9. *Large time behavior. Left: SO_2 concentration; right: calcite density, with respect to the $x/\sqrt{t}$-scale.*

If (s, c) is a solution of system (2.15)–(2.16) with $A = 1$, then the scaled function (s^k, c^k) defined in (5.4) is a solution of the same equations with $A = k^2$. Therefore, we have the following result.

PROPOSITION 5.1. *For any $p \in [1, +\infty]$, $\delta \geq 0$, $t \geq 0$,*

(5.19)
$$E_s(A, p, \delta, t) = A^{-1/2p} E_s(1, p, \delta, At), \quad E_c(A, p, \delta, t) = A^{-1/2p} E_c(1, p, \delta, At).$$

Suppose now that

$$E_s(A, p, \delta, t) = C_s(A, p, \delta) t^{-r_s}, \quad E_c(A, p, \delta, t) = C_c(A, p, \delta) t^{-r_c}.$$

Then

$$E_s(1, p, \delta, t) = A^{r_s + 1/2p} C_s(A, p, \delta) t^{-r_s},$$
$$E_c(1, p, \delta, t) = A^{r_c + 1/2p} C_c(A, p, \delta) t^{-r_c}.$$

Consequently, the convergence rates r_s, r_c do not depend on A. The other parameters are fixed by the physical properties of the calcite specimen and the SO_2. From these considerations, we conclude that these convergence rates may be considered as specific to the problem. We have determined them experimentally for $p = +\infty$, $p = 1$, $p = 2$.

As far as we are concerned with uniform convergence, the results are as expected—the SO_2 concentration converges uniformly on $\mathbb{R}^+$ and the calcite density does not:

$$\lim_{t \to +\infty} E_s(A, \infty, 0, t) = 0, \quad \lim_{t \to +\infty} E_c(A, \infty, 0, t) \neq 0.$$

We have then computed the maximum difference between the asymptotic limit and the computed solution, out of the front, for different values of δ in the range $[0.01\xi_0, 0.2\xi_0]$. For $\delta \geq 0.05\xi_0$ we observe that

$$\lim_{t \to +\infty} E_c(A, \infty, \delta, t) = 0.$$

These results are shown in Figure 10. In the legend we denoted $d = \delta/\xi_0$. The convergence rates are shown in Figure 11. As we evaluate them at each time step,

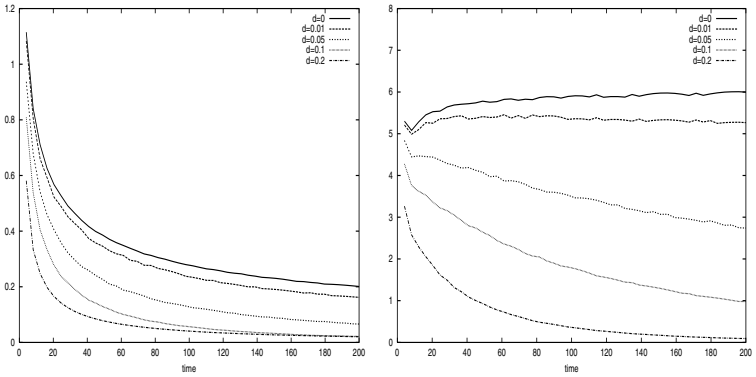


FIG. 10. *Maximal distance to asymptotic solution, out of the front, with respect to time. Left: SO_2 concentration; right: calcite density.*

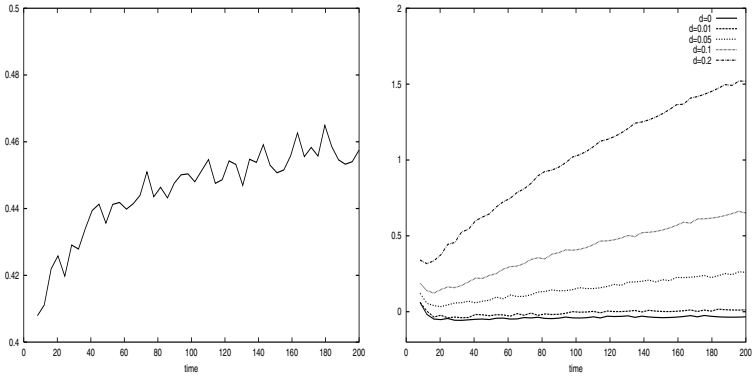


FIG. 11. *Convergence rate in maximum norm, with respect to time. Left: SO_2 concentration; right: calcite density.*

we obtain a function of time that is asymptotic to a constant value for large times. For SO_2, we show only the result for $\delta = 0$: we find that r_s is about 0.45. For the second variable, the convergence rate r_c is also about 0.4 for $\delta = 0.1\xi_0$. For the values considered here, it is an increasing function of δ. Observe that the rates begin to stabilize after the time $t = 5$, that is, the time for which we see the formation of the front.

The analysis of L^1 and L^2 convergences leads to similar results. In Figures 12 and 13, we represented the L^1 and L^2 distances of the computed solution to the theoretical limit, out of the front. As we already have uniform convergence of s on $\mathbb{R}^+$, it is clear that s converges also in L^p norm, and we actually observe it. With regard to the calcite density, we do not find out L^1 or L^2 convergence, where uniform convergence does not hold, although the curves for $\delta = 0.01\xi_0$ are slightly decreasing. Finally, Figures 14 and 15 give the convergence rates, and they are similar to the ones for the L^∞ norm. We also remark that these numerical results are in sharp contrast to the analytical results of [16, 4], where global strong convergence, i.e., up the front, of the reactive unknown c was proven for the case $\alpha = 0$, namely, in the constant porosity case. Actually, this convergence is observed also at the numerical level by using our

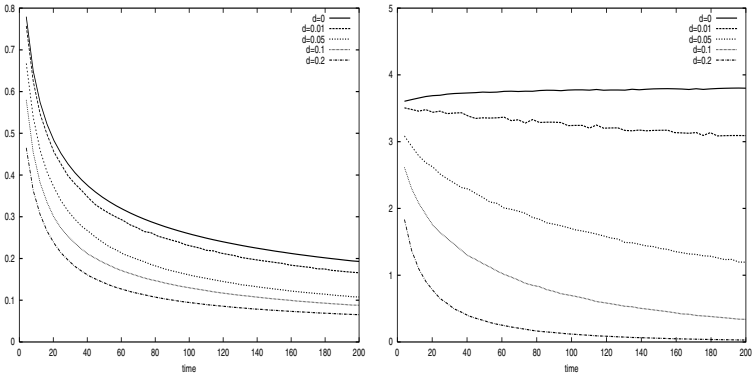


FIG. 12. *L^1 distance to asymptotic solution, out of the front, with respect to time. Left: SO_2 concentration; right: calcite density.*

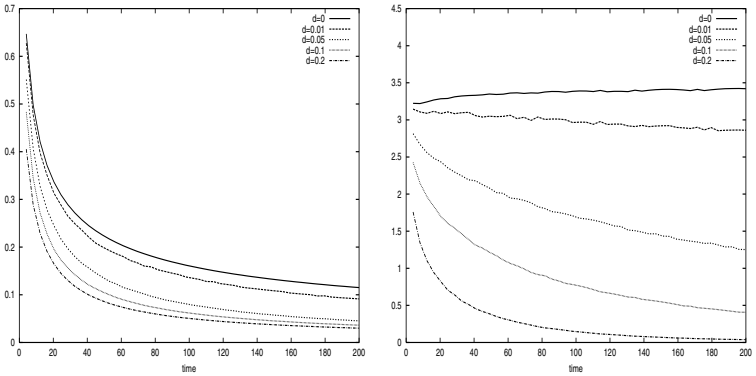


FIG. 13. *L^2 distance to asymptotic solution, out of the front, with respect to time. Left: SO_2 concentration; right: calcite density.*

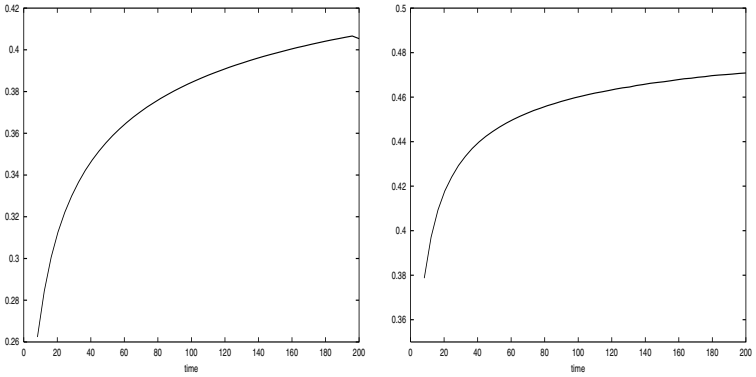


FIG. 14. *L^p convergence rate for SO_2 with respect to time. Left: $p = 1$; right: $p = 2$.*

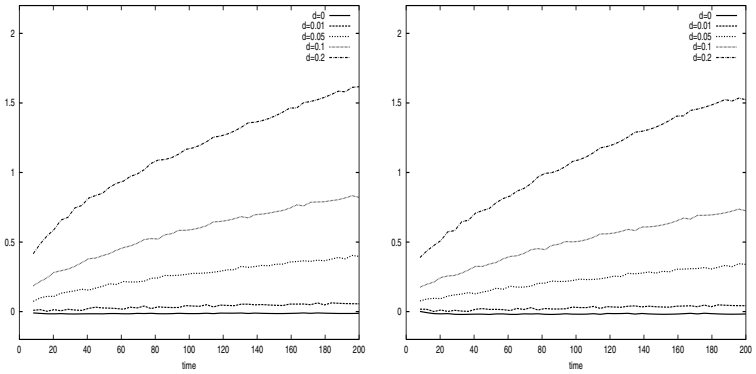


FIG. 15. *L^p convergence rate out of the front for calcite with respect to time. Left: $p = 1$; right: $p = 2$.*

schemes. The main difference between the constant and the nonconstant porosity case is in the singular nonlinear term $\alpha c_x s_x$, which appears in the nonconstant case by developing all the derivatives in (1.1). Actually, this is the main difficulty toward obtaining a rigorous proof of (5.18).

5.3. Comparison with experimental results. The asymptotic profile (5.16) of our solutions shows that there exists a clear front of gypsum, which evolves as a linear function of $\sqrt{t}$. Let us notice here that this behavior has been experimentally observed in many independent tests; see, for instance, [22, 23, 7, 8, 17]. In connection with the present research, some new laboratory tests were performed in [10], and great care was given to force the monoaxial symmetry of the experiment and to establish a clear dependence of the speed of the front on the physical parameters.

Figure 16 presents on the left the numerical simulation of the evolution of the front, plotted as a function of $\sqrt{t}$, while the right presents the gypsum thickness values at different times, in the same scale, as given by the preliminary experimental results given in [10].

It is possible to observe a good qualitative agreement between the numerical

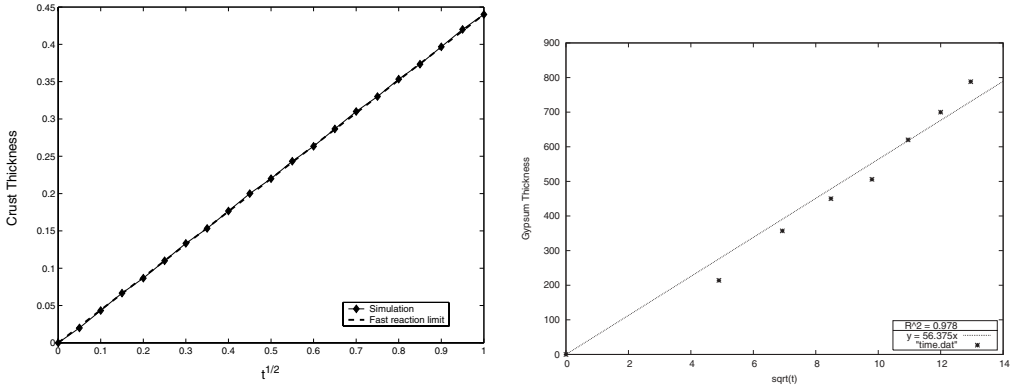


FIG. 16. *Left: Position of the calcite front as a function of $\sqrt{t}$. The simulated results were obtained using 300 finite elements, $\Delta t = 1/3000$, and $k = 10^5$. Right: Crust thickness measured in function of time reaction as obtained in the laboratory test in* [10].

prediction and the experimental data, which gives the possibility of a future calibration of the parameters on a larger set of data. In this way it will be possible to quantify the real damage phenomena on the stone materials.

6. Conclusion. We have considered a macroscopic hydrodynamic model for the evolution of the gypsum fronts in calcium carbonate stones, for which we have designed several finite element and finite difference discretizations. The finite element method involves the resolution of one tridiagonal linear system per time step. All the finite difference schemes have an explicit formulation and allow high order in time, but we have noticed that in our context, first order seems to be sufficient.

Our numerical experiments show that the approximations FD3 and FD4 are less efficient than the others. An interesting feature of the methods FE, FD1, and FD2 with $\theta = 1/2$ is that for short times they are numerically second order accurate. We have also observed that the implicit finite element method FE-1 can be used with a hyperbolic like time step $\Delta t = C.\Delta x$, which allows fast computations.

Numerical stability is established, and all those schemes give comparable shapes. We have produced asymptotic solutions for the model, and they are retrieved by the computational results. Moreover, this behavior is in qualitative agreement with the experimental tests. This is an important step in the validation of the model. In view of these encouraging facts, multidimensional computations on realistic complex geometries are now under consideration.

Acknowledgments. The authors would like to thank G. I. Barenblatt for valuable discussions about this work. We thank Rein van der Hout, Gianni Royer, Carlo Nitsch, and Maria Laura Santarelli, who read the first version of this manuscript and made many interesting remarks. We also thank Micaela Incitti and Vidar Furuholt for their active collaboration.

REFERENCES

[1] G. Alì, V. Furuholt, R. Natalini, and I. Torcicollo, *Numerical and Qualitative Analysis of a Mathematical Model of Sulphite Chemical Aggression of Limestones with High Permeability*, IAC report, Instituto per le Applicazioni del Calculo "Mauro Picone," Rome, Italy, 2004. Available online at http://www.iac.rm.cnr.it/∼natalini/ps/afnt.pdf.

[2] G. G. Amoroso and V. Fassina, *Stone Decay and Conservation—Atmospheric Pollution, Cleaning, Consolidation and Protection*, Elsevier, Amsterdam, 1983.

[3] G. I. Barenblatt, V. M. Entonov, and V. M. Ryzhik, *Theory of Fluid Flows through Natural Rocks*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[4] M. Belhadj, J.-F. Gerbeau, and B. Perthame, *A multiscale colloid transport model with anisotropic degenerate diffusion*, Asymptot. Anal., 34 (2003), pp. 41–54.

[5] R. Bugini, M. Laurenzi Tabasso, and M. Realini, *Rate of formation of black crusts on marble. A case study*, J. Cultural Heritage, 1 (2000), pp. 111–116.

[6] F. Garbassi, E. Mello, and M. Laurenzi Tabasso, *In situ XPS observation of the first stages of marble sulphation by atmospheric SO_2*, Durability Build. Mater., 3 (1985), pp. 51–58.

[7] K. L. Gauri and J. A. Gwinn, *Deterioration of marble in air containing* 5–10 *ppm SO_2 and NO_2*, Durability Build. Mater., 1 (1982/83), pp. 217–223.

[8] K. L. Gauri, N. P. Kulshreshtha, A. R. Punuru, and A. N. Chowdhury, *Rate of decay of marble in laboratory and outdoor exposure*, J. Mater. Civil Engrg., 1 (1989), pp. 73–85.

[9] K. L. Gauri, R. Popli, and A. C. Sarma, *Effect of relative humidity and grain size on the reaction rates of marble at high concentrations of SO_2*, Durability Build. Mater., 1 (1982/83), pp. 209–216.

[10] C. Giavarini, M. Incitti, M. L. Santarelli, R. Natalini, and V. Furuholt, *A Nonlinear Model of Sulphation of Calcium Carbonate Stones: Numerical Simulations and Preliminary Laboratory Assessments*, IAC report 19, Instituto per le Applicazioni del

Calculo "Mauro Picone," Rome, Italy, 2003. Available online at http://www.iac.rm.cnr.it/~natalini/ps/CIpre.pdf.

[11] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[12] F. R. Guarguaglini and R. Natalini, *Global Existence of Smooth Solutions to a Nonlinear Model of Sulphation Phenomena in Calcium Carbonate Stones*, IAC report 9, Instituto per le Applicazioni del Calculo "Mauro Picone," Rome, Italy, 2003. Available online at http://www.iac.rm.cnr.it/~natalini/postscript/solf4.pdf.

[13] S. M. Hassanizadeha and A. Leijnsea, *A non-linear theory of high-concentration-gradient dispersion in porous media*, Adv. Water Res., 18 (1995), pp. 203–215.

[14] F. H. Haynie, *Deterioration of marble*, Durability Build. Mater., 1 (1982/83), pp. 241–254.

[15] D. Hilhorst, R. van der Hout, and L. A. Peletier, *The fast reaction limit for a reaction-diffusion system*, J. Math. Anal. Appl., 199 (1996), pp. 349–373.

[16] D. Hilhorst, R. van der Hout, and L. A. Peletier, *Nonlinear diffusion in the presence of fast reaction*, Nonlinear Anal., 41 (2000), pp. 803–823.

[17] B. G. D. Hoke and D. L. Turcotte, *Weathering and damage*, J. Geophys. Res., 107 (2002), 2210.

[18] N. P. Kulshreshtha, A. R. Punuru, and K. L. Gauri, *Kinetics of reaction of SO_2 with marble*, J. Mater. Civil Engrg., 1 (1989), pp. 60–72.

[19] W. T. Lipfert, *Atmospheric damage to calcareous stones: Comparison and reconciliation of recent experimental findings*, Atmos. Environ., 23 (1989), pp. 415–429.

[20] A. M. Meirmananov, *The Stefan Problem*, de Gruyter, Berlin, 1992.

[21] D. A. Nield and A. Bejan, *Convection in Porous Media*, Springer, Berlin, 1992.

[22] Th. Skoulikidis and E. Papakonstantinou-Ziotis, *Mechanism of sulphation by atmospheric SO_2 of the limestones and marbles of the ancient monuments and statues*, I. *Observations in situ (Acropolis) and laboratory measurements*, British Corrosion J., 16 (1981), pp. 63–69.

[23] Th. Skoulikidis and D. Charalambous, *Mechanism of sulphation by atmospheric SO_2 of the limestones and marbles of the ancient monuments and statues*, II. *Hypothesis concerning the rate determining step in the process of sulphation, and its experimental confirmation*, British Corrosion J., 16 (1981), pp. 70–76.

[24] I. Stakgold, *Gas-solid reaction with porosity change*, in Proceedings of the Conference on Nonlinear Differential Equations, Electron. J. Differ. Equ. Conf. 5, Southwest Texas State University, San Marcos, TX, 2000, pp. 247–252.

[25] J. Szekely, J. W. Evans, and H. Y. Sohn, *Gas-Solid Reaction*, Academic Press, New York, 1976.

[26] S. Tambe, K. L. Gauri, S. Li, and W. G. Cobourn, *Kinetic study of SO_2 reaction with dolomite*, Environ. Sci. Technol., 25 (1991), pp. 2071–2075.

[27] R. S. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1668–1690

EUCLIDEAN SHIFT-TWIST SYMMETRY IN POPULATION MODELS OF SELF-ALIGNING OBJECTS*

PAUL C. BRESSLOFF†

Abstract. We consider the symmetry properties of a general class of nonlocal population models describing the aggregation and alignment of oriented objects in two dimensions. Such objects could be at the level of molecules, cells, or whole organisms. We show that the underlying interaction kernel is invariant under the so-called shift-twist action of the Euclidean group acting on the space $\mathbf{R}^2 \times S^1$. This group action was previously studied within the context of a continuum model of primary visual cortex. We use perturbation methods to solve the eigenvalue problem arising from linearization about a homogeneous state, and then use equivariant bifurcation theory to identify the various types of doubly periodic patterns that are expected to arise when the homogeneous state becomes unstable. We thus establish that two distinct forms of spatio-angular order can occur, corresponding to scalar and pseudoscalar representations of the Euclidean group.

Key words. cell alignment, actin cytoskeleton, animal aggregation, Euclidean symmetry, self-organization, population models, integro-differential equations

AMS subject classifications. 92B05, 37G40

DOI. 10.1137/S0036139903436017

1. Introduction.

A wide variety of self-organizing biological systems exhibit aggregation and alignment phenomena. These occur spontaneously due to mutual interactions between the individual elements of a population, in which both the relative position and the relative orientation of the individuals have a significant effect on the nature of the interactions. The underlying population can consist of molecules, cells, or whole organisms. A well-known example of the last category is the aggregation of animal herds, fish schools, and flocks of birds, in which the members of the group tend to align their bodies with each other and move in a common direction [14]. Such behavior provides a defense against predators. Examples at the cellular and molecular levels are the alignment of mammalian fibroblast cells within densely formed patches [10] and the alignment of actin filaments forming a scaffolding structure within a cell [22]. In order to investigate the important role of alignment in population survival and in the properties of biological materials, a number of continuum models of interacting oriented objects have been developed, with applications to animal social groups [19, 14, 6], fibroblasts [16, 17, 8], and actin [7, 12, 21]. All of these models are formulated in terms of integro-differential equations describing the evolution of the distribution of oriented elements in space. It is typically assumed that the interaction terms involve convolutions of the population distribution with some linear kernel. Convolutions with respect to orientation are natural, since an individual can interact with a neighboring cell having any relative orientation, whereas convolutions with respect to spatial position can be justified by assuming that signaling between individuals happens on a much faster time-scale than aggregation and alignment [15].

In this paper we investigate how symmetries of the interaction kernel determine the types of spatio-angular patterns that can emerge through a Turing-like instability of a homogeneous state. For concreteness, we focus on a diffusion-advection equation

*Received by the editors October 6, 2003; accepted for publication (in revised form) January 5, 2004; published electronically June 22, 2004.
http://www.siam.org/journals/siap/64-5/43601.html
†Department of Mathematics, University of Utah, 155 South 1400 East, Salt Lake City, UT 84112 (bressloff@math.utah.edu).

for a population of oriented objects distributed in the two-dimensional plane. This corresponds to the second of three classes of model previously studied by Mogilner and Edelstein-Keshet [18], in which nonlocal interactions induce a rotation about the center of mass of each element as well as a linear drift of the center of mass in the plane. (The other two models are less realistic in the sense that they treat the alignment process as instantaneous, although they do exhibit the same qualitative behavior since they have the same underlying symmetries.) Mogilner and Edelstein-Keshet [18] solved the eigenvalue problem arising from a linear stability analysis of the homogeneous state in the case of a separable interaction kernel, that is, one in which variations in the strength of interaction with respect to relative orientation and relative position are uncorrelated. Such a kernel is invariant under the action of the product group $\mathbf{E}(2) \times \mathbf{O}(2)$, where $\mathbf{E}(2)$ denotes the Euclidean group of rigid body motions in the plane $\mathbf{R}^2$ and $\mathbf{O}(2)$ consists of rotations and reflection on the circle S^1. However, it is often found that individuals with similar orientations have stronger interactions when they are collinear in the plane, implying that the interaction kernel is nonseparable [6]. (A specific example of a nonseparable kernel was briefly considered in [18], but general symmetry properties were not addressed.) Here we show that these more realistic kernels are invariant with respect to the so-called *shift-twist* action of $\mathbf{E}(2)$ acting on the space $\mathbf{R}^2 \times S^1$, and we explore the consequences of this symmetry for pattern formation. Note that the same group action has recently been analyzed within the context of continuum models of the visual cortex, where the nonlocal interactions are mediated by axonal connections between neurons that are tuned to respond to oriented visual stimuli [3, 4]. Shift-twist invariant kernels also play a central role in a recently proposed computational algorithm for grouping and joining edges that form the boundaries of objects in a visual image [24].

We begin by describing the nonlocal population model for the aggregation and alignment of oriented objects in two dimensions and discussing its symmetry properties (section 2). We show that the resulting diffusion-advection equation is equivariant with respect to the shift-twist action of the Euclidean group due to the invariance of the underlying interaction kernel. We then use perturbation methods to solve the eigenvalue problem arising from linearization about a homogeneous state and determine marginal stability conditions (section 3). Finally, we use equivariant bifurcation theory to identify the various types of doubly periodic patterns that are expected to arise when the homogeneous state becomes unstable, and thus establish that two distinct forms of spatio-angular order can occur, corresponding to the so-called scalar and pseudoscalar representations of the Euclidean group (section 4). Interestingly, analogous results [5] have been obtained for the Landau–de Gennes model of a nematic liquid crystal [9], where the oriented objects are rod-like molecules that interact by electrostatic attraction or repulsion.

2. Description of the model and its symmetries. Let $f(\mathbf{r}, \theta, t)$ denote the distribution of oriented objects in a two-dimensional domain $\mathcal{D} \subset \mathbf{R}^2$ at time t with $\mathbf{r} \in \mathcal{D}$ and $-\pi < \theta \leq \pi$. It is assumed that the total number of objects N is conserved; that is, $\dot{N} = 0$ with

$$N = \int_{\mathcal{D}} \int_{-\pi}^{\pi} f(\mathbf{r}, \theta, t) \frac{d\theta}{2\pi} d^2\mathbf{r}. \tag{2.1}$$

In cases where the population grows (due to cell proliferation, for example), N may still be treated as a constant, provided that the growth process is adiabatic. The pop-

ulation distribution f is taken to evolve according to the diffusion-advection equation

$$\frac{\partial f}{\partial t} = D_1 \frac{\partial^2 f}{\partial \theta^2} + D_2 \nabla^2 f - \frac{\partial J_\theta}{\partial \theta} - \nabla \cdot \mathbf{J}_r. \tag{2.2}$$

Here D_1 and D_2 are angular and spatial diffusion constants, J_θ is the flux arising from changes in the orientation of objects, and $\mathbf{J}_r$ is the flux arising from motion in the plane:

$$J_\theta = f \frac{d\theta}{dt}, \quad \mathbf{J} = f \frac{d\mathbf{r}}{dt}. \tag{2.3}$$

Following Mogilner and Edelstein-Keshet [16, 18], we assume that inertial forces can be neglected so that the velocities are simply proportional to the driving forces,

$$\frac{d\theta}{dt} = \eta_1 F_\theta, \quad \frac{d\mathbf{r}}{dt} = \eta_2 \mathbf{F}_r. \tag{2.4}$$

The forces are taken to be conservative; that is, each can be expressed in terms of the gradient of an underlying potential function V:

$$F_\theta = \frac{\partial V}{\partial \theta}, \quad \mathbf{F}_r = \nabla V, \tag{2.5}$$

where V is given by the integral of f with respect to a linear kernel W,

$$V(\mathbf{r}, \theta) = W * f(\mathbf{r}, \theta) \equiv \int_{\mathcal{D}} \int_{-\pi}^{\pi} W(\mathbf{r}, \theta | \mathbf{r}', \theta') f(\mathbf{r}', \theta') \frac{d\theta}{2\pi} d^2\mathbf{r}. \tag{2.6}$$

(One could also consider a more general situation in which angular and planar motion are generated by two distinct potentials [18].) Substituting (2.3), (2.4), (2.5), and (2.6) into (2.2) leads to the following model equation:

$$\frac{\partial f}{\partial t} = D_1 \frac{\partial^2 f}{\partial \theta^2} + D_2 \nabla^2 f - \eta_1 \frac{\partial}{\partial \theta} \left(f \frac{\partial (W * f)}{\partial \theta} \right) - \eta_2 \nabla \cdot (f \nabla (W * f)). \tag{2.7}$$

2.1. The interaction kernel. We now specify the form of the interaction kernel W. Consider a local patch of individuals at point $\mathbf{r}$ with orientation θ. These will move in the plane and reorient as a result of the influence from other patches at $\mathbf{r}'$ with orientation θ' (see Figure 2.1(a)). It is likely that the interactions depend on three distinct factors [6, 18]: (i) the Euclidean distance $|\mathbf{r} - \mathbf{r}'|$, (ii) the relative orientation $\theta - \theta'$, and (iii) the relative alignment in the plane $\psi = \arg(\mathbf{r} - \mathbf{r}') - \theta$. One way to understand the third factor is to consider the case of parallel objects that are equidistant in the plane but are either collinear or flanking each other (see Figure 2.1(b)). Given the fact that each object is elongated, it is possible that collinear objects tend to influence each other more strongly than flanking objects. In the cellular or molecular case this would arise due to differences in the contact areas of the individuals, whereas in animal social groups this would reflect differences in the ability to sense individuals in different relative directions. In the latter case, the influence of individuals behind a given animal would also tend to be weaker than on the sides or front. If we assume that the three effects are independent, then the interaction kernel can be decomposed into the product form (see [6])

$$W(\mathbf{r}, \theta | \mathbf{r}', \theta') = G(|\mathbf{r} - \mathbf{r}'|) H(\theta - \theta') \Delta(\arg(\mathbf{r} - \mathbf{r}') - \theta). \tag{2.8}$$

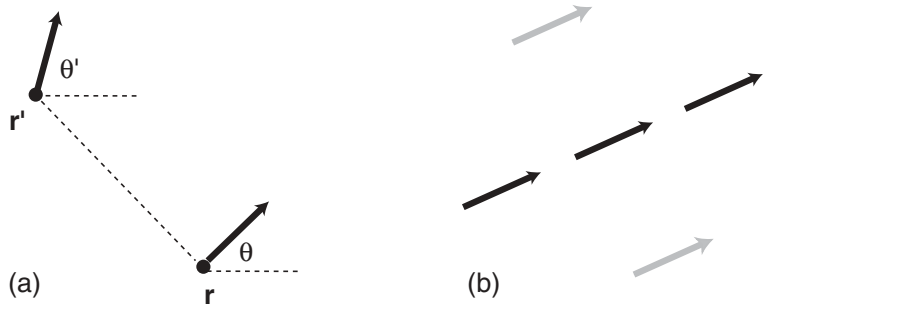


FIG. 2.1. (a) *Two oriented objects in the plane.* (b) *Difference in the influence of collinear (black) and flanking (gray) oriented objects.*

We now discuss each of the terms appearing in (2.8).

The strength of interactions decreases as a function of spatial separation so that $G(r)$ is a monotonically decreasing function of r. For concreteness, we take G to be a Gaussian

$$G(r) = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-r^2/2\sigma^2}, \tag{2.9}$$

where σ denotes the effective range of interactions. We fix the spatial scale by setting $\sigma = 1$. Throughout this paper we assume that the range of interactions is at least a few orders of magnitude smaller than the size of the domain $\mathcal{D}$. This allows us to treat the spatial domain as infinite so that we can ignore boundary effects. The angular contribution $H(\theta)$ is assumed to be an even function of θ with one or more maxima whose locations are model-dependent. In the case of fibroblasts [18], interactions tend to favor parallel alignment which may be "head-to-head" ($\theta = 0$) or "head-to-tail" ($\theta = \pi$). This can be modeled by taking a bimodal function of the form

$$H(\theta) = [\cos 2\theta - \cos 2\theta_0]_+, \quad \theta_0 < \frac{\pi}{4}, \tag{2.10}$$

where $[x]_+ = x$ if $x > 0$ and $[x]_+ = 0$ if $x \le 0$, and θ_0 determines the width of the two maxima. If only "head-to-head" alignment is favored, then $H(\theta)$ can be modeled by the unimodal function

$$H(\theta) = [\cos\theta - \cos\theta_0]_+, \quad \theta_0 < \frac{\pi}{2}. \tag{2.11}$$

In the case of actin fibers [7], crosslinking proteins allow fibers to interact and bind at different configurations that include both parallel and orthogonal alignment. An example of an orthogonal interaction kernel is

$$H(\theta) = [\cos 2\theta_0 - \cos 2\theta]_+, \quad \theta_0 > \frac{\pi}{4}. \tag{2.12}$$

All three cases are illustrated in Figure 2.2. The final factor Δ is expected to be a positive function that is greater for coaligned elements than for flanking elements (at least in the case of parallel alignment). One possibility is to take

$$\Delta(\psi) = 1 + \beta\cos 2\psi, \quad \beta < 1. \tag{2.13}$$

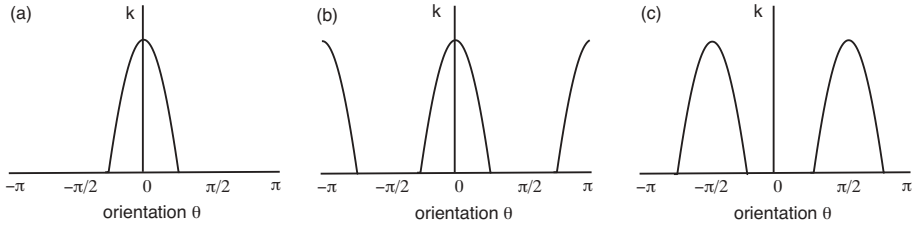


FIG. 2.2. *Orientation kernel $H(\theta)$ for different favored alignments:* (a) *unimodal parallel alignment,* (b) *bimodal parallel alignment,* (c) *orthogonal alignment.*

Note that as it stands the interaction kernel (2.8) is not symmetric under the interchange $(\mathbf{r},\theta) \leftrightarrow (\mathbf{r}',\theta')$. It is straightforward to modify the kernel so that it has such an exchange symmetry by taking

$$W(\mathbf{r},\theta|\mathbf{r}',\theta') = \frac{1}{2}H(\theta-\theta')\left[J(\mathbf{r}-\mathbf{r}',\theta)+J(\mathbf{r}-\mathbf{r}',\theta')\right], \tag{2.14}$$

where

$$J(\mathbf{r},\theta) = G(|\mathbf{r}|)\Delta(\arg(\mathbf{r})-\theta). \tag{2.15}$$

2.2. Euclidean symmetry. We now show that the nonseparable interaction kernel W given by (2.8) is invariant under the action of the Euclidean group $\mathbf{E}(2)$, which is composed of the (semidirect) product of $\mathbf{O}(2)$, the group of planar rotations and reflections, with $\mathbf{R}^2$, the group of planar translations. The action of the Euclidean group on $\mathbf{R}^2 \times \mathbf{S}^1$ is generated by

$$\begin{aligned} \mathbf{s}\cdot(\mathbf{r},\theta) &= (\mathbf{r}+\mathbf{s},\theta), \qquad \mathbf{s}\in\mathbf{R}^2, \\ \varphi\cdot(\mathbf{r},\theta) &= (R_\varphi\mathbf{r},\theta+\varphi), \qquad \varphi\in\mathbf{S}^1, \\ \kappa\cdot(\mathbf{r},\theta) &= (\kappa\mathbf{r},-\theta), \end{aligned} \tag{2.16}$$

where κ is the reflection $(x_1,x_2)\mapsto(x_1,-x_2)$ and R_φ is a rotation by φ. The corresponding group action on a function $a:\mathbf{R}^2\times\mathbf{S}^1\to\mathbf{R}$ is given by

$$\gamma\cdot a(P) = a(\gamma^{-1}\cdot P) \quad \text{for all } \gamma\in\mathbf{O}(2)\dot{+}\mathbf{R}^2, \tag{2.17}$$

where $P=(\mathbf{r},\theta)$, and the action on $W(P|P')$ is

$$\gamma\cdot W(P|P') = W(\gamma^{-1}\cdot P|\gamma^{-1}\cdot P').$$

The so-called shift-twist action in (2.16) reflects a crucial feature of the underlying interactions, namely that they tend to favor collinear parallel elements. This correlation between relative angular position and object orientation means that invariance of W under $\mathbf{E}(2)$ requires a rotation in the plane according to the *twist* $\mathbf{r}\to R_\varphi\mathbf{r}$ and a simultaneous rotation of object orientation according to the *shift* $\theta\to\theta+\varphi$ (see Figure 2.3). A similar argument holds for reflections.

Translation invariance of W given by (2.8) follows immediately from the spatial homogeneity of the interactions, which implies that

$$W(\mathbf{r}-\mathbf{s},\theta|\mathbf{r}'-\mathbf{s},\theta') = W(\mathbf{r},\theta|\mathbf{r}',\theta').$$

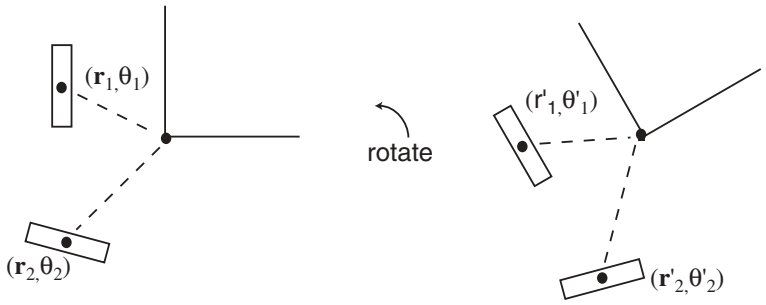


FIG. 2.3. *Action of a rotation by φ on two oriented objects located at planar positions $\mathbf{r}_1, \mathbf{r}_2$ and with internal orientations θ_1, θ_2. The action is of the form $(\mathbf{r}, \theta) \to (\mathbf{r}', \theta') = (R_\varphi \mathbf{r}, \varphi + \theta)$.*

Invariance with respect to a rotation by φ follows from

$$\begin{aligned} &W(R_{-\varphi}\mathbf{r}, \theta - \varphi | R_{-\varphi}\mathbf{r}', \theta' - \varphi) \\ &\quad = G(|R_{-\varphi}(\mathbf{r} - \mathbf{r}')|) H(\theta - \varphi - \theta' + \varphi) \Delta(\arg[R_{-\varphi}(\mathbf{r} - \mathbf{r}')] - \theta + \varphi) \\ &\quad = G(|\mathbf{r} - \mathbf{r}'|) H(\theta - \theta') \Delta(\arg(\mathbf{r} - \mathbf{r}') - \theta) \\ &\quad = W(\mathbf{r}, \theta | \mathbf{r}', \theta'). \end{aligned}$$

We have used the conditions $|R_\varphi \mathbf{r}| = |\mathbf{r}|$ and $\arg(R_{-\varphi}\mathbf{r}) = \arg(\mathbf{r}) - \varphi$. Finally, invariance under a reflection κ about the x-axis holds since

$$\begin{aligned} W(\kappa\mathbf{r}, -\theta | \kappa\mathbf{r}', -\theta') &= G(|\kappa(\mathbf{r} - \mathbf{r}')|) H(-\theta + \theta') \Delta(\arg[\kappa(\mathbf{r} - \mathbf{r}')] + \theta) \\ &= G(|\mathbf{r} - \mathbf{r}'|) H(\theta - \theta') \Delta(-\arg(\mathbf{r} - \mathbf{r}') + \theta) \\ &= W(\mathbf{r}, \theta | \mathbf{r}', \theta'). \end{aligned}$$

We have used the conditions $\arg(\kappa\mathbf{r}) = -\arg(\mathbf{r})$, $H(-\theta) = H(\theta)$, and $\Delta(-\psi) = \Delta(\psi)$. Finally, using identical arguments, it is straightforward to show that the modified kernel (2.14) is also invariant under the Euclidean group action (2.16).

Let us now determine how (2.7) transforms under the shift-twist action of the Euclidean group. Introducing the transformed coordinates $(\widetilde{\mathbf{r}}, \widetilde{\theta}) = \gamma^{-1}(\mathbf{r}, \theta)$ and setting $\widetilde{f}(\mathbf{r}, \theta) = f(\widetilde{\mathbf{r}}, \widetilde{\theta})$, etc., we see that (2.7) becomes

$$\frac{\partial \widetilde{f}}{\partial t} = D_1 \frac{\partial^2 \widetilde{f}}{\partial \widetilde{\theta}^2} + D_2 \widetilde{\nabla}^2 \widetilde{f} - \eta_1 \frac{\partial}{\partial \widetilde{\theta}} \left(\widetilde{f} \frac{\partial \widetilde{W * f}}{\partial \widetilde{\theta}} \right) - \eta_2 \widetilde{\nabla} \cdot (\widetilde{f} \widetilde{\nabla} \widetilde{W * f}). \tag{2.18}$$

Invariance of the weight kernel W implies that $\widetilde{W * f} = W * \widetilde{f}$:

$$\begin{aligned} W * f(\gamma^{-1}P, t) &= \int_{\mathbf{R}^2 \times \mathbf{S}^1} W(\gamma^{-1}P | P') f(P', t) dP' \\ &= \int_{\mathbf{R}^2 \times \mathbf{S}^1} W(P | \gamma P') f(P', t) dP' \\ &= \int_{\mathbf{R}^2 \times \mathbf{S}^1} W(P | P'') f(\gamma^{-1}P'', t) dP'', \end{aligned}$$

since $d[\gamma^{-1}P] = \pm dP$ and W is Euclidean invariant. It is also easy to establish that

all of the quadratic differential operators are Euclidean invariant, that is,

$$\frac{\partial}{\partial\widetilde{\theta}}\left(a\frac{\partial}{\partial\widetilde{\theta}}\right)=\frac{\partial}{\partial\theta}\left(a\frac{\partial}{\partial\theta}\right),\quad \widetilde{\nabla}\cdot(a\widetilde{\nabla})=\nabla\cdot(a\nabla) \tag{2.19}$$

for any scalar function $a(\mathbf{r},\theta)$. Hence,

$$\frac{\partial\widetilde{f}}{\partial t}=D_1\frac{\partial^2\widetilde{f}}{\partial\theta^2}+D_2\nabla^2\widetilde{f}-\eta_1\frac{\partial}{\partial\theta}\left(\widetilde{f}\frac{\partial W*\widetilde{f}}{\partial\widetilde{\theta}}\right)-\eta_2\nabla\cdot(\widetilde{f}\nabla W*\widetilde{f}). \tag{2.20}$$

If we rewrite (2.7) as an operator equation, namely,

$$\mathcal{F}_t[f]\equiv\frac{df}{dt}-\mathcal{F}[f]=0, \tag{2.21}$$

then it follows that $\gamma\mathcal{F}_t[f]=\mathcal{F}_t[\gamma f]=\mathcal{F}_t[\widetilde{f}]$. Thus $\mathcal{F}_t$ commutes with $\gamma\in\mathbf{E}(2)$, and $\mathcal{F}_t$ is said to be *equivariant* with respect to the symmetry group $\mathbf{E}(2)$ (see [13]). In sections 3 and 4 we show how the equivariance of the operator $\mathcal{F}_t$ with respect to the shift-twist action of $\mathbf{E}(2)$ has major implications for the nature of solutions bifurcating from a homogeneous steady state solution. In particular, equivariance implies that there exist two distinct forms of spatio-angular order, which are associated with scalar and pseudoscalar representations of the Euclidean group. Further details concerning the general approach used in this paper, as well as many illustrative examples, can be found in the recent excellent book on the role of symmetry in nonlinear dynamical systems by Golubitsky and Stewart [13].

3. Linear stability analysis. The first step in the analysis of pattern forming instabilities is to linearize (2.7) about the homogeneous solution $f(\mathbf{r},\theta)=\overline{f}$, where

$$\overline{f}=\frac{N}{2\pi A[\mathcal{D}]},\quad A[\mathcal{D}]=\int_{\mathcal{D}}d^2\mathbf{r}, \tag{3.1}$$

and to solve the resulting eigenvalue problem. In particular, we wish to find conditions under which the homogeneous solution becomes marginally stable due to the vanishing of one of the (degenerate) eigenvalues, and to identify the marginally stable modes. In the following we will consider the modified version of the interaction kernel given by (2.14).

3.1. Eigenvalue equation. Substitute

$$f(\mathbf{r},\theta,t)=\overline{f}+a(\mathbf{r},\theta)\mathrm{e}^{\lambda t} \tag{3.2}$$

into (2.7) and expand to first order in a. This generates the linear eigenvalue equation

$$\begin{aligned}\lambda a&=\widehat{L}a\\&\equiv D_1\frac{\partial^2 a}{\partial\theta^2}+D_2\nabla^2 a-\eta_1\overline{f}\frac{\partial^2(W*a)}{\partial\theta^2}-\eta_2\overline{f}\nabla^2(W*a).\end{aligned} \tag{3.3}$$

Since the homogeneous solution has full Euclidean symmetry, $\gamma\overline{f}=\overline{f}$ for all $\gamma\in\mathbf{E}(2)$, it follows that the linear operator $\widehat{L}$ is equivariant with respect to the Euclidean group action (2.16). This can be shown either by explicitly using (3.3) or by rewriting (2.7) in the form (2.21) and exploiting the equivariance of $\mathcal{F}$. In the latter case, linearizing both sides of the equation $\gamma\mathcal{F}[f]=\mathcal{F}[\gamma f]$ about $\overline{f}$ gives

$$\gamma\left(\mathcal{F}[\overline{f}]+D\mathcal{F}[\overline{f}](f-\overline{f})\right)=\mathcal{F}[\gamma\overline{f}]+D\mathcal{F}[\gamma\overline{f}](\gamma f-\overline{f}),$$

which implies that $\gamma\widehat{L} = \widehat{L}\gamma$ for all $\gamma \in \mathbf{E}(2)$, where $\widehat{L} = D\mathcal{F}[\overline{f}]$. Equivariance of $\widehat{L}$ determines the basic form of the eigenfunction solutions of (3.3); namely, they are given by irreducible representations of the group action (2.16) on the space $\mathbf{R}^2 \times \mathbf{S}^1$. We show this following similar arguments to those of Bressloff et al. [3, 4]. First, translation symmetry implies that the eigenfunctions can be expressed in the form

$$a(\mathbf{r},\theta) = u(\theta-\varphi)e^{i\mathbf{k}\cdot\mathbf{r}} + \text{c.c.}, \tag{3.4}$$

where c.c. denotes the complex conjugate, $\mathbf{k} = q(\cos\varphi, \sin\varphi)$, and

$$\begin{aligned}\lambda u(\theta) &= D_1\frac{\partial^2 u(\theta)}{\partial\theta^2} - D_2 q^2 u(\theta)\\ &\quad - \frac{\overline{f}}{2}\left[\eta_1\frac{\partial^2}{\partial\theta^2} - \eta_2 q^2\right]\int_{-\pi}^{\pi} H(\theta-\theta')\left[\widehat{J}(\mathbf{k},\theta+\varphi) + \widehat{J}(\mathbf{k},\theta'+\varphi)\right]u(\theta')\frac{d\theta'}{2\pi}.\end{aligned} \tag{3.5}$$

Here $\widehat{J}(\mathbf{k},\theta)$ is the Fourier transform of $J(\mathbf{r},\theta)$,

$$\widehat{J}(\mathbf{k},\theta) = \int_{\mathbf{R}^2} \mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}} J(\mathbf{r},\theta)d^2\mathbf{r}. \tag{3.6}$$

Euclidean symmetry further restricts the structure of the eigensolutions $u(\theta)$ of (3.5) as follows.

(i) The Fourier transform $\widehat{J}(\mathbf{k},\theta+\varphi)$ is independent of the direction $\varphi = \arg(\mathbf{k})$. This is easy to establish as follows:

$$\begin{aligned}\widehat{J}(\mathbf{k},\theta+\varphi) &= \int_{\mathbf{R}^2} \mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}} J(\mathbf{r},\theta+\varphi)d^2\mathbf{r}\\ &= \int_0^\infty\int_{-\pi}^{\pi} \mathrm{e}^{-iqr\cos(\psi-\varphi)} G(r)\Delta(\psi-\theta-\varphi)d\psi r dr\\ &= \int_0^\infty\int_{-\pi}^{\pi} \mathrm{e}^{-iqr\cos(\psi)} G(r)\Delta(\psi-\theta)d\psi r dr\\ &= \widehat{J}(q,\theta).\end{aligned} \tag{3.7}$$

Therefore, λ and $u(\theta)$ depend only on the magnitude $q = |\mathbf{k}|$ of the wavevector $\mathbf{k}$, and there is an infinite degeneracy due to rotational invariance. Note, however, that the eigenfunction (3.4) depends on $u(\theta-\varphi)$, which reflects the shift-twist action of the rotation group.

(ii) For each $\mathbf{k}$ the associated subspace of eigenfunctions

$$V_{\mathbf{k}} = \{u(\theta-\varphi)\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}} + \text{c.c.}\} \tag{3.8}$$

decomposes into two invariant subspaces,

$$V_{\mathbf{k}} = V_{\mathbf{k}}^+ \oplus V_{\mathbf{k}}^-, \tag{3.9}$$

corresponding to even and odd functions, respectively:

$$V_{\mathbf{k}}^+ = \{v \in V_{\mathbf{k}} : u(-\theta) = u(\theta)\} \quad \text{and} \quad V_{\mathbf{k}}^- = \{v \in V_{\mathbf{k}} : u(-\theta) = -u(\theta)\}. \tag{3.10}$$

This is a consequence of reflection invariance, as we now indicate. That is, let $\kappa_{\mathbf{k}}$ denote reflections about the wavevector $\mathbf{k}$ so that $\kappa_{\mathbf{k}}\mathbf{k} = \mathbf{k}$. Then $\kappa_{\mathbf{k}}a(\mathbf{r},\phi) =$

$a(\kappa_{\mathbf{k}}\mathbf{r}, 2\varphi - \phi) = u(\varphi - \phi)\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}+$ c.c. Since $\kappa_{\mathbf{k}}$ is a reflection, any space that it acts on decomposes into two subspaces—one on which it acts as the identity I and one on which it acts as $-I$. The even and odd functions correspond to scalar and pseudoscalar representations of the Euclidean group studied in a more general context by Vivancos, Chossat, and Melbourne [1].

A further reduction of (3.5) can be achieved by expanding the 2π-periodic function $u(\theta)$ as a Fourier series with respect to θ:

$$u(\theta) = \sum_{n\in\mathbf{Z}} U_n \mathrm{e}^{in\theta}. \tag{3.11}$$

This then leads to the matrix eigenvalue equation

$$\lambda(q)U_n = [\mathbf{L}(q)\mathbf{U}]_n \equiv -A_n(q)U_n + \frac{1}{2}B_n(q)\sum_{m\in\mathbf{Z}} \widehat{J}_{n-m}(q)[H_m + H_n]U_m, \tag{3.12}$$

where

$$A_n(q) = D_1 n^2 + D_2 q^2, \quad B_n(q) = \overline{f}(\eta_1 n^2 + \eta_2 q^2) \tag{3.13}$$

and

$$\widehat{J}_n(q) = \int_{-\pi}^{\pi} \mathrm{e}^{-in\theta}\widehat{J}(q,\theta)\frac{d\theta}{2\pi} = \Delta_n \widehat{G}_n(q) \tag{3.14}$$

with

$$\widehat{G}_n(q) = \int_0^{\infty}\int_{-\pi}^{\pi} \mathrm{e}^{-iqr\cos(\psi)}\mathrm{e}^{-in\psi}G(r)d\psi r dr. \tag{3.15}$$

We have used (3.7) together with the Fourier series expansions

$$H(\theta) = \sum_{n\in\mathbf{Z}} \mathrm{e}^{in\theta}H_n, \quad \Delta(\psi) = \sum_{n\in\mathbf{Z}} \mathrm{e}^{in\psi}\Delta_n. \tag{3.16}$$

Note that $H_n^* = H_{-n} = H_n$ and $\Delta_n^* = \Delta_{-n} = \Delta_n$, since $H(\theta)$ and $\Delta(\theta)$ are assumed to be real even functions of θ. Equations (3.14) and (3.15) imply that

$$\widehat{J}_n(q)^* = (-1)^n \widehat{J}_{-n}(q), \quad \widehat{J}_{-n}(q) = \widehat{J}_n(q). \tag{3.17}$$

Denote the set of solutions to (3.12) by $\{(\lambda_j(q), \mathbf{U}_j(q)),\ j \in \mathbf{Z}\}$. We now establish conditions under which the eigenvalues $\lambda_j(q)$ are real for all $q \in \mathbf{R}$. The case $q = 0$ is trivial because $\widehat{J}_n(0) \sim \delta_{n,0}$ so that

$$\lambda_n(0) = -A_n(0) + B_n(0)H_n. \tag{3.18}$$

Therefore, we take $q \neq 0$. Introduce the inner product of two periodic functions $V(\theta), U(\theta)$ according to

$$\langle V|U\rangle = \int_{-\pi}^{\pi} V^*(\theta)U(\theta)\frac{d\theta}{2\pi} = \sum_{n\in\mathbf{Z}} V_n^* U_n = \langle \mathbf{V}|\mathbf{U}\rangle, \tag{3.19}$$

where V^* denotes the complex conjugate of V. The adjoint matrix $\mathbf{L}(q)^\dagger$ is then given by

$$[\mathbf{L}(q)^\dagger \mathbf{V}]_n \equiv -A_n(q)V_n + \frac{1}{2}\sum_{m\in\mathbf{Z}} \widehat{J}_{m-n}(q)^* B_m(q)[H_m + H_n]V_m. \tag{3.20}$$

Equation (3.17) implies that $\mathbf{L}(q)^\dagger$ and $\mathbf{L}(q)$ have the same set of eigenvalues,

$$\mathbf{L}(q)\mathbf{U}_j(q) = \lambda_j(q)\mathbf{U}_j(q), \quad \mathbf{L}(q)^\dagger \mathbf{V}_j(q) = \lambda_j(q)\mathbf{V}_j(q), \tag{3.21}$$

with the corresponding eigenvectors related according to

$$U_{j,n}(q) = (-1)^n B_n(q) V_{j,n}(q). \tag{3.22}$$

This relationship is invertible since $B_n(q) > 0$ for all $n \in \mathbf{Z}$ and $q \neq 0$. It further follows that

$$\begin{aligned} \lambda_j \langle \mathbf{V}_{j'} | \mathbf{U}_j \rangle &= \langle \mathbf{V}_{j'} | \mathbf{L}\mathbf{U}_j \rangle \\ &= \langle \mathbf{L}^\dagger \mathbf{V} | \mathbf{U} \rangle \\ &= \lambda_{j'}^* \langle \mathbf{V}_{j'} | \mathbf{U}_j \rangle. \end{aligned} \tag{3.23}$$

Hence, if $\lambda_j(q) \neq \lambda_{j'}(q)$ for $j \neq j'$, then the vectors $\mathbf{U}_j(q)$ and $\mathbf{V}_j(q)$ form a biorthogonal system with

$$\langle \mathbf{V}_{j'}(q) | \mathbf{U}_j(q) \rangle = \chi_j(q)\delta_{j,j'} \tag{3.24}$$

and

$$\chi_j(q) = \langle \mathbf{V}_j(q) | \mathbf{U}_j(q) \rangle = \sum_n (-1)^n B_n(q) |V_{j,n}(q)|^2. \tag{3.25}$$

It also follows that $\lambda_j(q) = \lambda_j^*(q)$ if $\chi_j(q) \neq 0$. The latter condition certainly holds for all $q \neq 0$ and $j \in \mathbf{Z}$ when $\Delta(\psi)$ is π-periodic,

$$\Delta(\psi) = \Delta(\psi + \pi), \tag{3.26}$$

corresponding to the situation in which collinear interactions are equally strong at the front and at the back (see Figure 2.1(b)). In this case $\Delta_n = 0$ and $\widehat{J}_n(q) = 0$ for all odd integers n, and $\mathbf{L}(q)$ becomes a real matrix that only couples together even-to-even or odd-to-odd components U_n. Hence, $\chi_j(q) = \pm\sum_n^{\pm} B_n(q)|V_{j,n}(q)|^2 \neq 0$, where $\sum_n^{\pm}$ denotes the sum over even and odd integers, respectively. The corresponding eigenfunctions satisfy either $U(\theta + \pi) = U(\theta)$ or $U(\theta + \pi) = -U(\theta)$ and can be taken to be real-valued. If $\Delta(\psi)$ is not π-periodic, then the eigenvalues $\lambda_j(q)$ are still real, provided that $\chi_j(q) \neq 0$ except at isolated points, which follows from the observation that $\lambda_j(q)$ is a continuous function of q. In this more general situation, however, the eigenfunctions $U(\theta)$ will be complex-valued.

3.2. Perturbation expansion. The calculation of the eigenvalues and eigenfunctions of the linearized equation (3.5), and hence the derivation of conditions for the marginal stability of the homogeneous state, has been reduced to the problem of solving the matrix equation (3.12). In general it is not possible to solve this equation exactly. Here we will carry out a perturbation expansion under the assumption that

the dependence of the interactions on relative direction in the plane is weak. In other words, we write

$$\Delta(\psi) = 1 + \beta\Psi(\psi), \quad 0 \leq \beta \ll 1, \quad |\Psi(\psi)| \leq 1 \text{ for all } \psi, \tag{3.27}$$

with $\Psi_0 = \int_{-\pi}^{\pi} \Psi(\psi)d\psi = 0$. Equation (3.12) can then be rewritten in the form

$$[\lambda(q) - W_n(q)]\, U_n = \beta \sum_{n\in\mathbf{Z}} \widehat{W}_{nm}(q)U_m, \tag{3.28}$$

where

$$W_n(q) = -A_n(q) + B_n(q)H_n\widehat{G}_0(q) \tag{3.29}$$

and

$$\widehat{W}_{nm}(q) = \frac{1}{2}B_n(q)[H_m + H_n]\Psi_{n-m}\widehat{G}_{n-m}(q) \tag{3.30}$$

with $A_n(q), B_n(q), \widehat{G}_n(q)$ given by (3.13) and (3.15). Equation (3.28) can then be solved by expanding as a power series in β and using degenerate perturbation theory.

Case $\beta = 0$. In the limiting case that the interaction kernel is independent of relative direction in the plane, $\beta = 0$, (3.28) reduces to the result previously obtained for separable kernels [18]: the eigenvalues are

$$\lambda_n(q) = W_n(q) = -(D_1n^2 + D_2q^2) + \overline{f}(\eta_1n^2 + \eta_2q^2)H_n\mathrm{e}^{-q^2/2} \tag{3.31}$$

with corresponding eigenfunctions

$$a_{n,\mathbf{k}}(\mathbf{r},\theta) = \mathrm{e}^{in\theta}\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}, \quad |\mathbf{k}| = q. \tag{3.32}$$

We have used the result $\widehat{G}_0(q) = \mathrm{e}^{-q^2/2}$, which follows from substituting (2.9) into (3.15) and evaluating the resulting Gaussian integral. Note that the full interaction kernel W is invariant with respect to the group $\mathbf{E}(2)\times\mathbf{O}(2)$ so that the odd and even modes $a_{n,\mathbf{k}} \pm a_{-n,\mathbf{k}}$ are degenerate, that is, $\lambda_{-n}(q) = \lambda_n(q)$.

Case $\beta > 0$. For nonzero β, there is a q-dependent *splitting* of the pair of degenerate eigenvalues $\lambda_{\pm n}(q)$, $n \neq 0$, which separates out odd and even solutions. Denoting the characteristic size of such a splitting by $\delta\lambda = \mathcal{O}(\beta)$, we impose the condition that $\delta\lambda \ll \Delta W$, where $\Delta W = \min\{W_n - W_m, m \neq \pm n\}$. This ensures that the perturbation does not excite states associated with other eigenvalues of the unperturbed problem. We can then restrict ourselves to calculating perturbative corrections to the degenerate eigenvalues $\lambda_{\pm n}$ and their associated eigenfunctions. Therefore, introduce the power series expansions

$$\lambda_{\pm n} = W_n + \beta\lambda^{(1)}_{\pm n} + \beta^2\lambda^{(2)}_{\pm n} + \cdots \tag{3.33}$$

and

$$U_{\pm n,m} = z_{\pm n}\delta_{m,\pm n} + \beta U^{(1)}_{\pm n,m} + \beta^2 U^{(2)}_{\pm n,m} + \cdots, \tag{3.34}$$

where $\delta_{n,m}$ is the Kronecker delta function. Here $U_{n,m}$ is the mth component of the vector $\mathbf{U}_n$ associated with the eigenvalue λ_n. Substitute these expansions into the matrix eigenvalue equation (3.28) and systematically solve the resulting hierarchy of

equations to successive orders in β using degenerate perturbation theory along similar lines to [3]. This leads to the following results:

(i) the even (+) and odd (−) eigenvalues to $\mathcal{O}(\beta^2)$ are

$$\lambda_{\pm n}(q) = W_n(q) + \beta\left[\widehat{W}_{nn}(q) \pm \widehat{W}_{n,-n}(q)\right] + \beta^2 \sum_{0\le m\neq n} \frac{\left[\widehat{W}_{nm}(q) \pm \widehat{W}_{-n,m}(q)\right]\left[\widehat{W}_{mn}(q) \pm \widehat{W}_{m,-n}(q)\right]}{W_n - W_m}; \tag{3.35}$$

(ii) the corresponding eigenfunctions to $\mathcal{O}(\beta)$ are

$$a_{n,\mathbf{k}}(\mathbf{r},\theta) = \left[\cos(n\theta) + \beta \sum_{0\le m\neq n} U_m^+(q)\cos(m\theta)\right] \mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}; \tag{3.36}$$

$$a_{-n,\mathbf{k}}(\mathbf{r},\theta) = \left[\sin(n\theta) + \beta \sum_{0<m\neq n} U_m^-(q)\sin(m\theta)\right] \mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}} \tag{3.37}$$

with $|\mathbf{k}| = q$ and

$$U_0^+(q) = \frac{\widehat{W}_{0n}(q)}{W_n - W_0}, \quad U_m^\pm(q) = \frac{\widehat{W}_{mn}(q) \pm \widehat{W}_{m,-n}(q)}{W_n - W_m}, \quad 0 < m \neq n. \tag{3.38}$$

It is important to stress that the splitting of even and odd branches for $\beta > 0$ is a consequence of the underlying shift-twist symmetry and thus occurs beyond the small β regime.

3.3. Marginal stability. We now determine the marginal stability boundaries in parameter space that separate regions of stability from regions of instability. Crossing one of these boundaries signals that one or more eigenvalues become positive and their corresponding eigenfunctions start to grow, leading to the formation of a self-organizing pattern. (Conservation of population number implies that there always exists one neutrally stable mode $q = 0, n = 0$, that is, $\lambda_0(0) = 0$.) For concreteness, we treat the rescaled diffusion coefficients $D_1 \to D_1\overline{f}, D_2 \to D_2\overline{f}$ as bifurcation parameters. An adiabatic increase in the mean density $\overline{f}$ due to a growth in population number or a contraction in the area occupied by the population then corresponds to a reduction in D_1 and D_2. This reduction can lead to one of four distinct types of instability, depending on which eigenmodes are first excited:

(I) if $q \neq 0, n = 0$, then a spatially periodic pattern forms without any angular order (aggregation without orientation);

(II) if $q = 0, n \neq 0$, then a pattern with angular order forms that is spatially uniform (orientation without aggregation);

(IIIa,b) if $q \neq 0, n \neq 0$, then a pattern with spatio-angular order forms, in which the angle of preferred orientation changes periodically in space even though the spatial density remains homogeneous. The invariance of the interaction kernel under Euclidean shift-twist symmetry implies that when $\beta > 0$, there are two kinds of spatio-angular patterns, corresponding to (a) even and (b) odd eigenmodes, respectively. Examples of such eigenmodes are shown in Figure 3.1 under the simplifying assumption that a particular harmonic component n dominates. The crucial observation is that the direction of preferred orientation is correlated with the direction

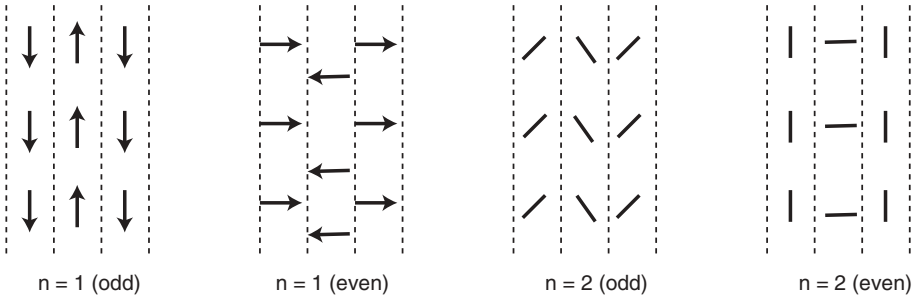


FIG. 3.1. *Examples of even ($a_{n,\mathbf{k}} = \cos(n\theta)e^{i\mathbf{k}\cdot\mathbf{r}} + c.c.$) and odd ($a_{n,\mathbf{k}} = \sin(n\theta)e^{i\mathbf{k}\cdot\mathbf{r}} + c.c.$) eigenmodes with spatio-angular order in the case of a horizontal wavevector $\mathbf{k} = (q, 0)$. If $n = 1$, then there is a unique direction of preferred orientation (indicated by an arrow). If $n = 2$, then there are two preferred directions of motion (indicated by a bar). The width of each vertical strip is π/q.*

of the wavevector $\mathbf{k}$, whereas there is no such correlation when $\beta = 0$ so that the distinction between odd and even solutions disappears.

3.3.1. Separable interaction kernel ($\beta = 0$). Before determining the effects of shift-twist symmetry on pattern forming instabilities of the homogeneous state, we first discuss the separable case previously analyzed by Mogilner and Edelstein-Keshet [18]. The eigenvalues for $\beta = 0$ are given by (3.31). In particular, when $n = 0$ we have (on setting $\overline{f} = 1$)

$$\lambda_0(q) = q^2\big(-D_2 + \eta_2 H_0 e^{-q^2/2}\big). \tag{3.39}$$

It is clear that $\lambda_0(q) < 0$ for all $q \neq 0$ if $D_2 > D_{2,0}$, where

$$D_{2,0} = \eta_2 H_0. \tag{3.40}$$

If $D_2 < D_{2,0}$, then the homogeneous state is unstable with respect to the excitation of eigenmodes over a range of wavenumbers that includes the origin so that $q \approx 0, n = 0$. Although this is a type I instability, the emerging pattern tends to involve long-wavelength spatial inhomogeneities. Therefore, the resulting stationary state could be treated as homogeneous on shorter spatial scales and thus be susceptible to secondary bifurcations associated with excitation of modes with $q \neq 0, n \neq 0$ (see below). Alternatively, the $n = 0$ modes could be stabilized by incorporating an additional contribution to the interaction kernel W in the form of a repulsive hard-core potential:

$$W(\mathbf{r}, \theta|\mathbf{r}', \theta') = G(\mathbf{r} - \mathbf{r}')H(\theta - \theta')\Delta(\arg(\mathbf{r} - \mathbf{r}') - \theta) - C\delta(\mathbf{r} - \mathbf{r}') \tag{3.41}$$

with $C > 0$. This modifies the $\beta = 0$ contribution to λ_0 according to

$$\lambda_0(q) = q^2\big(-D_2 + \eta_2 H_0 e^{-q^2/2} - H_0 C\big), \tag{3.42}$$

and the stability condition becomes $D_2 > D'_{2,0}$ with

$$D'_{2,0} = H_0(\eta_2 - C). \tag{3.43}$$

Now consider the branch of eigenmodes for a given n, $n \neq 0$, such that $H_n > 0$ (otherwise $\lambda_n(q) < 0$ for all q). Differentiating (3.31) with respect to q gives

$$\frac{d\lambda_n(q)}{dq} = q\left[-2D_2 + H_n\left(2\eta_2 - (\eta_1 n^2 + \eta_2 q^2)\right)\mathrm{e}^{-q^2/2}\right]. \tag{3.44}$$

This implies that $\lambda_n(q)$ has a maximum at $q = 0$ when $D_2 > D_{2,n}$ and a maximum at $q = q_c \neq 0$ when $D_2 < D_{2,n}$, where

$$D_{2,n} = H_n(\eta_2 - \eta_1 n^2/2). \tag{3.45}$$

Since

$$\lambda_n(0) = n^2[-D_1 + \eta_1 H_n], \tag{3.46}$$

it follows that $\lambda_n(q) < 0$ for all $n \neq 0$ when $D_2 > D_{2,n}$ and $D_1 > D_{1,n}$ with

$$D_{1,n} = \eta_1 H_n. \tag{3.47}$$

On the other hand, if $D_2 > D_{2,n}$ and $D_1 < D_{1,n}$, then there is a type II instability due to excitation of the eigenmode at $q = 0$. Finally, if $0 < D_2 < D_{2,n}$, then there is a type III instability due to excitation of the eigenmodes with wavenumber $q = q_c \neq 0$. This occurs at a critical value $D_1 = \mathcal{F}_n(D_2)$ with $\mathcal{F}_n$ a monotonically decreasing function such that $\mathcal{F}_n(D_{2,n}) = D_{1,n}$ and $\mathcal{F}_n(0) = D'_{1,n} > D_{1,n}$. Note that this last instability can occur only for integers n satisfying $n^2 < 2\eta_2/\eta_1$.

In order to determine the stability of the homogeneous state one now has to combine the stability conditions for all n. For concreteness, suppose the set of coefficients $\{H_n, n \neq 0\}$ has a maximum at $n = n_c$ such that $D_{j,n_c} > D_{j,n}$ for all $n \neq 0, n_c$. This leads to the stability diagram shown in Figure 3.2. In the absence of a repulsive contribution to the interaction kernel, the region of stability is given by the dark shaded region in Figure 3.2. Crossing the vertical boundary at $D_1 = D_{1,n_c}$ induces a type II instability, whereas crossing the horizontal boundary at $D_2 = D_{2,0}$ induces a type I instability. A type III instability can occur only through a secondary bifurcation. On the other hand, when there is a repulsive contribution to the potential, the stability region extends to include both the dark and light shaded regions. A type II instability now occurs on crossing the lower horizontal boundary at $D_2 = D'_{2,0}$, so that there is an additional curved boundary $D_1 = \mathcal{F}_{n_c}(D_2)$ for $D'_{2,0} < D_2 < D_{2,n_c}$. Crossing this boundary induces a type III instability, but there is no separation into even or odd patterns, since the separable interaction kernel is invariant under the standard Euclidean group action.

3.3.2. Nonseparable interaction kernel ($\beta > 0$). We now show that patterns with even or odd spatio-angular order can occur when $\beta > 0$. We take D_1 and D_2 to be close to the curved boundary of the stability region shown in Figure 3.2, where the unperturbed system undergoes a type III instability, and consider $\mathcal{O}(\beta)$ corrections to the eigenvalues $\lambda_{\pm n}(q)$ for $n = n_c$. Using (3.30) and (3.35) with $\Psi_0 = 0$, we find that $\lambda_{\pm n}(q) = -D_1 n^2 + \Lambda_{\pm n}(q) + \mathcal{O}(\beta^2)$, where

$$\Lambda_{\pm n}(q) = -D_2 q^2 + (\eta_1 n^2 + \eta_2 q^2) H_n\left[\mathrm{e}^{-q^2/2} \pm \beta \Psi_{2n} \widehat{G}_{2n}(q)\right]. \tag{3.48}$$

Suppose that $\Lambda_{\pm n}(q)$ has a unique maximum at $q = q_\pm \neq 0$. If $\Lambda_n(q_+) > \Lambda_{-n}(q_-)$, then the homogeneous state will become unstable at the critical point $D_1 = \Lambda_n(q_+)/n^2$

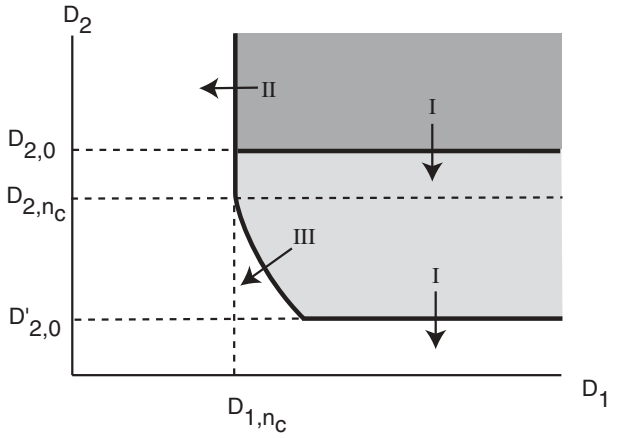


FIG. 3.2. *Stability diagram for the homogeneous state when $\beta = 0$. In the absence of a repulsive contribution to the interaction kernel, the region of stability is given by the dark shaded region. This extends to include the light shaded region when repulsion is included. The type of primary instability induced by crossing each boundary of the stability region is also indicated. (See text for details.)*

due to excitation of even eigenmodes with wavenumber q_+, whereas if $\Lambda_{-n}(q_-) > \Lambda_n(q_+)$, then the homogeneous state will become unstable at the critical point $D_1 = \Lambda_{-n}(q_-)/n^2$ due to excitation of odd eigenmodes with wavenumber q_-.

In order to determine whether an odd or even pattern arises, it is necessary to evaluate the function $\widehat{G}_{2n}(q)$ for the given Gaussian kernel (2.9). Rewriting (3.15) for even integers as

$$\widehat{G}_{2n}(q) = \int_0^\pi \mathrm{e}^{-2in\psi} \left[\int_0^\infty G(r) \cos(rq\cos\psi) r dr \right] d\psi$$

and using the Jacobi–Anger expansion

$$\cos(sq\cos\psi) = J_0(sq) + 2\sum_{m=1}^{\infty} (-1)^m J_{2m}(sq)\cos(2m\psi),$$

with $J_n(x)$ the Bessel function of integer order n, we find that $\widehat{G}_{2n}$ is related to G according to

$$\widehat{G}_{2n}(q) = (-1)^n \int_0^\infty G(r) J_{2n}(rq) r dr. \tag{3.49}$$

Substituting (2.9) into (3.49) and using standard properties of Bessel functions leads to the result

$$\widehat{G}_{2n}(q) = \frac{q\sigma\sqrt{2\pi}}{4} \mathrm{e}^{-\sigma^2 q^2/4} \left[I_{n-1/2}\left(\frac{\sigma^2 q^2}{4}\right) - I_{n+1/2}\left(\frac{\sigma^2 q^2}{4}\right) \right], \tag{3.50}$$

where I_ν is a modified Bessel function.

In Figure 3.3 we plot $\widehat{G}_{2n}(q)$ as a function of wavenumber q for $n = 0, 1, 2$. Note that $\widehat{G}_{2n}(q)$ alternates in sign with n, having a maximum for even n and a minimum for odd n. It follows that if $H_n \Psi_{2n} > 0$, then the homogeneous state destabilizes

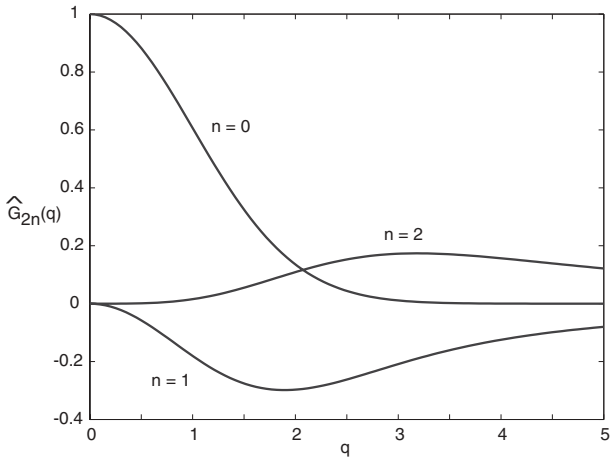


FIG. 3.3. *Plot of* $\widehat{G}_{2n}(q)$ *as a function of wavenumber* q *for* $n = 0, 1, 2$.

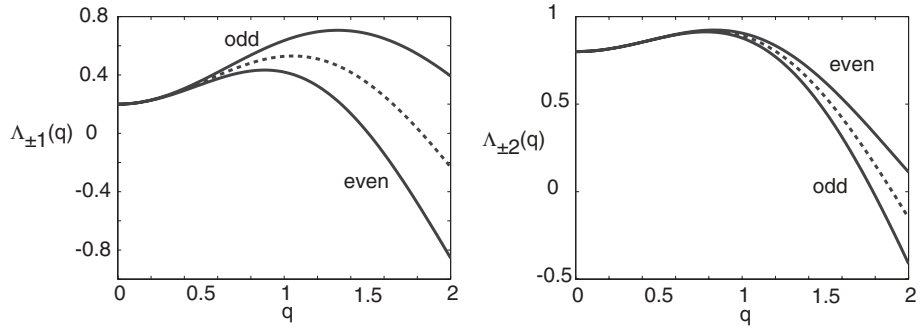


FIG. 3.4. *Plot of* $\Lambda_{\pm n}(q)$ *as a function of wavenumber* q *for* $n = 1, 2$. *Here* $D_2 = 0.2, \eta_1 = 0.3, \eta_2 = 1$, $H_n = 1$, *and* $\beta\Psi_{2n} = 0.5$. *The dashed curve represents the coalescence of the even and odd branches when* $\beta = 0$.

due to excitation of even (odd) eigenmodes when n is even (odd); the opposite result holds for $H_n\Psi_{2n} < 0$. Example dispersion curves $\Lambda_{\pm n}(q)$ are plotted in Figure 3.4 for $n = 1, 2$. Since $\max_q\{\Lambda_{-1}(q)\} > \max_q\{\Lambda_1(q), \Lambda_{\pm2}(q)/4\} > 0$ for the given parameter values, it follows that the critical eigenmodes are odd patterns with $n = 1$ (assuming $D_2 > D'_{2,0}$). Keeping the same parameters but setting $H_1 = 0$ would change the critical eigenmodes to even patterns with $n = 2$. For this particular example, the splitting of the even and odd modes around the critical point is small.

4. Doubly periodic patterns. Rotation symmetry implies that in the case of nonzero critical wavenumber q_c, the space of marginally stable eigenfunctions is infinite-dimensional, consisting of all solutions of the form $u(\theta - \varphi)e^{i\mathbf{k}_\varphi\cdot\mathbf{r}}$, where $u(\theta)$ is either an even or odd function of θ, $\mathbf{k}_\varphi = q_c(\cos\varphi, \sin\varphi)$, and $0 \le \varphi < 2\pi$. However, translation symmetry allows us to restrict the space of solutions of the original equation (2.7) to that of doubly periodic functions. This restriction is standard in many treatments of spontaneous pattern formation, but as yet it has no formal justification. However, there is a wealth of evidence from experiments on convecting fluids and chemical reaction-diffusion systems [23] indicating that such systems tend to generate

TABLE 4.1
Generators for the planar lattices and their dual lattices.

Lattice	$\boldsymbol{\ell}_1$	$\boldsymbol{\ell}_2$	$\hat{\boldsymbol{\ell}}_1$	$\hat{\boldsymbol{\ell}}_2$
Square	$(1,0)$	$(0,1)$	$(1,0)$	$(0,1)$
Hexagonal	$(1,\frac{1}{\sqrt{3}})$	$(0,\frac{2}{\sqrt{3}})$	$(1,0)$	$\frac{1}{2}(-1,\sqrt{3})$
Rhombic	$(1,-\cot\eta)$	$(0,\csc\eta)$	$(1,0)$	$(\cos\eta,\sin\eta)$

doubly periodic patterns in the plane when the homogeneous state is destabilized. Given such a restriction, the associated space of marginally stable eigenfunctions is finite-dimensional. A finite set of specific eigenfunctions can then be identified as candidate planforms, in the sense that they approximate time-independent solutions of (2.7) sufficiently close to the critical point where the homogeneous state loses stability.

Let $\mathcal{L}$ be a planar lattice; that is, choose two linearly independent vectors $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ and let

$$\mathcal{L} = \{2\pi d(m_1\boldsymbol{\ell}_1 + m_2\boldsymbol{\ell}_2) : m_1, m_2 \in \mathbf{Z}\},$$

where d is the lattice spacing. Note that $\mathcal{L}$ is a subgroup of the group of planar translations. A function $f : \mathbf{R}^2 \times \mathbf{S}^1 \to \mathbf{R}$ is *doubly periodic with respect to* $\mathcal{L}$ if

$$f(\mathbf{r}+\boldsymbol{\ell},\theta) = f(\mathbf{r},\theta)$$

for every $\boldsymbol{\ell} \in \mathcal{L}$. Let θ be the angle between the two basis vectors $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$. We can then distinguish three types of lattice according to the value of θ: square ($\theta = \pi/2$), rhombic ($0 < \theta < \pi/2$, $\theta \neq \pi/3$), and hexagonal ($\theta = \pi/3$). After rotation, the generators of the planar lattices are given in Table 4.1 (for unit lattice spacing). Also shown are the generators of the dual lattice

$$\hat{\mathcal{L}} = \{d^{-1}(m_1\hat{\boldsymbol{\ell}}_1 + m_2\hat{\boldsymbol{\ell}}_2) : m_1, m_2 \in \mathbf{Z}\}$$

with $\boldsymbol{\ell}_i \cdot \hat{\boldsymbol{\ell}}_j = \delta_{i,j}$. Restriction to double periodicity means that the original Euclidean symmetry group is now restricted to the symmetry group of the lattice, $\Gamma = H_{\mathcal{L}} \dot{+} \mathbf{T}^2$, where $H_{\mathcal{L}}$ is the *holohedry* of the lattice, the subgroup of $\mathbf{O}(2)$ that preserves the lattice, and $\mathbf{T}^2$ is the two torus of planar translations modulo the lattice. Thus, the holohedry of the rhombic lattice is $\mathbf{D}_2$, the holohedry of the square lattice is $\mathbf{D}_4$, and the holohedry of the hexagonal lattice is $\mathbf{D}_6$.

Imposing double periodicity on the marginally stable eigenfunctions means restricting the lattice spacing d so that the critical wavevector $\mathbf{k}$ lies on the dual lattice. There are infinitely many choices for the lattice size that satisfy this constraint—we select the one for which q_c is the shortest length of a dual wavevector, that is, $q_c = d^{-1}$. Linear combinations of eigenfunctions that generate doubly periodic solutions corresponding to dual wavevectors of shortest length are given by

$$a(\mathbf{r},\theta) = \sum_{j=1}^{N} z_j u(\theta - \varphi_j) \mathrm{e}^{i\mathbf{k}_j\cdot\mathbf{r}} + \text{c.c.}, \tag{4.1}$$

where the z_j are complex amplitudes. Here $N = 2$ for the square lattice with $\mathbf{k}_1 = \mathbf{k}_c$ and $\mathbf{k}_2 = R_{\pi/2}\mathbf{k}_c$, where R_ξ denotes rotation through an angle ξ. Similarly, $N = 3$ for the hexagonal lattice with $\mathbf{k}_1 = \mathbf{k}_c$, $\mathbf{k}_2 = R_{2\pi/3}\mathbf{k}_c$, and $\mathbf{k}_3 = R_{4\pi/3}\mathbf{k}_c = -\mathbf{k}_1 - \mathbf{k}_2$. It follows that the space of marginally stable eigenfunctions can be identified with

the N-dimensional complex vector space spanned by the vectors $(z_1, \dots, z_N) \in \mathbf{C}^N$, with $N = 2$ for square or rhombic lattices and $N = 3$ for hexagonal lattices. It can be shown that these form Γ-irreducible representations. The actions of the group Γ on $\mathbf{C}^N$ can then be explicitly written down for both the odd and even cases [3, 4]. For example, on a hexagonal lattice, a translation $(\mathbf{r}, \theta) \to (\mathbf{r} + \mathbf{s}, \theta)$ induces the action

$$\gamma \circ (z_1, z_2, z_3) = (z_1 \mathrm{e}^{-i\xi_1}, z_2 \mathrm{e}^{-i\xi_2}, z_3 \mathrm{e}^{i(\xi_1+\xi_2)}), \tag{4.2}$$

where $\xi_j = \mathbf{k}_j \cdot \mathbf{s}$, a rotation $(\mathbf{r}, \theta) \to (R_{2\pi/3}\mathbf{r}, \theta + 2\pi/3)$ induces the action

$$\gamma \circ (z_1, z_2, z_3) = (z_3, z_1, z_2), \tag{4.3}$$

and a reflection κ across the x-axis (assuming $\mathbf{k}_c = q_c(1, 0)$) induces the action

$$\gamma \circ (z_1, z_2, z_3) = (z_1, z_3, z_2). \tag{4.4}$$

The next important observation is that, using weakly nonlinear analysis and perturbation methods, it is possible to reduce the infinite-dimensional system (2.7) to a finite set of coupled ODEs constituting an amplitude equation for $\mathbf{z}$,

$$\frac{dz_j}{dt} = F_j(\mathbf{z}), \quad j = 1, \dots, N, \tag{4.5}$$

which is equivariant with respect to the induced shift-twist action of the group Γ on $\mathbf{C}^N$. One can now use techniques from symmetric bifurcation theory to determine the equilibrium solutions that are likely to bifurcate from the homogeneous fixed point $\mathbf{z} = 0$. This analysis has been carried out elsewhere within the context of a continuum model of visual cortex [3, 4]. Since the population model (2.7) has the same Euclidean shift-twist symmetry as the cortical model, it has the same restrictions regarding the types of bifurcations from a homogeneous state that can occur. However, which particular bifurcation scenario is realized in practice may differ in the two models. That is, although symmetry considerations restrict the form of the nonlinear functions F_j appearing in the amplitude equation (4.5) [3, 4], the values of the coefficients multiplying terms at a particular order in z_j will be model-dependent. Determining these coefficients would require carrying out an explicit perturbation calculation. Here we focus on general aspects of the bifurcating solutions that can be deduced from symmetry principles. For completeness we briefly review a few basic definitions and results from equivariant bifurcation theory [13].

Isotropy subgroups. The symmetries of any particular equilibrium solution $\mathbf{z}$ form a subgroup called the *isotropy* subgroup of $\mathbf{z}$ defined by

$$\Sigma_{\mathbf{z}} = \{\sigma \in \Gamma : \ \sigma\mathbf{z} = \mathbf{z}\}. \tag{4.6}$$

More generally, we say that Σ is an isotropy subgroup of Γ if $\Sigma = \Sigma_{\mathbf{z}}$ for some $\mathbf{z} \in V$. Isotropy subgroups are defined up to some conjugacy. A group Σ is conjugate to a group $\widehat{\Sigma}$ if there exists $\sigma \in \Gamma$ such that $\widehat{\Sigma} = \sigma^{-1}\Sigma\sigma$. The *fixed-point subspace* of an isotropy subgroup Σ, denoted by $\mathrm{Fix}(\Sigma)$, is the set of points $\mathbf{z} \in V$ that are invariant under the action of Σ,

$$\mathrm{Fix}(\Sigma) = \{\mathbf{z} \in V : \ \sigma\mathbf{z} = \mathbf{z} \ \forall \ \sigma \in \Sigma\}. \tag{4.7}$$

Finally, the *group orbit* through a point $\mathbf{z}$ is

$$\Gamma\mathbf{z} = \{\sigma\mathbf{z} : \ \sigma \in \Gamma\}. \tag{4.8}$$

TABLE 4.2

Even planforms with $u(-\theta) = u(\theta)$. The hexagon solutions (0) and (π) have the same isotropy subgroup, but they are not conjugate solutions.

Lattice	Name	Planform eigenfunction
Square	Even square	$u(\theta)\cos x + u\left(\theta - \frac{\pi}{2}\right)\cos y$
	Even roll	$u(\theta)\cos x$
Rhombic	Even rhombic	$u(\theta)\cos(\mathbf{k}_1\cdot\mathbf{r}) + u(\theta-\eta)\cos(\mathbf{k}_2\cdot\mathbf{r})$
	Even roll	$u(\theta)\cos(\mathbf{k}_1\cdot\mathbf{r})$
Hexagonal	Even hexagon (0)	$u(\theta)\cos(\mathbf{k}_1\cdot\mathbf{r}) + u\left(\theta+\frac{\pi}{3}\right)\cos(\mathbf{k}_2\cdot\mathbf{r}) + u\left(\theta-\frac{\pi}{3}\right)\cos(\mathbf{k}_3\cdot\mathbf{r})$
	Even hexagon (π)	$u(\theta)\cos(\mathbf{k}_1\cdot\mathbf{r}) + u\left(\theta+\frac{\pi}{3}\right)\cos(\mathbf{k}_2\cdot\mathbf{r}) - u\left(\theta-\frac{\pi}{3}\right)\cos(\mathbf{k}_3\cdot\mathbf{r})$
	Even roll	$u(\theta)\cos(\mathbf{k}_1\cdot\mathbf{r})$

If $\mathbf{z}$ is an equilibrium solution of (4.5), then so are all other points of the group orbit (by equivariance). One can now adopt a strategy that restricts the search for solutions of (4.5) to those that are fixed points of a particular isotropy subgroup. In general, if a dynamical system is equivariant under some symmetry group Γ and has a solution that is a fixed point of the full symmetry group, then we expect a loss of stability to occur upon variation of one or more system parameters. Typically such a loss of stability will be associated with the occurrence of new solution branches with isotropy subgroups Σ smaller than Γ. One says that the solution has spontaneously broken symmetry from Γ to Σ. Instead of a unique solution with the full set of symmetries Γ, a set of symmetrically related solutions (orbits under Γ modulo Σ) each with symmetry group (conjugate to) Σ is observed.

Equivariant branching lemma (see [13]). The system of equations (4.5) has a fixed point $z = 0$ of the full symmetry group Γ. The equivariant branching lemma states that generically there exists a (unique) equilibrium solution bifurcating from the fixed point for each of the axial subgroups of Γ under the given group action—a subgroup $\Sigma \subset \Gamma$ is *axial* if $\dim \mathrm{Fix}(\Sigma) = 1$. The heuristic idea underlying this lemma is as follows. Let Σ be an axial subgroup and $\mathbf{z} \in \mathrm{Fix}(\Sigma)$. Equivariance of F then implies that

$$\sigma F(\mathbf{z}) = F(\sigma\mathbf{z}) = F(\mathbf{z}) \tag{4.9}$$

for all $\sigma \in \Sigma$. Thus $F(\mathbf{z}) \in \mathrm{Fix}(\Sigma)$ and the system of coupled ODEs (4.5) can be reduced to a single equation in the fixed-point space of Σ. Thus one can systematically identify the various expected primary bifurcation branches by constructing the associated axial subgroups and finding their fixed points. The calculation of these subgroups has been carried out elsewhere [3, 4], and the resulting even and odd planforms are listed in Tables 4.2 and 4.3.

One way to represent the planforms graphically is to indicate the direction(s) of preferred orientation at each point in space $\mathbf{r}$, that is, the orientations that maximize the state $a(\mathbf{r}, \theta)$ for fixed $\mathbf{r}$. This has been carried out elsewhere in the case of cortical patterns, where the preferred orientation corresponds to the orientation of a local visual stimulus that elicits the maximum response of a neuron at a particular location in the cortex [3, 4]. From a mathematical rather than a physical viewpoint, the only difference between the cortical patterns and those of the population model is that in the former case the functions $u(\theta)$ are always restricted to be π-periodic, and hence, the resulting spatio-angular patterns are line fields. As a simple example, consider a square lattice with $u(\theta) = \cos 2\theta$ or $u(\theta) = \sin 2\theta$ and $\mathbf{k}_c = 2\pi(1, 0)$. The

TABLE 4.3
Odd planforms with $u(-\theta) = -u(\theta)$.

Lattice	Name	Planform eigenfunction
Square	Odd square	$u(\theta)\cos x - u\left(\theta - \frac{\pi}{2}\right)\cos y$
	Odd roll	$u(\theta)\cos x$
Rhombic	Odd rhombic	$u(\theta)\cos(\mathbf{k}_1\cdot\mathbf{r}) + u(\theta-\eta)\cos(\mathbf{k}_2\cdot\mathbf{r})$
	Odd Roll	$u(\theta)\cos(\mathbf{k}_1\cdot\mathbf{r})$
Hexagonal	Odd hexagon	$u(\theta)\cos(\mathbf{k}_1\cdot\mathbf{r}) + u\left(\theta+\frac{\pi}{3}\right)\cos(\mathbf{k}_2\cdot\mathbf{r}) + u\left(\theta-\frac{\pi}{3}\right)\cos(\mathbf{k}_3\cdot\mathbf{r})$
	Triangle	$u(\theta)\sin(\mathbf{k}_1\cdot\mathbf{r}) + u\left(\theta+\frac{\pi}{3}\right)\sin(\mathbf{k}_2\cdot\mathbf{r}) + u\left(\theta-\frac{\pi}{3}\right)\sin(\mathbf{k}_3\cdot\mathbf{r})$
	Patchwork quilt	$u\left(\theta+\frac{\pi}{3}\right)\cos(\mathbf{k}_2\cdot\mathbf{r}) - u\left(\theta-\frac{\pi}{3}\right)\cos(\mathbf{k}_3\cdot\mathbf{r})$
	Odd roll	$u(\theta)\cos(\mathbf{k}_1\cdot\mathbf{r})$

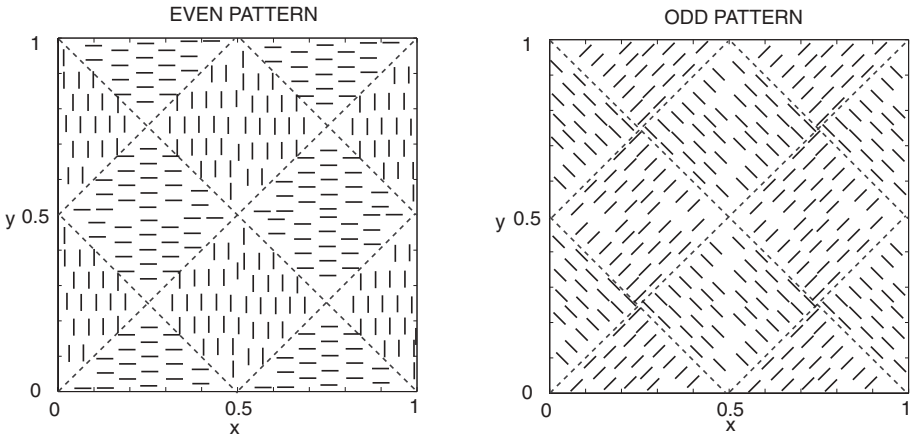


FIG. 4.1. *Line fields for even patterns ($u(\theta) = \cos 2\theta$) and odd patterns ($u(\theta) = \sin 2\theta$) in a fundamental domain of a square lattice. The domain is divided up into subregions where the preferred orientation (*mod π*) is uniform. The direction of orientation within each subregion is indicated by the small parallel bars, which could be interpreted as aligned objects at discretely sampled points within the subregion. Dashed lines indicate line singularities separating regions of different orientation.*

corresponding even and odd planforms (modulo an arbitrary translation) are

$$\begin{aligned} a_+(\mathbf{r},\theta) &= \cos 2\theta(\sin 2\pi x - \sin 2\pi y), \\ a_-(\mathbf{r},\theta) &= \sin 2\theta(\sin 2\pi x - \sin 2\pi y). \end{aligned} \tag{4.10}$$

In the case of the even eigenmode $a_+(\mathbf{r},\theta)$, the preferred orientation (mod π) at (x,y) is $\theta = 0$ when $\sin 2\pi x > \sin 2\pi y$ and $\theta = \pi/2$ when $\sin 2\pi x < \sin 2\pi y$. The corresponding preferred orientations of the odd eigenmode $a_-(\mathbf{r},\theta)$ are $\theta = \pi/4, 3\pi/4$. Note that line singularities occur for $\sin 2\pi x = \sin 2\pi y$, across which there are jumps in orientation preference. The resulting even and odd line fields are shown in Figure 4.1. Inclusion of higher harmonic contributions to the function $u(\theta)$ can lead to point rather than line singularities as well as sites containing more than one preferred orientation [4]. If there is no distinction between "head" and "tail," then solutions of the population model (2.7) will also be π-periodic. On the other hand, if there is such a distinction, then generically the resulting patterns will be represented by vector fields rather than line fields. In the case of a square lattice with $u(\theta) = \cos\theta$ or $u(\theta) = \sin\theta$ and $\mathbf{k}_c = 2\pi(1,0)$, the corresponding even and odd planforms (modulo

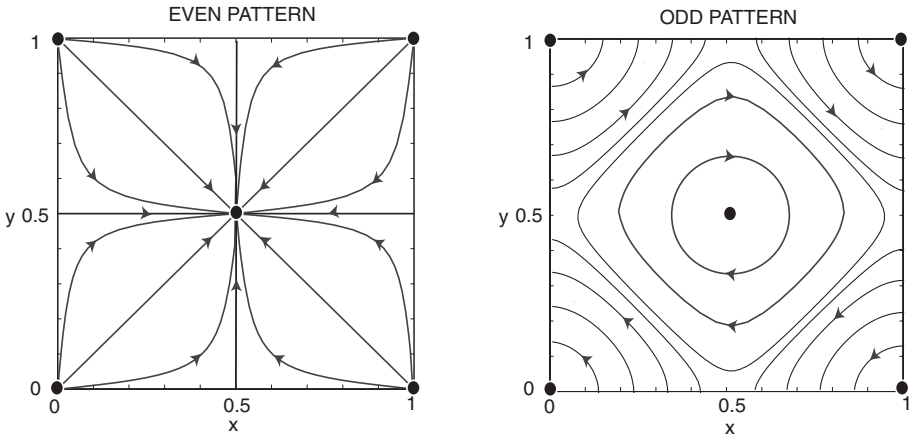


FIG. 4.2. *Vector fields for even patterns ($u(\theta) = \cos\theta$) and odd patterns ($u(\theta) = \sin\theta$) in a fundamental domain of a square lattice. The preferred orientation at a given point in the plane is given by the direction tangential to the flow line passing through that point. There is no preferred orientation at the singularities (indicated by filled circles).*

an arbitrary translation) are

$$\begin{aligned} a_+(\mathbf{r},\theta) &= \cos\theta \sin 2\pi x + \sin\theta \sin 2\pi y = A(\mathbf{r})\cos[\theta - \theta_+(\mathbf{r})], \\ a_-(\mathbf{r},\theta) &= \sin\theta \sin 2\pi x - \cos\theta \sin 2\pi y = A(\mathbf{r})\cos[\theta - \theta_-(\mathbf{r})], \end{aligned} \tag{4.11}$$

with $A(\mathbf{r}) = \sqrt{\sin^2 2\pi x + \sin^2 2\pi y}$ and

$$\theta_+(\mathbf{r}) = \tan^{-1}\frac{\sin 2\pi y}{\sin 2\pi x}, \quad \theta_-(\mathbf{r}) = \tan^{-1}\frac{-\sin 2\pi x}{\sin 2\pi y}. \tag{4.12}$$

It follows that the preferred orientation at position $\mathbf{r}$ is $\theta(\mathbf{r})$, provided that $A(\mathbf{r}) \neq 0$; otherwise there is no preferred orientation. Another way to state this is that the preferred orientation is determined by the flow lines of the vector fields,

$$\mathbf{V}_+ = \sin 2\pi x \frac{\partial}{\partial x} + \sin 2\pi y \frac{\partial}{\partial y}, \quad \mathbf{V}_- = -\sin 2\pi y \frac{\partial}{\partial x} + \sin 2\pi x \frac{\partial}{\partial y}, \tag{4.13}$$

except at the singularities $x = m\pi, y = m'\pi$ for integers m, m'. The resulting vector fields are shown in Figure 4.2. Note that if higher harmonics are included in the function $u(\theta)$, then it is possible for the flow lines to intersect, indicating that there can be more than one preferred orientation away from singularities.

5. Discussion. In this paper we have shown that a wide class of self-organizing biological systems have interactions that are invariant with respect to the shift-twist action of the Euclidean group, and that this has major implications for the types of patterns that can arise in these systems. Our main prediction is that patterns with spatio-angular order should exhibit correlations between the directions of preferred orientation and the underlying spatial orientation of the pattern (as determined by the wavevectors of the excited eigenmodes), and that there are two distinct types of correlation corresponding to scalar and pseudoscalar representations of the Euclidean group. In other words, given a spatially periodic variation in preferred orientation,

there is a correlation between the preferred orientation within a patch and the orientation of the boundaries of that patch. Whether or not such correlations are actually observed in real systems remains to be seen, although the line fields shown in Figure 4.1 are very suggestive of certain arrangements of fibroblasts where the cells appear to be oriented at approximately 45° to line singularities [18].

One of the simplifying assumptions in our analysis has been to restrict the spatial domain to be two-dimensional. This is appropriate for cells grown in vitro on a flat surface and for animal herds. On the other hand, cells in vivo and fish schools or bird flocks [20] are better described by a three-dimensional domain. In the three-dimensional case, the orientation of an individual is specified by points on a sphere (ϕ, θ) with $\phi \in [0, \pi]$ and $\theta \in [0, 2\pi)$. This suggests that the underlying symmetry group is $\mathbf{E}(3) \times \mathbf{O}(3)$ when the corresponding interaction kernel is separable with respect to spatial and angular coordinates. An interesting problem that follows from this is to determine the appropriate three-dimensional analogue of Euclidean shift-twist symmetry when the kernel is taken to be nonseparable. Another factor that would modify the symmetry group is the presence of an environmental gradient that biases the selection of a direction with which to align. Examples include migrating birds using the earth's magnetic field as a directional cue and fibroblasts aligning strongly with grooves on an artificial substrate.

REFERENCES

[1] I. B. Vivancos, P. Chossat, and I. Melbourne, *New planforms in systems of partial differential equations with Euclidean symmetry*, Arch. Ration. Mech. Anal., 131 (1995), pp. 199–224.

[2] W. H. Bosking, Y. Zhang, B. Schofield, and D. Fitzpatrick, *Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex*, J. Neurosci., 17 (1997), pp. 2112–2127.

[3] P. C. Bressloff, J. D. Cowan, M. Golubitsky, P. J. Thomas, and M. Wiener, *Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex*, Phil. Trans. Roy. Soc. London B, 356 (2001), pp. 299–330.

[4] P. C. Bressloff, J. D. Cowan, M. Golubitsky, and P. J. Thomas, *Scalar and pseudoscalar bifurcations: Pattern formation on the visual cortex*, Nonlinearity, 14 (2001), pp. 739–775.

[5] D. Chilingworth and M. Golubitsky, *Symmetry and pattern formation in a planar layer of nematic liquid crystal*, J. Math. Phys., 44 (2003), pp. 4201–4219.

[6] J. Cook, *Waves of alignment in populations of interacting, oriented individuals*, Forma, 10 (1995), pp. 171–203.

[7] G. Civelekoglu and L. Edelstein-Keshet, *Modelling the dynamics of F-actin in the cell*, Bull. Math. Biol., 56 (1994), pp. 587–616.

[8] J. C. Dallon and J. A. Sherratt, *A mathematical model for spatially varying extracellular matrix alignment*, SIAM J. Appl. Math., 61 (2000), pp. 506–527.

[9] P. De Gennes, *The Physics of Liquid Crystals*, Clarendon Press, Oxford, UK, 1974.

[10] T. R. Elsdale and J. B. L. Bard, *Collagen substrata for studies on cell behavior*, J. Cell Biol., 54 (1972), pp. 626–537.

[11] G. B. Ermentrout and J. D. Cowan, *A mathematical theory of visual hallucination patterns*, Biol. Cybernetics, 34 (1979), pp. 137–150.

[12] E. Geigant, K. Ladizhansky, and A. Mogilner, *An integrodifferential model for orientation distributions of f-actin in cells*, SIAM J. Appl. Math., 59 (1998), pp. 787–809.

[13] M. Golubitsky and I. Stewart, *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*, Birkhäuser, Basel, 2002.

[14] D. Grunbaum and A. Okubo, *Modelling social animal aggregations*, in Frontiers of Theoretical Biology, Lecture Notes in Biomath. 100, S. A. Levin, ed., Springer-Verlag, Berlin, 1994, pp. 296–325.

[15] C. T. Lee, M. F. Hoopes, J. Diehl, W. Gilliland, G. Huxel, E. V. Leaver, K. McCann, J. Umbanhowar, and A. Mogilner, *Non-local concepts and models in biology*, J. Theoret. Biol., 210 (2001), pp. 201–219.

[16] A. Mogilner and L. Edelstein-Keshet, *Selecting a common direction* I: *How orientational order can arise from simple contact responses between interacting cells*, J. Math. Biol., 33 (1995), pp. 619–660.
[17] A. Mogilner, L. Edelstein-Keshet, and G. B. Ermentrout, *Selecting a common direction* II: *Peak-like solutions representing total alignment of cell clusters*, J. Math. Biol., 34 (1996), pp. 811–842.
[18] A. Mogilner and L. Edelstein-Keshet, *Spatio-angular order in populations of self-aligning objects: Formation of oriented patches*, Phys. D, 89 (1996), pp. 346–367.
[19] A. Okubo, *Dynamical aspects of animal grouping: Swarms, schools, flocks and herds*, Adv. Biophys., 22 (1986), pp. 1–94.
[20] J. Parrish and W. Hammner, eds., 3*D Animal Aggregations*, Cambridge University Press, Cambridge, UK, 1994.
[21] J. A. Sherratt and M. E. Lewis, *Stress-induced alignment of actin filaments and the mechanics of cytogel*, Bull. Math. Biol., 55 (1993), pp. 637–654.
[22] T. P. Stossel, *How cells crawl*, Amer. Sci., 78 (1990), pp. 408–423.
[23] D. Walgraef, *Spatio-Temporal Pattern Formation*, Springer-Verlag, Berlin, 1997.
[24] L. R. Williams and D. W. Jacobs, *Stochastic completion fields: A neural model of illusory contour shape and salience*, Neural Comput., 9 (1999), pp. 849–870.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1691–1708

BAYESIAN VIDEO DEJITTERING BY THE BV IMAGE MODEL*

JIANHONG SHEN†

Dedicated to all pioneering mathematicians in image and vision analysis, on whose shoulders we the younger generations are standing

Abstract. Line jittering, or random horizontal displacement in video images, occurs when the synchronization signals are corrupted in video storage media, or by electromagnetic interference in wireless video transmission. The goal of intrinsic video dejittering is to recover the ideal video directly from the observed jittered and often noisy frames. The existing approaches in the literature are mostly based on local or semilocal filtering techniques and autoregressive image models and are complemented by various image processing tools. In this paper, based on the statistical rationale of Bayesian inference, we propose the first variational dejittering model based on the bounded variation (BV) image model, which is global, clean and self-contained, and intrinsically combines dejittering with denoising. The mathematical properties of the model are studied based on the direct method of calculus of variations. We design one effective algorithm and present its computational implementation based on techniques from numerical partial differential equations (PDEs) and nonlinear optimizations.

Key words. video, line jittering, Bayesian, bounded variation, variational, partial differential equations

AMS subject classifications. Primary, 94A08; Secondary, 68U10, 65K10

DOI. 10.1137/S0036139902418699

1. Introduction.

The best way to describe video jittering is to quote from the recent remarkable monograph by Kokaram on motion picture restoration [20]:

> Video signals must contain synchronization information to allow the video display to properly locate lines and frames relative to each other in space and time. Noise in the video signal, or degradation of the storage medium on which the signal is stored (video tape) can cause the synchronization signals to be corrupted. This can cause the loss of "lock" in video digitizing and playback apparatus. The loss of line synchronization pulses will prevent the video manipulation device from locating the actual start and end of each line thus yielding random line displacements (line jitter) in the observed video images.

Figure 1.1 displays a typical jittered video frame, in which horizontal image lines are randomly shuffled. In the current paper, we will not consider interframe correlation and jittering. Therefore the dejittering problem, i.e., to recover the original ideal image frame u from the observed jittered (and often noisy due to medium corruption) image frame u_0, is fundamentally a still image restoration problem.

For real analog videos, it is possible to recover the line synchronization information by cleaning the nonpicture parts of the video signals. This is the method of *time base correctors* [20]. It demands the availability of signal information that has nothing to do with the video or image content.

*Received by the editors November 27, 2002; accepted for publication (in revised form) November 18, 2003; published electronically July 2, 2004. This research was supported by the Program of Applied Mathematics of the NSF under grant DMS-0202565.
http://www.siam.org/journals/siap/64-5/41869.html
†School of Mathematics, University of Minnesota, 206 Church Street SE, Minneapolis, MN 55455 (jhshen@math.umn.edu).

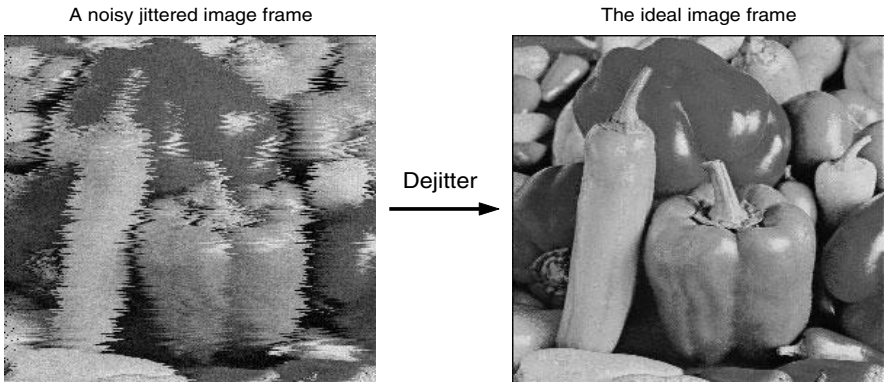


FIG. 1.1. *A noisy jittered image frame and the goal of dejittering.*

The idea of *intrinsic* video dejittering [21, 22], on the other hand, is to restore the ideal image frames directly from the observed jittered data. That is, as in the classical denoising or deblurring problems, one attempts to recover the ideal images solely based on intrinsic image structures (i.e., image prior models).

Thus in comparison, the intrinsic approach is more flexible and is applicable in more general settings. For instance, due to electromagnetic interference in the environment (especially intentional interference, as in battle fields), wireless image signals can experience very similar jittering problems, completely or partially. It is impossible to reconstruct the random and dynamic process of environmental interference, but it is still feasible to recover the original images based on their intrinsic structures.

In the pioneering works by Kokaram and colleagues [21, 22], intrinsic dejittering models are developed based on autoregressive image models, line registration, image interpolation, parameter estimation, and the filtering technique. The goal of the current paper is to address the dejittering problem with the help of calculus of variations and partial differential equations (PDEs), two novel modern tools in mathematical image and vision analysis (Miva).

Our main contributions are highlighted as follows.

(a) Treating dejittering as an image restoration problem, we propose the first variational dejittering model in the literature. The rationale of this deterministic model is the general statistical framework of Bayesian inference.
(b) The model is clean and self-contained, meaning that no other pre- or postprocessing steps are needed. Furthermore, it intrinsically combines the two processes of denoising and dejittering.
(c) We propose to apply the BV (bounded variation) image prior model for faithfully recovering the regularity of the jittered boundaries of image objects. The BV image model was first applied by Rudin and Osher [26] and Rudin, Osher, and Fatemi [27] for image denoising and deblurring. The most beautiful attribute of the BV image model is that it takes care of object boundaries automatically, without the pain of separating them from the *interior* homogeneous regions, which considerably eases the burden of both theory and computation.
(d) By applying the direct method of calculus of variations, we attempt to reveal some important mathematical properties of the proposed dejittering model, including uniqueness, existence, and convergence.

(e) For the nonlinear and nonconvex objective of our model, we design an iterative algorithm which alternately optimizes the image estimation and jittering estimation. This algorithm is then numerically implemented by techniques from computational PDEs and nonlinear optimization.

The organization of the paper is as follows. The statistical assumptions on the jittering and intensity noisy models are stated in section 2. In section 3, the proposed BV-based dejittering model is developed from the Bayesian rationale in decision and inference theory. In section 4, we study the associated admissible conditions and some fundamental properties of the model. The algorithm and its computational implementation are detailed in section 5, and are followed by three typical numerical examples. Section 6 concludes the paper with a brief summary.

2. The statistical jittering model. We begin with some necessary statistical assumptions or models for the line jittering process. For convenience, the image domain is assumed to be a horizontal stripe: $\Omega = \mathbb{R} \times (0, H)$, in which each point (x, y) is called a pixel.

Let $s = s(y)$ be a Gaussian homogeneous white noise on $y \in (0, H)$. That is, there exists no correlation between the jitters of two distinct horizontal lines:

$$\mathbf{E}[s(y_1)s(y_2)] = \sigma_s^2 \delta_{y_1, y_2},$$

where $\delta_{y_1,y_2} = 1$ if $y_1 = y_2$ and 0 otherwise, and σ_s^2 denotes the shared jittering variance. We shall always assume that $s(y)$ has zero mean.

The model of an infinitely long stripe domain Ω is then theoretically consistent with the assumption of Gaussian jitters, since the latter can be any real numbers.

Now suppose $u(x, y), (x, y) \in \Omega$ is the original ideal image to be displayed or transmitted. The horizontal line jittering process is modeled by $u \to u_s$:

$$u_s(x, y) = u(x + s(y), y), \quad (x, y) \in \Omega.$$

Thus generally u_s becomes a random field on Ω. In practice, due to electromagnetic or medium noise, the jittered image u_s is further polluted in terms of grey levels to u_0:

$$u_0(x, y) = u_s(x, y) + n(x, y), \quad (x, y) \in \Omega.$$

Here n denotes Gaussian homogenous white noise with mean 0 and variance σ_n^2, and has been assumed additive.

It shall be further assumed that the jittering s and intensity noise n are independent, since their physical causes are often uncorrelated in applications.

The dejittering problem can then be stated as follows. Suppose only one *single* observation u_0 is made. Find a suitable way to restore the original image u (to certain commercially acceptable standards, say).

This naturally falls into the scope of Bayesian inference: decision making or feature (in our case u) extraction based on observed data. Thus the spirit of our approach is tightly rooted in the Bayesian restoration framework.

3. Bayesian dejittering for BV images. The goal of dejittering is to estimate both the original ideal image u and the particular jitter s involved in the given single observation u_0. In the Bayesian framework, we are to maximize the posterior probability

$$p(u, s \mid u_0) = \frac{p(u_0 \mid u, s)p(u, s)}{p(u_0)}.$$

Jittering is caused by the corruption of the synchronizing signatures and is therefore independent of the image u, which leads to $p(u,s) = p(u)p(s)$. In addition, once the observation u_0 is given, $p(u_0)$ is simply a normalization constant and has no influence in terms of probability maximization. Taking either the logarithmic likelihood function, or formally in terms of statistical mechanics, the Gibbs' ensemble energy $E[\cdot] = -\frac{1}{\beta}\ln p(\cdot)$ (β denoting the reciprocal of the absolute temperature), we are to minimize the posterior "energy"

$$E[u,s \mid u_0] = E[u_0 \mid u,s] + E[s] + E[u].$$

The equality holds up to an additive "grounding" energy level. Throughout this paper, the notation $E[A|B]$ always denotes a functional of A, which depends on the *given* B as well. Both A and B can contain multiple components.

The *data model* $E[u_0 \mid u,s]$ easily follows from the Gaussian assumption:

$$u_0 = u_s + n = u(x + s(y), y) + n(x,y).$$

That is,

$$E[u_0 \mid u,s] = \lim_{R\to\infty} \frac{1}{2\sigma_n^2|\Omega_R|}\int_{\Omega_R}(u_0 - u_s)^2 dxdy, \tag{3.1}$$

where $\Omega_R = (-R,R)\times(0,H)$, and $|\Omega_R| = 2RH$ denotes its Lebesgue measure.

We have two priors: the line jittering model $E[s]$ and the image model $E[u]$. Since the former has been assumed Gaussian, we must have

$$E[s] = \frac{1}{2\sigma_s^2 H}\int_0^H s^2(y)dy. \tag{3.2}$$

Therefore, the key to the Bayesian dejittering approach is to adopt an appropriate image prior $E[u]$.

Identification of effective image priors is probably the most fundamental problem in the entire field of Miva. See, for example, the author's recent expository article in *SIAM News*, in which this viewpoint was explicitly expressed and emphasized [28]. There are a number of valuable image priors in the literature, which can be coarsely classified into two categories: stochastic and deterministic (see, e.g., the recent survey paper by Chan, Shen, and Vese [10]).

Stochastic image priors are typically based on either the lattice model and Gibbs' fields in statistical mechanics (Geman and Geman [14]), or statistical learning via techniques like multiscale filtering and the maximum entropy principle (Zhu and Mumford [32] and Zhu, Wu, and Mumford [33]).

Deterministic image priors emphasize more the property of regularities, in contrast to the focus on various spectral or multiscale statistical features in stochastic priors. Well-known examples include (a) the Sobolev image model $E_2[u] = \alpha\int_\Omega |\nabla u|^2 dxdy < \infty$ in the classical linear filtering theory; (b) the BV [15] model $E_{\rm tv}[u] = \alpha\int_\Omega |Du| < \infty$, first introduced to image restoration by Rudin, Osher, and Fatemi [27]; and (c) the Mumford–Shah [25] object-edge free boundary model

$$E_{\rm ms}[u,\Gamma] = E[u|\Gamma] + E[\Gamma] = \alpha\int_{\Omega\setminus\Gamma}|\nabla u|^2 dxdy + \beta\mathcal{H}^1(\Gamma) < \infty,$$

where Γ denotes the jump set or the collection of "edges," $\mathcal{H}^1$ the one-dimensional Hausdorff measure, and α and β two tunable weights. Recently, motivated by the image inpainting problem, we have also investigated high order geometric image models

such as the elastica model (Chan, Kang, and Shen [5]) and the Mumford–Shah–Euler image model (Esedoglu and Shen [13]), where high order geometric information such as the mean curvature is also taken into account.

In the current paper, we choose to work with the Rudin–Osher–Fatemi BV image model

$$E_{\mathrm{tv}}[u] = \alpha \int_{\Omega} |Du| < 0. \tag{3.3}$$

Notice that here $|Du|$ denotes the Radon measure of a BV function [15]. For the more restricted Sobolev $W^{1,1}$ images, $E_{\mathrm{tv}}[u]$ is simply the ordinary L^1 norm of ∇u. Numerous applications have demonstrated that the BV image model is well balanced in terms of fidelity in approximating generic images (especially those mainly containing man-made objects), theoretical tractability, and computational complexity [1, 3, 4, 6, 7, 8, 11, 18, 26, 27, 29]. Such advantages can also be witnessed in the rest of the paper. We refer to our most recent survey paper [9] for a concise overview of the role of the BV image model in Miva.

It is worth pointing out that *digitally* the BV image model can be approximately realized by certain adaptive (and thus nonlinear) autoregressive (AR) models (see, e.g., [2, 6, 20]). However, in the continuum limit, the rich mathematical theory of BV functions provides an independent, clean, and rigorous framework in both theory and computation [15, 4, 9].

The combination of the three models (3.1), (3.2), (3.3) leads to the complete Bayesian posterior "energy" to be minimized:

$$E[u, s \mid u_0] = \lim_{R\to\infty} \frac{\lambda_R}{2} \int_{\Omega_R} (u_0 - u_s)^2 dxdy + \frac{\mu}{2} \int_0^H s^2(y)dy + \alpha \int_{\Omega} |Du|, \tag{3.4}$$

where $\lambda_R = 1/(\sigma_n^2|\Omega_R|)$ and $\mu = 1/(\sigma_s^2 H)$. Notice that the total variation weight α is the only tunable parameter. The impossibility of having a universally working α is closely connected to the undefinability of a scale-invariant probability measure over "all" images (see the extraordinary recent work by Mumford and Gidas [24]).

The rest of the paper focuses on the properties and computation of this Bayesian dejittering model. In what follows, we first argue that model (3.4) has to be modified for allowing nontrivial solutions.

4. Properties of the model. So far the dejittering model (3.4) has been purely inspired by the statistical jittering model and the Gibbs' energy rationale. In this section, we (a) first rigorously define its admissible space, (b) then argue that model (3.4) has to be modified to yield meaningful solutions, and finally (c) study some important properties of the modified model.

4.1. Admissible conditions and correcting model (3.4). From the squared integration term in (3.4) defining the jittering variance s, it is necessary that $s = s(y) \in L^2(0, H)$.

The BV prior naturally requires that $u \in \mathrm{BV}(\Omega)$. By the Sobolev embedding theorem [15] (and its extension to the BV space), $\mathrm{BV}(\Omega)$ is embedded in $L^2(\Omega)$, implying that $\int_\Omega u^2(x, y)dxdy < \infty$.

For any $s \in L^2(0, H)$, define the jittering transform $T_s : \Omega \to \Omega$ by

$$T_s : (x, y) \to (x + s(y), y). \tag{4.1}$$

Then $T_s^{-1} = T_{-s}$. We now show that T_s is a Lebesgue isomorphism. Let $E \subset \Omega$ be any measurable set, and $|E|$ its Lebesgue measure. Denote the characteristic function or indicator of E by $1_E(x, y)$. Then it is easy to see that

$$1_{T_s E}(x, y) = 1_E \circ T_{-s}(x, y).$$

By Fubini's theorem,

$$\begin{aligned} |T_s E| &= \int_\Omega 1_E \circ T_{-s}(x, y)dxdy = \int_0^H dy \int_{\mathbb{R}} 1_E(x - s(y), y)dx \\ &= \int_0^H dy \int_{\mathbb{R}} 1_E(x, y)dx = \int_\Omega 1_E(x, y)dxdy = |E|. \end{aligned}$$

The jittering transform T_s is therefore a Lebesgue isomorphism. In particular, for any $u \in \mathrm{BV}(\Omega) \subset L^2(\|\cdot\|, \Omega)$,

$$u_s = u \circ T_s \in L^2(\Omega) \quad \text{and} \quad \|u_s\| = \|u\|. \tag{4.2}$$

Finally, the data model in (3.4) has been formally motivated by the *law of large numbers*:

$$\sigma_n^2 = \lim_{R\to\infty} \frac{1}{|\Omega_R|} \int_{\Omega_R} n^2(x, y)dxdy.$$

Due to the averaging over the entire infinite domain, we now show that model (3.4) leads to an unexpected dead end. The problem is fixed in the next subsection.

THEOREM 4.1. *Let u_0 be a given noisy jittered image in $L^2_{\mathrm{loc}}(\Omega)$. Suppose that there exists at least one $w(x, y) \in \mathrm{BV}(\Omega)$ such that*

$$\lim_{R\to\infty} \frac{1}{|\Omega_R|} \int_{\Omega_R} n^2(x, y)dxdy = \sigma_n^2$$

exists, with $n = u_0 - w_s = u_0 - w \circ T_s$. Then

$$(u = 0,\ s = 0) = \operatorname{argmin} E[u, s \mid u_0], \qquad E \text{ as in model (3.4)}.$$

Proof. The conclusion follows directly from

$$\lim_{R\to\infty} \frac{1}{|\Omega_R|} \int_{\Omega_R} (u_0 - w_s)^2 dxdy = \lim_{R\to\infty} \frac{1}{|\Omega_R|} \int_{\Omega_R} (u_0 - u_s)^2 dxdy,$$

for any $u \in \mathrm{BV}(\Omega)$, which we are now to prove. Define

$$\langle f, g\rangle_R = \int_{\Omega_R} f(x, y)g(x, y)dxdy, \quad \|f\|_R^2 = \langle f, f\rangle_R.$$

Notice that

$$\|u_0 - u_s\|_R^2 = \|u_0 - w_s\|_R^2 - 2\langle u_0 - w_s, u_s - w_s\rangle_R + \|u_s - w_s\|_R^2. \tag{4.3}$$

Now that $u, w \in \mathrm{BV}(\Omega)$, we must have $u_s - w_s = (u - w)_s \in L^2(\|\cdot\|, \Omega)$. Therefore,

$$\lim_{R\to\infty} \frac{1}{|\Omega_R|} \|u_s - w_s\|_R^2 = \lim_{R\to\infty} \frac{1}{|\Omega_R|} \|u_s - w_s\|^2 = 0.$$

For the cross term in (4.3), by the Schwarz inequality,

$$\frac{1}{|\Omega_R|}|\langle u_0 - w_s, u_s - w_s\rangle_R| \le \frac{1}{|\Omega_R|}\|u_0 - w_s\|_R \, \|u_s - w_s\|_R$$
$$= \left[\frac{1}{|\Omega_R|}\,\|n\|_R^2\right]^{\frac{1}{2}} \left[\frac{1}{|\Omega_R|}\,\|u_s - w_s\|_R^2\right]^{\frac{1}{2}}.$$

As $R \to \infty$, the first term converges to the standard deviation σ_n according to the assumption, but the second term vanishes as just shown. Therefore,

$$\lim_{R\to\infty} \frac{1}{|\Omega_R|}\langle u_0 - w_s, u_s - w_s\rangle_R = 0.$$

In combination, we are able to conclude that

$$\lim_{R\to\infty} \frac{1}{\|\Omega_R\|}\,\|u_0 - w_s\|_R^2 = \lim_{R\to\infty} \frac{1}{\|\Omega_R\|}\,\|u_0 - u_s\|_R^2.$$

This completes the proof. □

4.2. The corrected model and its properties. The problem is caused by averaging near infinity. In applications, images are given only on a bounded domain Ω_R and can be extended over the infinite stripe domain Ω by zero-padding. Thus to diminish the unnecessary role of the infinity, we propose to modify model (3.4) to

$$E[u, s \mid u_0] = \frac{\lambda}{2}\int_\Omega (u_0 - u_s)^2 dxdy + \frac{\mu}{2}\int_0^H s^2(y)dy + \alpha\int_\Omega |Du|, \tag{4.4}$$

where $\mu = 1/(\sigma_s^2 H)$ and $\lambda = \beta/\sigma_n^2$. Besides α, a new tunable parameter β is thus introduced. It is now easy to collect all the admissible conditions:

(a) the given jittered image $u_0 \in L^2(\Omega)$ and

(b) $u \in \mathrm{BV}(\Omega)$, and $s \in L^2(0, H)$.

As in the statistical framework, we define the corresponding "conditional" energies for $E[u, s \mid u_0]$:

$$E[u \mid u_0, s] = \frac{\lambda}{2}\int_\Omega (u_0 - u_s)^2 dxdy + \alpha\int_\Omega |Du| \quad \text{when } s \text{ is known}, \tag{4.5}$$

$$E[s \mid u_0, u] = \frac{\lambda}{2}\int_\Omega (u_0 - u_s)^2 dxdy + \frac{\mu}{2}\int_0^H s^2(y)dy \quad \text{when } u \text{ is known}. \tag{4.6}$$

Since $s \in L^2(0, H)$ is to at least model the rapid oscillations of random jittering, differential constraints (such as Sobolev norms or the BV norm) are generally inappropriate. (Of course, one is able to introduce them with small weights merely for the sake of Tikhonov regularization in inverse problems.) This makes the energy $E[u, s \mid u_0]$ in (4.4) lack the necessary compactness properties for establishing a general existence theorem. Uniqueness is jeopardized as well because of the lack of convexity (especially in s).

However, it is indeed possible to say more about the two "conditional" energies just defined, which will also be important for our algorithm in the next section.

THEOREM 4.2. *For any given jittering s, the minimizer in* $\mathrm{BV}(\Omega)$ *for* $E[u \mid u_0, s]$ *is unique. Furthermore if there exists a minimizing sequence which is bounded in* $L^1(\Omega)$*, then the minimizer indeed exists.*

We shall explain right after the proof why we need the extra boundedness condition for the existence part, which is often unnecessary for image processing problems on finite domains (see Chambolle and Lions [4], for example).

Proof. The jittering operator is linear: $(u+v)_s = u_s + v_s$ for any $u, v \in \mathrm{BV}(\Omega)$. Therefore it is straightforward to establish the strict convexity of $E[u \mid u_0, s]$, which leads to the uniqueness. We now prove the existence.

Let $(u^n)_n \subset \mathrm{BV}(\Omega)$ be a minimizing sequence of $E[u \mid u_0, s]$:

$$\lim_{n\to\infty} E[u^n \mid u_0, s] = \inf_{u\in \mathrm{BV}(\Omega)} E[u \mid u_0, s] < \infty.$$

In addition, assume that the sequence is bounded in $L^1(\Omega)$. There must exist an upper bound M, so that $e_n = E[u^n \mid u_0, s] \le M, n = 1, 2, \dots$, and

$$\int_\Omega |Du^n| \le \frac{1}{\alpha} e_n \le \frac{M}{\alpha}.$$

Therefore, $(u^n)_n$ is bounded in $\mathrm{BV}(\Omega)$. By the weak compactness property, there exist some $w = w(x, y) \in \mathrm{BV}(\Omega)$ and a subsequence, still denoted by $(u^n)_n$ for convenience, such that

$$u^n \to w \ \text{ in } \ L^1(\Omega) \quad \text{and} \quad \int_\Omega |Dw| \le \liminf_{n\to\infty} \int_\Omega |Du^n|.$$

The second half follows from the L^1 lower semicontinuity in $\mathrm{BV}(\Omega)$.

Possibly with another step of subsequence refinement, we can assume that

$$u^n(x, y) \to w(x, y), \quad \text{a.e.} \quad \text{on } \ \Omega.$$

Since the jittering transform $T_s : \Omega \to \Omega$ is a Lebesgue isomorphism, we must have as well

$$u^n_s(x, y) \to w_s(x, y), \quad \text{a.e.} \quad \text{on } \ \Omega.$$

Application of Fatou's lemma to the nonnegative sequence $g_n = (u_0 - u^n_s)^2$ leads to

$$\int_\Omega (u_0 - w_s)^2 dxdy \le \liminf_{n\to\infty} \int_\Omega (u_0 - u^n_s)^2 dxdy.$$

Eventually the above results enable us to conclude that

$$E[w|u_0, s] \le \liminf_{n\to\infty} E[u^n|u_0, s] = \inf_{u\in \mathrm{BV}(\Omega)} E[u \mid u_0, s].$$

Thus w has to be the (unique) minimizer of $E[u \mid u_0, s]$ in $\mathrm{BV}(\Omega)$. □

Remark. We now explain why the L^1 boundedness condition seems to be necessary. It is true that we can bound the L^2 norms of any minimizing sequence similarly to what has been done for the total variation norms in the proof:

$$\|u^n\|^2 = \|u^n_s\|^2 \le 2\|u_0\|^2 + \frac{4}{\lambda} e_n \le 2\|u_0\|^2 + \frac{4M}{\lambda}, \quad n = 1, 2, \dots .$$

If the domain Ω is finite, a Schwarz inequality immediately leads to a common bound on the L^1 norms:

$$\left[\int_\Omega |u(x, y)| dxdy\right]^2 \le \|u\|^2 \, |\Omega|.$$

However, in our case, $\Omega = \mathbb{R} \times (0, H)$ is an infinitely long stripe domain, for which the L^1 norms can indeed be unbounded. For example, define

$$u^n(x,y) = \frac{1}{(1+x^2)^{\frac{n+1}{2n}}}, \quad n = 1, 2, \dots,$$

which are translation invariant along the y direction. Then it is easy to show that

(a) the L^2 norms are bounded: $\|u^n\|^2 \le \pi H$;
(b) all $u^n \in \mathrm{BV}(\Omega)$ and their total variations are always $2H$ exactly; but
(c) their L^1 norms diverge to ∞ by the *monotone convergence theorem:*

$$\int_\Omega |u^n(x,y)|dxdy = H \int_{\mathbb{R}} \frac{1}{(1+x^2)^{\frac{n+1}{2n}}} dx \to H \int_{\mathbb{R}} \frac{1}{(1+x^2)^{\frac{1}{2}}} dx = \infty.$$

Thus it cannot be a Cauchy sequence in $L^1(\Omega)$. By the *monotone convergence theorem* again, it is not compact even in the topology of $L^1_{\mathrm{loc}}(\Omega)$.

One can say much less about the other conditional energy $E[s \mid u_0, u]$, due to the lack of convexity (for uniqueness) and enough regularity (for existence).

One nice property of $E[s \mid u_0, u]$ is that it is *separable* in terms of the horizontal and vertical directions. That is, it can be written as

$$E[s \mid u_0, u] = \int_0^H e(s(y), y)dy, \quad \text{with}$$
$$e(s(y), y) = \frac{\mu}{2} s(y)^2 + \frac{\lambda}{2} \int_{\mathbb{R}} (u_0(x,y) - u(x + s(y), y))^2 dx.$$

Having y fixed, we define $f_0(x) = u_0(x,y), f(x) = u(x,y)$, and $e(t) = e(t,y)$. Then the minimization of $E[s \mid u_0, u]$ is reduced to the minimization of every such single variable function $e(t)$ associated to each y. Notice that

$$e(t) = \frac{\mu}{2} t^2 + \frac{\lambda}{2} \int_{\mathbb{R}} (f_0(x) - f(x+t))^2 dx, \tag{4.7}$$

which is well defined for almost every $y \in (0, H)$ following Fubini's theorem. Notice that $e(t)$ is generally nonconvex, which leads to the uncertainty of uniqueness.

However, we are still able to establish the existence theorem and give an a priori bound on the minimizers.

THEOREM 4.3. *Suppose that $f_0, f \in L^2(\mathbb{R})$ in (4.7). Then*

(a) *$e(t)$ is a continuous function. In particular, the minimizer exists.*
(b) *Suppose $t = s$ is one minimizer of $e(t)$. Then*

$$|s| \le \sqrt{\lambda/\mu}\ (\|f_0\| + \|f\|).$$

Proof. For (a), take the continuity at $t = 0$, for example. Notice that

$$|\, \|f_0(x) - f(x+t)\| - \|f_0(x) - f(x)\| \,| \le \|f(x+t) - f(x)\|.$$

It is well known from Lebesgue integration theory that $f(x+t)$ converges to $f(x)$ in $L^p(\mathbb{R})$ for any $p \in [1, \infty)$ (excluding L^∞) as $t \to 0$. ($p = 2$ in our case.) Therefore $e(t)$ is indeed a continuous function. Since $e(t) \to +\infty$ as $t \to \pm\infty$, the global minima must be attainable at some finite locations. (b) follows easily from $\frac{\mu}{2} s^2 \le e(s) \le e(0)$. ∎

5. The algorithm and numerical results. In this section, we present an iterative algorithm to minimize $E[u, s \mid u_0]$ in (4.4). As one of the referees has kindly pointed out to us, such an algorithm generally falls into the category known in computer sciences as ICM, *iterated conditional models*.

The plan is to alternately minimize the two conditional energies $E[u \mid u_0, s]$ and $E[s \mid u_0, u]$ in (4.5) and (4.6), which were being studied above only in theory. Starting from a pair of initial guesses (u^0, s_0), we generate (u^{n+1}, s_{n+1}) from (u^n, s_n), $n = 0, 1, \dots$, by

$$u^{n+1} = \operatorname{argmin} E[u \mid u_0, s_n], \quad \text{followed by} \tag{5.1}$$

$$s_{n+1} = \operatorname{argmin} E[s \mid u_0, u^{n+1}]. \tag{5.2}$$

The motivation is clear: the current best jittering estimation s_n leads to an improved estimation of the target image u^{n+1}, which in return contributes to the updating of the jittering itself.

Since the energy $E[u, s \mid u_0]$ may have many local minima, convergence to the global minimum is generally unguaranteed. However, we are still able to show that the sequence $(u^n, s_n)_n$ is consistently "down-hill."

THEOREM 5.1. *Let $(u^n, s_n)_n$ be the sequence derived from the above iterative algorithm. Then for $n = 0, 1, \dots$,*

$$E[u^{n+1}, s_{n+1} \mid u_0] \le E[u^n, s_n \mid u_0].$$

Proof. From the energy definitions (3.2), (3.3), (4.4), (4.5), and (4.6), one has

$$\begin{aligned} E[u, s \mid u_0] &= E_{\mathrm{tv}}[u] + E[s \mid u_0, u] \\ &= E[s] + E[u \mid u_0, s]. \end{aligned}$$

Based on these two relations as well as the iteration formulae (5.1) and (5.2), one obtains

$$\begin{aligned} E[u^{n+1}, s_{n+1} \mid u_0] &= E_{\mathrm{tv}}[u^{n+1}] + E[s_{n+1} \mid u_0, u^{n+1}] \\ &\le E_{\mathrm{tv}}[u^{n+1}] + E[s_n \mid u_0, u^{n+1}] && \text{by (5.2)} \\ &= E[u^{n+1} \mid u_0, s_n] + E[s_n] \\ &\le E[u^n \mid u_0, s_n] + E[s_n] && \text{by (5.1)} \\ &= E[u^n, s_n \mid u_0]. \end{aligned}$$

This verifies the claim. ∎

The convergence of the sequence $(u^n, s_n)_n$ is still unclear, although our numerical results always seem to confirm it. (That is, numerically, the sequence invariably converges to some pair (u, s), which appears to be visually meaningful as judged by human observers.) What we are able to establish is the following weak theorem on convergence.

THEOREM 5.2. *Suppose the jittered image $u_0 \in L^2(\Omega)$ and $u_0(\cdot, y)$ is a continuous function in x for almost every $y \in (0, H)$. Let (u^n, s_n) be the sequence generated by the iterative scheme (5.1) and (5.2). Suppose that $(s_n)_n$ a.e. converges to some $s(y) \in L^2(0, H)$ and that $(\|u^n\|_{L^1})_n$ does not converge to ∞. Then there must exist a subsequence $(n_k)_k$ and some $u \in \mathrm{BV}(\Omega)$ so that $u^{n_k} \to u$ in $L^1(\Omega)$ and*

$$E[u, s \mid u_0] \le \lim_{n \to \infty} E[u^n, s_n \mid u_0].$$

Proof. By the preceding theorem, for any n,

$$E[u^n \mid u_0, s_n] \le E[u^n, s_n \mid u_0] \le E[u^0, s_0 \mid u_0].$$

Thus the bound on the total variations is immediate:

$$\int_\Omega |Du^n| \le \frac{1}{\alpha} E[u^0, s_0 \mid u_0], \qquad n = 1, 2, \dots .$$

Since $(\|u^n\|_{L^1})_n$ does not converge to ∞, there must exist a subsequence $(n'_k)_k$ so that $(u^{n'_k})_k$ is a bounded sequence in $L^1(\Omega)$. Therefore, $(u^{n'_k})_k$ is bounded in $\mathrm{BV}(\Omega)$. By the properties of weak compactness and L^1 lower semicontinuity, there must exist a refined subsequence $(n_k)_k$ and some $u \in \mathrm{BV}(\Omega)$ so that $u^{n_k} \to u$ in $L^1(\Omega)$ as $k \to \infty$ and

$$\int_\Omega |Du| \le \liminf_{k\to\infty} \int_\Omega |Du^{n_k}|. \tag{5.3}$$

Possibly with an extra step of subsequence refinement, we can also assume that $u^{n_k} \to u$, a.e. on Ω.

Secondly, since $s_n(y) \to s(y)$ a.e., by Fatou's lemma,

$$\int_0^H s^2(y)dy \le \liminf_{k\to\infty} \int_0^H s_{n_k}^2(y)dy. \tag{5.4}$$

As discussed in section 4.1, the jittering transform

$$T_s : (x, y) \to (x + s(y), y), \qquad (x, y) \in \Omega,$$

is a Lebesgue isomorphism, and $T_s^{-1} = T_{-s}$, which implies that

$$\int_\Omega (u_0 - u^n \circ T_{s_n})^2 dxdy = \int_\Omega (u_0 \circ T_{-s_n} - u^n)^2 dxdy.$$

Now that $s_n \to s$ a.e. on $(0, H)$, we must have, as $n \to \infty$,

$$T_{-s_n}(x, y) = (x - s_n(y), y) \to (x - s(y), y) = T_{-s}(x, y), \qquad \text{a.e. on } \Omega.$$

Since it is assumed that the observed data $u_0(\cdot, y)$ is a continuous function in x for *almost every* $y \in (0, H)$, we must have

$$u_0 \circ T_{-s_n} \to u_0 \circ T_{-s} \qquad \text{a.e. on } \Omega.$$

In combination with the a.e. convergence condition on $(u^{n_k})_k$, this implies that

$$u_0 \circ T_{-s_{n_k}} - u^{n_k} \to u_0 \circ T_{-s} - u \qquad \text{a.e. on } \Omega.$$

Therefore, Fatou's lemma again leads to

$$\int_\Omega (u_0 \circ T_{-s} - u)^2 dxdy \le \liminf_{k\to\infty} \int_\Omega (u_0 \circ T_{-s_{n_k}} - u^{n_k})^2 dxdy,$$

or equivalently,

$$\int_\Omega (u_0 - u_s)^2 dxdy \le \liminf_{k\to\infty} \int_\Omega (u_0 - u_{s_{n_k}}^{n_k})^2 dxdy. \tag{5.5}$$

The combination of the three bounds (5.3), (5.4), and (5.5) eventually completes the proof:

$$E[u, s \mid u_0] \le \liminf_{k\to\infty} E[u^{n_k}, s_{n_k} \mid u_0] = \lim_{n\to\infty} E[u^n, s_n \mid u_0].$$

The last equality follows from the monotonicity in the preceding theorem. □

Notice in the proof how the Lebesgue isomorphism property of the jittering transform allows us to stay away from the touchy issue of the convergence of $u^n_{s_n}$ to u_s. From the proof it is clear that the condition "$u_0(\cdot, y)$ is continuous in x" could be relaxed to "$u_0(\cdot, y)$ is continuous for almost every $x \in \mathbb{R}$."

Next we discuss the computational strategies for the two "conditional" optimization problems (4.5) and (4.6).

5.1. Algorithm for minimizing $E[u \mid u_0, s]$. Recall that, given u_0 and the current available jittering estimation s, the conditional energy for u is given by

$$E[u \mid u_0, s] = \frac{\lambda}{2}\int_\Omega (u_0 - u_s)^2 dxdy + \alpha \int_\Omega |Du|,$$

which is similar to the classical total variation (TV) denoising model of Rudin, Osher, and Fatemi [27], except that now the jittering $u_s = u \circ T_s$ is involved.

Formally, or assuming that $u \in W^{1,1}(\Omega)$ and $\int_\Omega |Du| = \int_\Omega |\nabla u| dxdy$, the first order variation on total variation leads to the well-known curvature derivative (see [27]) in the distributional sense:

$$-\alpha \nabla \cdot \left[\frac{\nabla u}{|\nabla u|}\right].$$

Secondly, the variation on the data model term gives

$$\lambda T_s^*(T_s u - u_0),$$

where $T_s u = u \circ T_s$ and T_s^* denotes the adjoint of T_s. Notice that $T_s^* = T_{-s} = T_s^{-1}$. Eventually we obtain the (formal) Euler–Lagrange equation

$$\frac{\partial E}{\partial u}[u \mid u_0, s] = \lambda(u - u_{0,-s}) - \alpha \nabla \cdot \left[\frac{\nabla u}{|\nabla u|}\right] = 0, \tag{5.6}$$

where $u_{0,-s} = (u_0)_{-s} = u_0 \circ T_{-s}$, with the natural Neumann boundary condition $\frac{\partial u}{\partial \vec{n}} = 0$.

Define $v_0 = u_{0,-s}$ for a given jitter s. To our great surprise, the Euler–Lagrange equation here is identical to that of the Rudin–Osher–Fatemi denoising model when applied to v_0.

Consequently, we are entitled to apply all computational techniques from the rich literature of TV denoising. First, instead of solving the singular equation (5.6), we solve its viscosity approximation:

$$\lambda(u - u_{0,-s}) - \alpha \nabla \cdot \left[\frac{\nabla u}{|\nabla u|_\epsilon}\right] = 0, \quad |a|_\epsilon = \sqrt{a^2 + \epsilon^2}, \tag{5.7}$$

for some small regularizing positive parameter ϵ. Furthermore, this nonlinear elliptic equation is solved by *lagged diffusivity iterations* [12], which is a natural linearization

technique. Let u^n be the current estimation for (5.7). Then u^n is updated to u^{n+1} by

$$\lambda(u^{n+1} - u_{0,-s}) - \alpha\nabla \cdot \left[\frac{\nabla u^{n+1}}{|\nabla u^n|_\epsilon}\right] = 0,$$

with the associated Neumann boundary conditions. To u^{n+1}, the original "diffusivity coefficient" $1/|\nabla u^{n+1}|_\epsilon$ is replaced by that from the previous step $1/|\nabla u^n|_\epsilon$. Convergence of the algorithm is well studied in [4, 12].

5.2. Algorithm for minimizing $E[s \mid u_0, u]$. Given the current best image estimation u, the "conditional" energy for the jittering s is given by

$$E[s \mid u_0, u] = \frac{\lambda}{2}\int_\Omega (u_0 - u_s)^2 dxdy + \frac{\mu}{2}\int_0^H s^2(y)dy. \tag{5.8}$$

As in the last part of section 4, for almost every (in the Lebesgue sense) given $y \in (0, H)$, define $f_0(x) = u_0(x, y)$ and $f(x) = u(x, y)$. Then the minimization of the functional $E[s \mid u_0, u]$ is reducible to one-dimensional energy functions in the form of

$$e(s) = \frac{\mu}{2}s^2 + \frac{\lambda}{2}\int_{\mathbb{R}} (f_0(x) - f(x+s))^2 dx. \tag{5.9}$$

This is a nonlinear function well defined for any given $f_0, f \in L^2(R)$. From the viscosity approximation, we can assume that u belongs to the Sobolev space $H^1(\Omega)$ [15]. By Fubini's theorem, for almost every $y \in (0, H)$, $f(x) = u(x, y) \in H^1(R)$.

The optimal line jittering s must satisfy

$$e'(s) = \mu s - \lambda\int_{\mathbb{R}} (f_0(x) - f(x+s))f'(x+s)dx = 0.$$

Notice that the integration is indeed well defined since $f_0, f(x+s), f'(x+s)$ are all square integrable. Let $\langle f, g\rangle$ denote the inner product in the Hilbert space $L^2(R)$. Then we can rewrite it by

$$e'(s) = \mu s - \lambda\langle f_0 - f(x+s), f'(x+s)\rangle. \tag{5.10}$$

If, furthermore, we assume that the current estimation $u \in H^2(\Omega)$, then Fubini's theorem again implies that $f'' \in L^2(R)$. We can then take the second order derivative following (5.10):

$$e''(s) = \mu + \lambda\langle f'(x+s), f'(x+s)\rangle - \lambda\langle f_0 - f(x+s), f''(x+s)\rangle. \tag{5.11}$$

Assume either the Neumann condition or vanishing condition at $\pm\infty : ff'(\pm\infty) = 0$. Then integration by parts gives

$$\langle -f(x+s), f''(x+s)\rangle = \langle f'(x+s), f'(x+s)\rangle,$$

and eventually

$$e''(s) = \mu + \lambda\langle f_0,\ -f''(x+s)\rangle. \tag{5.12}$$

Our algorithm for minimizing $e(s)$ is then based on the Newton–Raphson method. Starting from an initial guess s_0, we update s_n to s_{n+1} by

$$s_{n+1} = s_n - \frac{e'(s_n)}{e''(s_n)} = \frac{s_n\langle f_0, -f''(x+s_n)\rangle + \langle f_0 - f(x+s_n), f'(x+s_n)\rangle}{\langle f_0, -f''(x+s_n)\rangle + (\mu/\lambda)}. \tag{5.13}$$

Notice that $\lambda \propto 1/\sigma_n^2$ (see section 2). In the absence of intensity noise (corresponding to $\sigma_n = 0$), the ratio $\mu/\lambda = 0$. Then we have a much simpler formula:

$$s_{n+1} = s_n + \frac{\langle f_0 - f(x+s_n), f'(x+s_n)\rangle}{\langle f_0, -f''(x+s_n)\rangle}. \tag{5.14}$$

The feasibility of the algorithm relies on how robust the denominator $e''(s)$ stays away from 0 (i.e., pure convexity or concavity), at least when s is close to the optimal jittering. We now argue heuristically that indeed $e''(s)$ is reasonably well behaved, which has been observed from our numerical implementation as well. Suppose that the observed jittered one-dimensional image $f_0(x)$ has been generated from an image $g(x) \in H^1(R)$ with a jittering t, i.e., $f_0(x) = g(x+t) + n(x)$, where n denotes the Gaussian intensity white noise. Since n is independent of both g and f, or has rapid oscillatory behavior [23], we have

$$\langle f_0(x), -f''(x+s)\rangle = \langle g(x+t), -f''(x+s)\rangle = \langle g'(x+t), f'(x+s)\rangle,$$

where the last equality follows from integration by parts and the vanishing conditions at $\pm\infty$. Therefore, as the estimation pair (f, s) gets close to the genuine pair (g, t), $e''(s)$ is in the order of

$$\mu + \lambda\langle g'(x+t), g'(x+t)\rangle = \mu + \lambda\|g'(x)\|^2,$$

which certainly robustly stays away from the zero. As a byproduct, we also see that the quantity $\|g'(x)\|$ functions like an information measure: larger values mean richer variations in the image and more clues for robustly recovering the jittering.

5.3. Numerical simulation and results. We now briefly discuss some issues in the implementation of the above algorithms.

5.3.1. Neumann boundary jittering model. In numerical simulation and real applications, images are given on a finite square domain $\Omega_R = (-R, R) \times (0, H)$. We therefore need a boundary jittering model. In our simulation, we adopt what we have called the *Neumann boundary jittering model*. For any $s(y)$ and $(x, y) \in \Omega_R$: if $|x+s(y)| \le R$, we define $u_s(x, y) = u(x+s(y), y)$; otherwise, suppose $\pm(x+s(y)) > R$, then we define $u_s(x, y) = u(\pm R, y)$ in the sense of traces for BV functions [15]. The intensity noise model remains untouched:

$$u_0(x, y) = u_s(x, y) + n(x, y), \qquad \text{with Gaussian white noise } n(x, y). \tag{5.15}$$

5.3.2. Parameter tuning. On the fixed finite image domain Ω_R, the data model in both (3.4) and (4.4) bears the exact form following (5.15):

$$E[u_0 \mid u, s] = \frac{\lambda_R}{2}\int_{\Omega_R}(u_0 - u_s)^2 dxdy,$$

with $\lambda_R = 1/(\sigma_n^2|\Omega_R|)$. Therefore in terms of numerical simulation, the dejittering model becomes

$$E[u, s \mid u_0] = \frac{\lambda_R}{2}\int_{\Omega}(u_0 - u_s)^2 dxdy + \frac{\mu}{2}\int_0^H s(y)^2 dy + \alpha\int_{\Omega}|Du|.$$

It allows only one tunable parameter α, since both λ_R and μ are completely determined by the noise model (σ_n^2) and the jittering model (σ_s^2):

$$\lambda_R = \frac{1}{2\sigma_n^2 RH}, \qquad \mu = \frac{1}{\sigma_s^2 H}.$$

Numerically the variances are obtained from any statistical estimators.

FIG. 5.1. *TV-based Bayesian dejittering.*

5.3.3. Random jittering generation. In analog media, in principle a jitter s could be any real number. However in digital implementation, where the image domain Ω_R becomes a matrix of pixel dots, jitters can be restricted only to integers. One could simulate such integer jitters by the continuous Gaussian noise followed by a quantization step. In this paper, instead, we directly make use of the binomial noise $B(p, N)$ with an even integer $N = 2n$:

$$\text{Prob}(s = k) = \binom{N}{n+k} p^{n+k}(1-p)^{n-k}, \quad k = -n, -n+1, \dots, n-1, n.$$

From probability theory, one has

$$\mathbf{E}(s) = N(1-p) - n = n(1-2p), \qquad \sigma_s^2 = Np(1-p).$$

Since it is assumed in section 2 that the jittering has zero mean, we must have $p = 1/2$ and $\sigma_s^2 = N/4$. The *central limit theorem* confirms that Gaussian is still a good approximation. In our simulations, N is in the order of 100, which brings the standard deviation σ_s close to 5 pixels. Jittering with a several-pixel magnitude is already severe for real video tapes or television signals (see [20], for example).

5.3.4. Simulation examples. Finally, we present three numerical results derived from the model and its algorithm detailed above, and discuss their implications.

In Figure 5.1, from left to right are the initial noisy jittered image u_0, the first image estimation $u^1 = \text{argmin } E[u \mid u_0, s_0]$ with zero jittering $s_0(y) \equiv 0$, and the final dejittered output.

Ideally, we would expect that the standard deviation $\text{std}(s_n)$ for each intermediate jittering estimation s_n is exactly σ_s. But computationally there is always some deviation, possibly due to the random number generator being used, the limited number of samples in digital implementation, and the Newton–Raphson algorithm itself. Therefore, we could enforce $\text{std}(s_n) = \sigma_s$ by introducing an extra normalization step. The quality of the output indeed improves in such circumstances (see Figure 5.2).

In Figure 5.2, from left to right are the initial noisy jittered image u_0, the image estimation after the fifth round $u^6 = \text{argmin } E[u \mid u_0, s_5]$, and the similar image estimation based on the normalized jittering $\hat{s}_5 = s_5/\text{std}(s_5) \times \sigma_s$. (Here std denotes the statistical standard deviation.) If one compares the dejittered edges, the normalization step clearly improves their qualities.

Figure 5.3 shows the output when the TV dejittering model is applied to another standard test image. It is well known in the literature of variational methods

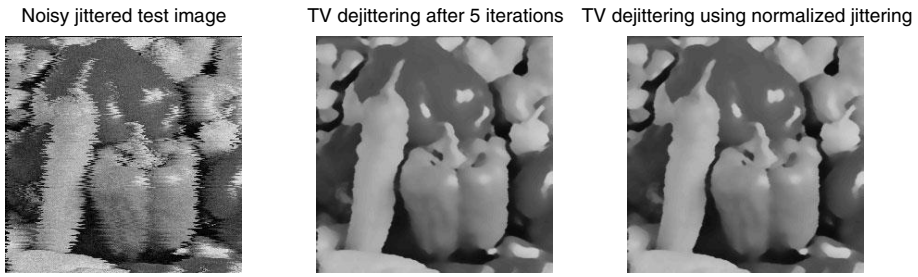


FIG. 5.2. *TV-based Bayesian dejittering with jittering normalization.*

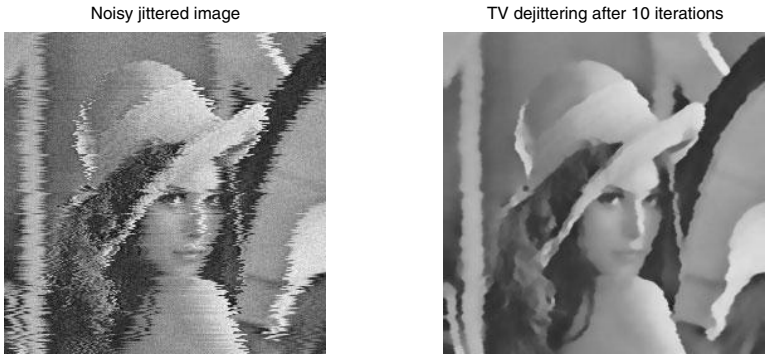


FIG. 5.3. *TV-based Bayesian dejittering applied to another standard test image. Pay attention to thin vertical structures and nonsmooth hair textures, to which the BV image model is only approximative.*

[3, 4, 9, 23] that the BV image model is only approximative in modeling textures like the hair region in the current image. Therefore, the smoothing effect (on highly oscillatory textures) in the dejittering output is intrinsic and unsurprising. Textures always impose great challenges in a variety of image processing tasks such as denoising, deblurring, inpainting, and segmentation [8, 16, 17, 23, 32, 33].

6. Conclusion. In this paper, based on the Bayesian rationale, we have proposed a novel variational model for video dejittering. The image model of bounded variations that we have applied, first introduced into image processing by Rudin and Osher [26] and Rudin, Osher, and Fatemi [27], is a powerful tool for restoring the regularity of randomly jittered objects.

We have studied the mathematical properties of the model based on the direct method of calculus of variations, the theory of functions with bounded variations, and various tools from analysis.

For the nonlinear and nonconvex energy functional, we have designed an algorithm based on alternately optimizing the image and jittering estimations. The algorithm is then numerically implemented by solving nonlinear PDEs (for image estimations) and by Newton–Raphson iterations. Typical numerical results are demonstrated.

This work again demonstrates the power of a good image model in image analysis. We expect that if the BV image model is replaced by the Mumford–Shah image model [25], many results should remain similar. (For example, such exchange has been very successful in image inpainting [13].)

If the random jitterings are correlated among different video frames, we expect that dynamic tools such as the Kalman filter [19, 30, 31] can play an important role in the process of modeling and computation.

Finally, it must be made clear that the current work only attempts to model some parts of the complex behavior of *real* jittering problems. For the sake of both enlightenment and inspiring future work, the author would like to quote an elegant paragraph from one of the referees (without editing):

> In real systems, jitter is almost always the combination of a random displacement and an underlying harmonic component. Any dejittering algorithm is going to have to rely on vertical image structure in order to realign the lines. All variants of single frame dejittering algorithms attempt to shift lines so as to increase vertical image smoothness. The heart of the dejittering problem is to tell the difference between vertical image features that are not straight and the harmonic component of jitter. This is extremely difficult to do, since there is (almost) no way to distinguish between dejittered images containing wavy but smooth vertical structures, and (those containing) straight and smooth structures. To add a further complication to the mix, the jitter is invariably caused by noise on the sync-tip signal. When that happens, not only is the line shifted, but the reference grey level is wrong, and then the actual DC line level changes from line to line.

As an intermediate stage between the current mathematical model and the realistic complexity elaborated in the preceding quotation, one could, for instance, study the scenario when the jittering s consists of two components:

$$s(y) = s_m(y) + s_n(y), \quad y \in (0, H),$$

where $s_m(y)$ denotes the *mean-field* or smooth jittering (with respect to the vertical variable y), and $s_n(y)$ still the Gaussian white noise. Then the prior model for jittering should be upgraded from (3.2) to

$$E[s] = E[s_m] + E[s_n] = \gamma \int_0^H |s_m'(y)|^p dy + \frac{1}{2\sigma_s^2 H} \int_0^H s_n^2(y) dy,$$

where γ denotes some suitable regularity weight, $p \geq 1$, and σ_s^2 the variance of the random component s_n. We leave further development along this line to interested readers.

Acknowledgments. The author would like to thank Gilbert Strang, Tony Chan, Stan Osher, and David Mumford for their invaluable teaching and inspirations, as well as the referees for their generous efforts in uplifting the quality of the current paper.

REFERENCES

[1] F. Andreu, V. Caselles, J. I. Diaz, and J. M. Mazon, *Some qualitative properties for the total variation flow*, J. Funct. Anal., 188 (2002), pp. 516–547.

[2] S. Armstrong, A. Kokaram, and P. J. W. Rayner, *Nonlinear interpolation of missing data using min-max functions*, in Proceedings of the IEEE International Conference on Nonlinear Signal and Image Processing, Mackinack Island, Boston, 1997.

[3] G. Aubert and L. Vese, *A variational method in image recovery*, SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[4] A. Chambolle and P. L. Lions, *Image recovery via total variational minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.

[5] T. F. Chan, S. H. Kang, and J. Shen, *Euler's elastica and curvature-based inpaintings*, SIAM J. Appl. Math., 63 (2002), pp. 564–592.

[6] T. F. Chan, S. Osher, and J. Shen, *The digital TV filter and nonlinear denoising*, IEEE Trans. Image Process., 10 (2001), pp. 231–241.
[7] T. Chan and J. Shen, *Variational restoration of nonflat image features: Models and algorithms*, SIAM J. Appl. Math., 61 (2000), p. 1338–1361.
[8] T. F. Chan and J. Shen, *Mathematical models for local nontexture inpaintings*, SIAM J. Appl. Math., 62 (2002), pp. 1019–1043.
[9] T. F. Chan and J. Shen, *On the role of the BV image model in image restoration*, in Recent Advances in Scientific Computing and Partial Differential Equations, Contemp. Math. 330, S. Y. Cheng, C.-W. Shu, and T. Tang, eds., AMS, Providence, RI, 2003, pp. 25–41.
[10] T. F. Chan, J. Shen, and L. Vese, *Variational PDE models in image processing*, Amer. Math. Soc. Notices, 50 (2003), pp. 14–26.
[11] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, *Deterministic edge-preserving regularization in computed imaging*, IEEE Trans. Image Process., 6 (1997), pp. 298–311.
[12] D. C. Dobson and C. R. Vogel, *Convergence of an iterative method for total variation denoising*, SIAM J. Numer. Anal., 34 (1997), pp. 1779–1791.
[13] S. Esedoglu and J. Shen, *Digital inpainting based on the Mumford–Shah–Euler image model*, European J. Appl. Math., 13 (2002), pp. 353–370.
[14] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Machine Intell., 6 (1984), pp. 721–741.
[15] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser Boston, Cambridge, MA, 1984.
[16] Y. Gousseau and J.-M. Morel, *Are natural images of bounded variation?*, SIAM J. Math. Anal., 33 (2001), pp. 634–648.
[17] H. Igehy and L. Pereira, *Image replacement through texture synthesis*, in Proceedings of the IEEE International Conference on Image Processing, Santa Barbara, CA, 1997, pp. 186–189.
[18] K. Ito and K. Kunisch, *An active set strategy based on the augmented Lagrangian formulation for image restoration*, Math. Model. Numer. Anal., 33 (1999), pp. 1–21.
[19] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Trans. ASME J. Basic Eng., 82 (1960), pp. 34–45.
[20] A. Kokaram, *Motion Picture Restoration*, Springer-Verlag, London, 1998.
[21] A. Kokaram and P. Rayner, *An algorithm for line registration of TV images based on a 2-D AR model*, in Proceedings of the Sixth European Signal Processing Conference on Theories and Applications, Brussels, Belgium, 1992, J. Vondewalle, R. Boite, M. Moonen, and A. Oosterlinck, eds., Elsevier, Amsterdam, 1992, pp. 1283–1286.
[22] A. Kokaram, P. M. B. Roosmalen, P. Rayner, and J. Biemond, *Line registration of jittered video*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, 1997, pp. 2553–2556.
[23] Y. Meyer, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*, Univ. Lecture Ser. 22, AMS, Providence, RI, 2001.
[24] D. Mumford and B. Gidas, *Stochastic models for generic images*, Quart. Appl. Math., 59 (2001), pp. 85–111.
[25] D. Mumford and J. Shah, *Optimal approximations by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.
[26] L. Rudin and S. Osher, *Total variation based image restoration with free local constraints*, in Proceedings of the 1st IEEE International Conference on Image Processing, Austin, TX, 1994, pp. 31–35.
[27] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[28] J. Shen, *Inpainting and the fundamental problem of image processing*, SIAM News, 36, June, 2003.
[29] J. Shen, *On the foundations of vision modeling* I. *Weber's law and Weberized TV restoration*, Phys. D, 175 (2003), pp. 241–251.
[30] G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, MA, 1993.
[31] G. Strang, *Block tridiagonal matrices and the Kalman filter*, in Wavelet Analysis: Twenty Years' Developments, Series in Analysis 1, World Scientific, River Edge, NJ, 2002, pp. 266–280.
[32] S. C. Zhu and D. Mumford, *Prior learning and Gibbs reaction-diffusion*, IEEE Trans. Pattern Anal. Machine Intell., 19 (1997), pp. 1236–1250.
[33] S. C. Zhu, Y. N. Wu, and D. Mumford, *Minimax entropy principle and its applications to texture modeling*, Neural Computation, 9 (1997), pp. 1627–1660.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1709–1736

THE MILNE PROBLEM FOR HIGH FIELD KINETIC EQUATIONS*

N. BEN ABDALLAH†, I. M. GAMBA‡, AND AXEL KLAR§

Abstract. Half space problems of the linear Boltzmann equation with a constant driving force are considered. Such problems model boundary layers between kinetic zones and fluid zones described by a high field limit of the Boltzmann equation. Existence, uniqueness, and asymptotic behavior of solutions are studied for positive and negative driving forces. In the positive case, the force field accelerates the particles, and we show that the solution of the half space problem is determined only by the inflow data. In contrast, for negative forces, the behavior at infinity has to be prescribed in order to insure uniqueness. Due to the nonvanishing forces, the problem does not possess any entropy. The existence and uniqueness issues are dealt with by supersolution techniques, while the asymptotic behavior is analyzed by semiexplicit integration of the equations along the characteristics. In the case of relaxation time approximation, a fast numerical method for computing the asymptotic state method is presented and tested.

Key words. kinetic theory of plasmas, nonequilibrium statistical mechanics, boundary layer problems in kinetic equations, numerical methods for kinetic equations

AMS subject classifications. 82B, 82C, 82D

DOI. 10.1137/S0036139902408898

1. Introduction.

Macroscopic fluid models are usually obtained from kinetic equations in collision dominated situations. Diffusion scalings are used when the equilibrium states (for which collisions are transparent) carry no current. Depending on the specific collision phenomena taken into account, asymptotics methods based on scaling assumptions lead to various diffusion models like the drift-diffusion [29], the energy-transport [8, 18], or the spherical harmonics expansion (SHE) model [32, 31, 4, 21, 17, 16].

When the driving forces are strong enough that their effect is of the same order of magnitude as collisional effects, another scaling, called high field scaling, has to be used. For the linear Boltzmann operator, the limit equation has been formally shown by Poupaud [30] to be a linear convection equation with the convective term depending on the force field. When the force field is the gradient of a potential coupled through a mean field approximation, a nonlinear system is obtained with a first order correction corresponding to augmented diffusion and transport [12, 13].

Kinetic high field models and associated macroscopic models have been considered in [13, 12, 34, 30, 5]. Recent comparisons between kinetic multiscale domain decomposition and the Monte Carlo method were presented in [20, 1, 10].

*Received by the editors June 3, 2002; accepted for publication (in revised form) November 17, 2003; published electronically July 2, 2004. The first and the third authors were supported by the T.M.R. network *Asymptotic Methods in Applied kinetic Theory* # ERB FMRXCT97 0157, run by the European Community. The first two authors were supported by the NSF-CNRS cooperation project entitled *"Modeling, Analysis and Simulation of Hybrid Quantum Models with Applications to Semiconductor Devices."* This work was also supported by the Texas Institute for Computational Engineering and Sciences/Austin.

http://www.siam.org/journals/siap/64-5/40889.html

†Mathématiques pour l'Industrie et la Physique, UMR CNRS 5640, Université Paul Sabatier, 31062 Toulouse, France (naoufel@mip.ups-tlse.fr).

‡Texas Institute for Computational and Applied Math and Department of Mathematics, University of Texas, Austin, TX 78712 (gamba@math.utexas.edu). This author was supported by the NSF under grant DMS 9971779 and by TARP under grant 003658-0459-1999.

§Fachbereich Mathematik, TU Darmstadt, D-64289 Darmstadt, Germany (klar@mathematik.tu-darmstadt.de).

However, up to now, no analysis of the kinetic boundary layer problem to find the correct boundary conditions for the fluid approximation has been performed. Such an analysis is also required if one wants to solve the matching problem for kinetic and macroscopic equations. Here, an interface region between the two equations has to be considered. The matching problem has to be solved, for example, for domain decomposition approaches simultaneously solving kinetic and macroscopic equations in different regions of the computational domain.

Boundary and interface regions are described by a transition layer where a stationary kinetic equation is solved. A standard assumption is that the layer has a slab symmetry, that is, the particle distribution is constant on surfaces parallel to the interface. Rigorous analysis of boundary value problems for linear transport kinetic equations in the absence of forces, known as the half space problem, and its corresponding limiting behavior in a strong collisional regime and long time scaling linear, as the length of the transition layer is comparable to the reference collision frequency, known as the Milne problem, was initiated [6] by means of spectral methods and semigroup theory.

For charged transport models, the force field gradient of the electrostatic potential is bounded along flat boundaries where the potential either is prescribed or is a solution of the corresponding mean field equation. In both cases, the force field will become a constant in the rescaled layer. In in the drift-diffusion regime due to weak force field forces, the rescaled force field vanishes. The corresponding half space and Milne problem was studied in [28], and computations for the corresponding fluid kinetic interface procedure for numerical implementations of hybrid methods were due to [35, 24].

For the case of strong force field regimes, one expects a slab symmetry whenever the curvature of the interface is small compared to the reciprocal of the mean free path and when the force field is normal to the interface. Consequently, the space coordinate reduces to, say, x the distance to the boundary or interface. After scaling it like $\frac{x}{\epsilon}$, where ϵ is the order magnitude of the mean free path, one has to solve a kinetic half space problem.

These strong force field scalings are characterized by nonstatistical equilibrium states $P = P(v)$; that is, they are $L^1_k(\mathcal{R}_v)$ space homogeneous solutions to the layer problem, with nonvanishing mean or first moment, which depend on the force field and on the Maxwellian in the kernel of the collision operator and the scattering function. This problem was treated by Trugman and Taylor [34] for the relaxation operator, and by Poupaud [30] for the general linear operator in three dimensions.

The first part of this paper treats the existence and uniqueness results for half space problems corresponding to strong force field scaling, both for positive and negative forces, and describes their corresponding asymptotic behavior, since from a practical point of view, the objects of great interest in obtaining boundary or matching conditions are the asymptotic states and the outgoing distribution (Albedo operator). The only assumption is that the boundary incoming data is a positive L^1_k bounded by a multiple of the state P.

In the case of positive forces, the force field accelerate the particles. Here we show that the solution of the half space problem is determined only by the inflow data. In fact, we prove that the unique solution $f(x, v)$ of the half space problem satisfies the condition that f_∞/P belongs to $L^\infty(\mathcal{R}^+_x \times \mathcal{R}_v)$. In addition $\lim_{x\to\infty} f(x, \cdot)/P$ converges to a proper factor n_∞. This factor is uniquely determined by the quotient of the mean of the solution, which is space independent, and the mean of the nonstatistical equilibrium state P, and thus it depends on the boundary data. This result indicates that under such strong forced scaling, the kinetic equation will admit an asymptotic

non-Maxwellian homogeneous stationary state; that is, under *strong acceleration*, the unscaled original kinetic solution should take a local stationary state which does not correspond to statistical equilibrium, whose asymptotic limit is a singularly perturbed augmented transport-diffusion converging to convective transport.

In contrast, for the case of strong negative forces, particles are slowed down, and under the same conditions for existence, one needs to prescribe the behavior of f at the right end of the layer in order to get uniqueness of the half space problem. In fact, we prove that for *any* given constant parameter n_∞, there exists a unique positive solution f_∞/P, belonging to $L^\infty(\mathcal{R}_x^+ \times \mathcal{R}_v)$, to the stationary problem in the rescaled layer, such that $\lim_{x\to\infty} f(x,\cdot)/P = n_\infty$. This essentially indicates that the behavior at infinity does not depend on the inflow boundary data.

A way to see the difference between the positive and negative force is that, in the former, the characteristic curves passing at $x = 0$ for the first order layer equation grow to $+\infty$ for $v > 0$ as $x \to \infty$, and come from $-\infty$ for $v < 0$. However, for the negative forced equation, the characteristic curves passing at $x = 0$ for $v > 0$ will turn back to intersect the axis $x = 0$ for $v < 0$. In particular for this second case, one may prescribe the behavior at infinity.

This anisotropic nature of the problem has as a consequence the lack of a natural entropy functional that controls the decay in space, such as it is possible to obtain in the low field scaling case. This motivates us to introduce new analytical methods based on comparison techniques by super- and subsolutions, namely, a maximum principle for solutions to kinetic stationary boundary value half space problems, basically introduced by Poupaud in [29] in order to treat boundary value problems for the stationary Vlasov–Maxwell system.

We recall that in the low field scaling case, the characteristic curves passing at $x = 0$ for the first order layer equation are all constant straight lines $v = v_o$ for all v_o, that is, all parallel to the x-axis. In particular, it has been shown that a corresponding boundary layer problem has a solution to the Milne problem given by an asymptotic behavior approaching a Maxwellian state, independent of forces [28]. In this case the boundary layer problem is similar to the one for a kinetic equation in the absence of forces, as treated in [6]. In both cases a diffusion limit arises, which may have a weak drift proportional to the field, corresponding to low field scaling.

In the second part of the paper, we describe a numerical procedure which computes n_∞, depending on the initial data, for the case of positive forces and a relaxation collision operator. It uses a classical Chapman–Enskog–type expansion to approximate the solution. We obtain a force field modified Marshak condition, which is a higher order correction to prescription of incoming fluxes. Our calculation recovers the classical Marshak condition for diffusion approximations as the force fields tends to zero. The method is seen to converge very fast numerically. It seems to give accurate results when compared to the available explicit solutions in some special cases. For approaches to the numerical solution of the standard half space problem in gas dynamics and semiconductor equations, we refer the reader to [2, 14, 22, 33], and for a mathematical investigation, to [3, 15, 23]. We expect a future implementation of very efficient hybrid computational schemes that will be able to link nonstatistical equilibrium scales by their anisotropic diffusion convective limits, as well as to solve the coupling of convective regions to diffusion regions by transition layer or interfaces, as is steadily observed in strongly doped device simulation under hot-electron regimes.

The paper is organized as follows. In section 2 we present the strong force field equations. Section 3 contains an analytical investigation of the half space problem for both the negative and positive forces. In both cases, existence and uniqueness results

with the asymptotic behavior at infinity are investigated. In section 4 the numerical procedure and some numerical results are presented in the case of relaxation operators.

2. High field kinetic equations. The drift-collision balance regime. We consider the semiclassical linear Boltzmann equation in dimensionless variables for an electron gas for a semiconducting material in the parabolic band approximation, with a strong force field scaling

$$\eta\partial_t f + v\cdot\nabla_z f + \frac{\eta}{\epsilon}E(z,t)\cdot\nabla_v f = \frac{1}{\epsilon}Q(f) \tag{2.1}$$

with $z, v \in \mathcal{R}^3$. The general linear collision operator under consideration is

$$Q(f) = \int s(v,v')[M(v)f(v') - M(v')f(v)]\,dv' = Q^+(f) - \sigma(v)f. \tag{2.2}$$

The scattering function $s(v,v')$ is symmetric and satisfies

$$0 < s_0 \le s(v,v') \le s_1 < +\infty \quad \text{and} \quad s(v,v') = s(v',v), \tag{2.3}$$

and σ denotes the collision frequency

$$\sigma(v) = \int s(v,v')M(v')\,dv', \tag{2.4}$$

whereas

$$Q^+(f) = \int s(v,v')f(v')M(v)\,dv' \tag{2.5}$$

is the gain operator. Throughout this paper the notation $\langle h\rangle$ stands for $\int h(v)\,dv$.

Here we use the standard notation $M(v)$ for the centered, reduced Maxwellian $M = (2\pi)^{-\frac{3}{2}}\exp(-\frac{v^2}{2})$.

As a motivation for this problem, we look at semiconductor modeling as the main example of a transport phenomenon that exhibits stationary nonequilibrium statistical states. Usually, the vector field $E = E(z,t) = -\nabla_z\Phi$ denotes the scaled electric force, which is determined by a Poisson equation for the potential Φ:

$$\nabla_z\cdot(\nabla_z\Phi) = \gamma\left(\frac{1}{\eta^d}\int_{\mathcal{R}^3} f dv - C(z)\right),$$

where γ is the inverse to the scaled Debye length of the device and $C(z)$, which denotes the ion background, is bounded, measurable, and largely varying. It is worth mentioning that strong force field scalings are present due to space inhomogeneities, such as short base channel devices, under strong forward bias that produces a region of positive charges inside the channel (i.e., $\gamma^{-1} \gg 0$). Such an effect is known as hot-electron transport. Under these assumptions on $C(z)$, classical potential theory implies that the solution of the Poisson equation in a bounded channel-like region yields a continuous bounded force field $E(z)$.

Because of this effect, assume dimensionless parameters η and γ both of order $O(1)$ and that ϵ, the scaled mean free path, is small. This scaling assumption corresponds to the drift-collision balance scaling introduced in [30, 13]. Such a scaling is realized, for instance, in the modeling of silicon doped diodes with $0.4\,\mu$m channel [20, 1, 10] under potential bias of 1eV. These simulations exhibit the formation of transition

layers in the drain junction, with a clear jump from a close to convective state in the channel region to a diffusion equilibrium at the contact. In addition, inside the channel, there is a clear region where the numerical probability distribution function, the solution to the approximated kinetic Poisson system, takes a definite state away from statistical equilibrium. Such a configuration corresponds to a relatively strong forced field scale with respect to collisions against a background. This may be the case for other collisional plasma physics applications under strong force fields.

The problem we want to study is related to the solution and its asymptotic behavior in a given layer of length ϵ. This layer is inversely proportional to the drift-collision scale associated with the reciprocal of the scaled mean free path of $o(\epsilon)$ for the kinetic problem under such a regime.

From now we focus on the problem of having a force field $E(x,t)$ given in a transition layer or boundary with a slab symmetry; that is, the particle distribution is constant on surfaces parallel to the interface. For the case of strong force field regimes, one expects such a slab symmetry whenever the curvature of the interface is small compared to the reciprocal of the mean free path and when the force field is normal to the interface.

In order to obtain the boundary or interface layer equations in a slab geometry, fix a point $\hat{z}$ on the boundary, assume that the electric force is orthogonal to the interface, and rescale as usual the space coordinate in the layer normal to the boundary with the mean free path ϵ, introducing the new coordinate x orthogonal to the boundary:

$$x = \frac{(\hat{z} - z) \cdot n}{\epsilon}.$$

Here, n denotes the outer normal to the boundary or interface. This transformation yields the new coordinates $(x, \hat{z})$ instead of z in the slab layer. To $O(1)$ one obtains, after applying the transformation to (2.1),

$$v \cdot n \partial_x f + \eta E \cdot \nabla_v f, = Q(f),$$

where, as $\epsilon \to 0$, the variable $x \in [0, \infty)$, and the field $E = E(x = 0, \hat{z}, t)$ does not depend on x and thus is constant. This problem has to be supplied with the ingoing function at the boundary, i.e., at $x = 0$; that is, $f(0, v)$, $v \cdot n > 0$, with n the outer normal to the boundary at $x = 0$. In order to have the force field E constant it is enough that the potential Φ is regular enough so that $\nabla\Phi$ is bounded at the slab boundary.

To simplify the problem, we assume from now on that the z_1-coordinate points in the direction of the normal, so that that $E = (E_1, 0, 0)$ and that $\tau = 1$, $\eta = 1$. Then the above reduces to the following one-dimensional problem:

$$v_1 \partial_{z_1} f + E_1 \partial_{v_1} f = Q^+(f) - \sigma(v) f,$$

with $x \in [0, \infty)$, $v_1 \in \mathcal{R}$, $f = f(z_1, v_1)$. Then $M(v)$ in the definition of Q^+ and σ reduces to the one-dimensional Maxwellian $M(v_1) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{v_1^2}{2})$ for $v_1 \in \mathcal{R}$.

From now on, we use the notation $(x, v) \in \mathcal{R}^+ \times \mathcal{R}$ rather than (z_1, v_1). The Milne problem takes the following form:

$$\begin{aligned} v \partial_x f + E \partial_v f = Q(f) = Q^+(f) - \sigma(v) f, \\ \varphi(0, v) = k(v), \quad v > 0, \end{aligned} \tag{2.6}$$

with $x \in [0, \infty)$, $v \in \mathcal{R}$, and an ingoing positive function k satisfying the conditions stated below.

Before announcing the main theorems, we define the homogeneous solution $P_{\sigma,E}(v)$ as the unique function which satisfies

$$E\partial_v P = Q(P) = Q^+(P) - \sigma(v)P, \qquad \text{with } \langle P\rangle = 1, \tag{2.7}$$

using the notation

$$\langle P\rangle = \int P(v)dv. \tag{2.8}$$

In addition, for any integrable solution f of (2.6),

$$j = \langle vf\rangle \quad \text{is } x\text{-independent.} \tag{2.9}$$

The proof of this statement is trivial for any integrable solution.

Solvability of problem (2.7) in $L^\infty \cap L^1$ can be found in Trugman and Taylor [34] for the relaxation-type operator in one dimension. It has also been discussed in Frosali, Van der Mee, and Paveri-Fontana [19]. The most general result has been obtained by Poupaud [30], who finds solutions to (2.7) in L^1 for general linear collision operators in higher dimensions, depending on the integrability of the collision frequency. In addition he shows that the solution function P is unique and positive. Recently, this result has been generalized to the collision operators with Pauli-exclusion terms [7].

For completeness we recall the Poupaud solution representation to problem (2.7), obtained via spectral analysis [30] of the following linear integral operator:

$$\begin{aligned} P_E(v) &= L_E(Q^+(P))(v) \\ &= \int_0^\infty \exp\left(-\int_0^\tau \sigma(v-\mu E)d\,\mu\right)\int_{\mathcal{R}} s(v-\tau E,w)P_E(w)d\,w\,M(v-\tau E)d\,\tau \end{aligned} \tag{2.10}$$

for $E \neq 0$ such that $\langle P\rangle = 1$. The operator $L_E : L^1 \to L^1_\sigma$ is the inversion operator to $E\cdot\nabla_v + \sigma(v)$, defined by

$$L_E(f)(v) = \int_0^\infty \exp\left(-\int_0^\tau \sigma(v-\mu E)d\,\mu\right) f(v-\tau E)d\,\tau. \tag{2.11}$$

Poupaud proves that the integral equation (2.10) has a unique integrable (L^1) positive solution if and only if

$$\int_0^\infty \sigma(v+\mu E)d\,\mu = +\infty \qquad \text{a.e.} \tag{2.12}$$

In addition, the solution satisfies the property

$$P_E(v) = P_{-E}(-v). \tag{2.13}$$

It is clear that in our case, by (2.3), the scattering function $s(v,v')$ is bounded above and below by positive constants, so that the collision frequency $\sigma(v)$ function as defined in (2.4) satisfies the infinite integrability compatibility condition (2.12).

Moreover, the unique solution P to problem (2.7) has all moments bounded. Indeed by (2.2) and (2.3) the following moment recursion inequality holds:

$$\langle v^k P\rangle \le \frac{s_1}{s_0}\langle v^k M\rangle + E\langle v^{k-1}P\rangle. \tag{2.14}$$

In the particular case of a relaxation collision operator, when $s(v', v) = \tau^{-1}$,

$$Q(f)(v) = \frac{1}{\tau}\left(M(v)\int_{\mathcal{R}^3} f(v')dv' - f(v)\right) = \frac{\langle f\rangle M - f}{\tau}, \tag{2.15}$$

with τ the relaxation time, one obtains an explicit formula for the dominant P state, as the right-hand side of (2.10) is computable. Setting $\tau = 1$, without loss of generality, the probability distribution function P, a solution to (2.7) with the collisional form (2.15), is explicitly given, as originally computed in [19], by

$$P_E(v) = \frac{1}{2E}\exp\left(\frac{-\lambda}{E}\right) erfc\left(\lambda\sqrt{\frac{1}{2}}\right) \tag{2.16}$$

with $E > 0$, $\lambda = \frac{2}{E} - v$, and $erfc(x) = \frac{1}{2} + \frac{1}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$.

In addition, moments are explicitly computed by a recursion formula [13], and, in the one-dimensional case, the first three satisfy

$$\begin{aligned}\langle vP_E\rangle &= E,\\ \langle v^2P_E\rangle &= 1 + 2E^2,\\ \langle v^3P_E\rangle &= 3E + 6E^3.\end{aligned}$$

The main result for the first part of the paper, *the Milne problem for strong force fields*, is stated as follows.

THEOREM 1 (positive force field). *Let $E > 0$ be a given positive real number. Let $P_{\sigma,E}$ (we shall also use the short notation P) be the solution of the space homogeneous equation* (2.7). *Assume that $0 \le k(v) \le KP(v)$ for some constant K. Then,* (2.6) *has a unique positive solution such that $f/P \in L^\infty(\mathcal{R}_x^+ \times \mathcal{R}_v)$. Moreover, there exists a constant $n_\infty = \frac{\langle vf\rangle}{\langle vP\rangle}$ such that*

$$\lim_{x\to+\infty} f(x, v) = n_\infty P(v) \quad \textit{pointwise.} \tag{2.17}$$

THEOREM 2 (negative force field). *Let $E < 0$ be a given negative real number. Let $P_{\sigma,E}$ (we shall also use the short notation P) be the solution of the space homogeneous equation* (2.7). *Assume that $0 \le k(v) \le KP(v)$ for some constant K. Then, for any given $n_\infty \in \mathcal{R}^+$, there exists a unique positive solution f_{n_∞} of* (2.6) *such that $f_{n_\infty}/P \in L^\infty(\mathcal{R}_x^+ \times \mathcal{R}_v)$ and*

$$\lim_{x\to+\infty} f_{n_\infty}(x, v) = n_\infty P(v) \quad \textit{pointwise.}$$

In both cases the integrability of f follows from the integrability of P.

3. Analysis of the Milne problem.

3.1. Properties independent of $sgn(E)$. We first start by showing that the current carried by the homogeneous solution P, that is $\langle vP\rangle$, has the same sign as E. Namely, we claim the following.

LEMMA 1. *The solution P of problem* (2.7) *with the linear collisional form* (2.2) *satisfies $E\langle vP\rangle > 0$ and $0 < E\langle v^3\, P\rangle < K < \infty$ if $E \neq 0$.*

Proof. In the case of relaxation the statement is trivial since, by (2.17), $E\langle vP\rangle = E^2$ and $E\langle v^3\, P\rangle = E^2 + 3E^4$.

For the general linear case, the collision operator is self-adjoint in the weighted space $L^2_M = \{f \in L^1_{loc},\ \int_{\mathcal{R}} f^2 M^{-1}\, dv < +\infty\}$, so that

$$\langle gQ(f)M^{-1}\rangle = \langle fQ(g)M^{-1}\rangle. \tag{3.1}$$

Since, by symmetrization,

$$\int Q(f)g\, dv = -\frac{1}{2}\int sMM'\left(\frac{f'}{M'} - \frac{f}{M}\right)(g' - g)\, dv\, dv, \tag{3.2}$$

it follows that for all monotone increasing H

$$\int Q(f)H\left(\frac{f}{M}\right) dv \le 0 \tag{3.3}$$

and

$$\int Q(f)H\left(\frac{f}{M}\right) dv = 0 \quad \text{if and only if} \quad f(v) = cM(v)$$

for any constant c.

Now, taking $H(\tau) = \ln \tau$, we obtain

$$\int Q(P)\ln\left(\frac{P}{M}\right) dv = \int Q(P)\left(\ln P + \frac{v^2}{2}\right) dv \le 0. \tag{3.4}$$

In addition, by (2.7), $E\frac{\partial P}{\partial v} = Q(P)$; then

$$0 \ge \int Q(P)\left(\ln P + \frac{v^2}{2}\right) dv = \int E\frac{\partial P}{\partial v}\left(\ln P + \frac{v^2}{2}\right) dv.$$

Since by integrability of P and $P\ln P$ the identity

$$\int \ln P \frac{\partial P}{\partial v} dv = \int_R \frac{\partial}{\partial v}(P\ln P - P) dv = 0$$

holds, then

$$E\int \frac{v^2}{2}\ \frac{\partial P}{\partial v}\, dv \le 0.$$

Thus, integrating by parts yields the first inequality

$$E\int vP\, dv \ge 0. \tag{3.5}$$

Next, we show that (3.5) cannot be zero if E is not zero. Indeed, if $E\int vP\, dv = 0$, then, by (3.4), $\int Q(P)\ln\frac{P}{M}\, dv = 0$, which implies $P = c\ M$, and thus P is a multiple of the Maxwellian. Therefore $E\frac{\partial M}{\partial v} = 0$ and $E \neq 0$, which is a contradiction since $\partial M/\partial v$ does not vanish.

Finally, the finiteness of moments for all orders follows from the moments recursion formula (2.14). □

THEOREM 3 (existence). *Let E be a given real number. Let P be the solution of the space homogeneous equation* (2.7). *Assume that $K_1P(v) \le k(v) \le K_2P(v)$ for*

some positive constants K_1 and K_2. Then there exist two solutions $(\underline{f}, \overline{f})$ of (2.6) called minimal and maximal solutions such that

$$K_1 P(v) \le \underline{f}(x,v) \le \overline{f}(x,v) \le K_2 P(x,v)$$

and such that any solution f of (2.6), such that $K_1 P(v) \le f(x,v) \le K_2 P(v)$, is trapped between $\underline{f}$ and $\overline{f}$:

$$\underline{f}(x,v) \le f(x,v) \le \overline{f}(x,v).$$

To construct a solution on the half real line $\mathcal{R}^+$, we first solve the problem on the interval $[0, L]$ and then let L tend to $+\infty$. To this end, we consider the problem

$$\begin{cases} v\varphi_x + E\varphi_v = Q(\varphi)(x,v), \\ \varphi(0,v) = k_1(v) \text{ for } v > 0, \\ \varphi(L,v) = k_2(v) \text{ for } v < 0. \end{cases} \tag{3.6}$$

LEMMA 2. *Assume that $K_1 P(v) \le k_{1,2}(v) \le K_2 P(v)$. Then (3.6) admits a unique solution φ such that $\varphi(x,v)/P(v) \in L^\infty([0,L] \times \mathcal{R}_v)$. Moreover,*

$$K_1 P(v) \le \varphi(x,v) \le K_2 P(x,v).$$

Proof. To prove the existence of a solution, we consider the mapping T_L defined by $f = T_L(g)$, where f is the unique solution of

$$\begin{cases} \sigma(v) f + v f_x + E f_v = Q^+(g)(x,v), \\ f(0,v) = k_1(v) \text{ for } v > 0, \\ f(L,v) = k_2(v) \text{ for } v < 0. \end{cases} \tag{3.7}$$

The function f exists by virtue of [29] and is unique since $\sigma \ge s_0 > 0$. Moreover, the maximum principle insures that $f \ge K_1 P(v)$ if $g \ge K_1 P(v)$ (and $f \le K_2 P(v)$ if $g \le K_2 P(v)$). Starting from $f_1(x,v) = K_1 P(v)$, we proceed as in [29] and define $f^n = T_L f^{n-1}$ and set $\varphi = \lim_{n \to +\infty} \frac{1}{n} \sum_{l=1}^n f_l$. It is then clear that $K_1 P \le f_n \le K_2 P$ and φ is a solution of (3.7) which satisfies $K_1 P \le \varphi \le K_2 P$.

The uniqueness follows by an entropy argument developed in [9]. For the sake of completeness, we detail this argument. We set h to be the difference between two solutions. Then h is a solution of

$$\begin{cases} v h_x + E h_v = Q(h), \\ h(0,v) = 0 \text{ for } v > 0, \\ h(L,v) = 0 \text{ for } v < 0. \end{cases}$$

Using the inequality $\int Q(h) sgn(h)\, dv \le 0$ and the fact that equality holds if and only if the sign of $h(x,v)$ does not depend on v, we obtain

$$\int_0^{+\infty} v|h(L,v)|\, dv - \int_{-\infty}^0 v|h(0,v)|\, dv = \int_0^L \int_{\mathcal{R}} Q(h) sgn(h)\, dv\, dx,$$

which implies that $h(0,v) = 0$ and $h(L,v) = 0$ for $v \in \mathcal{R}$ and that the sign of $h(x,v)$ does not depend on v. Setting $H = |h|$, since the collisional form is linear, then $Q^+(h) sgn(h) = Q^+(H)$. Therefore,

$$\begin{cases} \sigma(v) H + v H_x + E H_v = Q^+(H), \\ H(0,v) = 0 \text{ for } v \in \mathcal{R}, \\ H(L,v) = 0 \text{ for } v \in \mathcal{R}. \end{cases}$$

This implies that $H = 0$ after an integration along the characteristics.

Proof of existence Theorem 3. The maximal and minimal solutions are respectively obtained by solving problem (3.6), with $k_1 = k$ and $k_2 = K_2P$ (for the maximal solution) and $k_1 = k$ and $k_2 = K_1P$ (for the minimal solution). Indeed, define $f^+ = T_L^+(g)$ and $f^- = T_L^-(g)$ as the unique solutions of

$$\begin{cases} \sigma(v)f^+ + vf_x^+ + Ef_v^+ = Q^+(g)(x,v), \\ f^+(0,v) = k(v) \text{ for } v > 0, \\ f^+(L,v) = K_2P(v) \text{ for } v < 0, \end{cases} \tag{3.8}$$

$$\begin{cases} \sigma(v)f^- + vf_x^- + Ef_v^- = Q^+(g)(x,v), \\ f^-(0,v) = k(v) \text{ for } v > 0, \\ f^-(L,v) = K_1P(v) \text{ for } v < 0. \end{cases} \tag{3.9}$$

Then the maximal and minimal solutions are defined by $\overline{f}_L = T_L^+(\overline{f}_L)$ and $\underline{f}_L = T_L^-(\underline{f}_L)$, and so

$$K_1P(v) \le \underline{f}_L \le \overline{f}_L \le K_2P(v).$$

Moreover, $\overline{f}_L = \lim_{n\to+\infty}(T_L^+)^n(K_2P)$ (the sequence $((T_L^+)^n(K_2P))$ being pointwise decreasing) and $\underline{f}_L = \lim_{n\to+\infty}(T_L^-)^n(K_1P)$ (the sequence being pointwise increasing). The above constructed sequences satisfy the following monotonicity properties.

LEMMA 3. *If $L_1 \le L_2$, then $\overline{f}_{L_1} \ge \overline{f}_{L_2}$ and $\underline{f}_{L_1} \le \underline{f}_{L_2}$ on $[0, L_1] \times \mathcal{R}$.*

Proof. Let $L_1 \le L_2$ and $H = \overline{f}_{L_1} - \overline{f}_{L_2}$ on $[0, L_1] \times \mathcal{R}_v$; then H^m is the solution of

$$\begin{cases} vH_x + EH_v = Q(H), \\ H(0,v) = 0 \text{ for } v > 0, \\ H(L_1,v) = K_2P(v) - \overline{f}_{L_2}(L_1,v) \ge 0 \quad \text{for } v > 0. \end{cases}$$

Therefore $H \ge 0$ by virtue of Lemma 2. The inequality for $\underline{f}_L$ is obtained analogously. ∎

Let us now pass to the limit $L \to +\infty$. For this purpose, we notice that $K_1P(v) \le \underline{f}_L \le \overline{f}_L \le K_2P(v)$ and that $\underline{f}_L$ is increasing with respect to L, while $\overline{f}_L$ is decreasing with respect to L. The pointwise limits $\underline{f}$ and $\overline{f}$ of $\underline{f}_L$ and $\overline{f}_L$ as L tends to $+\infty$ are obviously solutions of the problem (2.6) and satisfy

$$K_1P \le \underline{f} \le \overline{f} \le K_2P.$$

The only thing left to show now is that any solution $f \in [K_1P, K_2P]$ of (2.6) is trapped between $\underline{f}$ and $\overline{f}$. To this aim, set $g = f - T_L^-(f)$. Then g is the solution of

$$\begin{cases} vg_x + Eg_v + \sigma g = 0, & \\ g(0,v) = 0, & v > 0, \\ g(L,v) = f(L,v) - K_1P(v) \ge 0, & v < 0, \end{cases}$$

which implies $g \ge 0$. Hence $T_L^-(f) \le f$. The maximum principle insures that $T_L^\pm(g_1) \le T_L^\pm(g_2)$ whenever $g_1 \le g_2$. Therefore $(T_L^-)^m(f) \le f$. However, $(T_L^-)^m(f) \ge (T_L^-)^m(K_1P)$. Since $\underline{f}_L = \lim_{m\to+\infty}(T_L^-)^m(K_1P)$, we deduce from the above inequality that $f \ge \underline{f}_L$ on $[0, L] \times \mathcal{R}$, which leads to $f \ge \underline{f}$. The inequality $f \le \overline{f}$ is obtained analogously. The proof of Theorem 3 is now complete. ∎

Next, we study the uniqueness and the asymptotic behavior for the solutions.

3.2. The Milne problem for strong positive forces. The aim of this section is to complete the proof of Theorem 1. First, we show *uniqueness*, that is, $\overline{f} = \underline{f}$ for arbitrary K_2. This proof, which is rather short, uses the asymptotic behavior to be shown next. However, we leave the asymptotic behavior for last, since its proof does not require uniqueness of the solutions.

THEOREM 4 (uniqueness for the case of strong positive forces). *Assume that* $E > 0$ *and that* $0 \le k \le KP$. *Then* $\overline{f}$ *and* $\underline{f}$ *coincide.*

First we prove the following proposition.

PROPOSITION 1. *Let* $h = \overline{f} - \underline{f}$. *Then* $\partial_x h \ge 0$, *and there exists* $\alpha \ge 0$ *such that*

$$\lim_{x \to +\infty} h(x, v) = \alpha P(v).$$

Proof. Take the function $h_{L,a}(x,v) = \overline{f}_{L+a}(x+a,v) - \underline{f}_{L+a}(x+a,v) - \overline{f}_L(x,v) + \underline{f}_L(x,v)$ for $a > 0$. Then $h_{L,a}$ satisfies

$$v\partial_x h_{L,a} + E\partial_v h_{L,a} = Q(h_{L,a})$$

and $h_{L,a}(0,v) \ge 0$ for $v > 0$, while $h_{L,a}(L,v) = 0$ for $v < 0$. Therefore $h_{L,a} \ge 0$ uniformly in L, which implies by passing to the limit $L \to +\infty$ that $h(x+a,v) - h(x,v) \ge 0$. Since a is arbitrary, then $\partial_x h \ge 0$. ∎

Proof of Theorem 4. Let $h = \overline{f} - \underline{f}$. By construction, $h \ge 0$.

However, because of the boundary condition at $x = 0$, $h(0,v) = 0$ for $v > 0$, and consequently the associated first moment j, which is x-independent, is nonpositive since $j = \langle vh \rangle = \int_{-\infty}^{0} vh \le 0$.

On the other hand, by Proposition 1, on the one hand, $\lim_{x \to +\infty} h(x,v) = \alpha P(v)$ for $\alpha \ge 0$, and on the other hand, by Lemma 1, $\alpha \langle vP \rangle \ge 0$.

Now, since $j = \langle vh \rangle$ is x-independent due to is limit at infinity, $0 \ge j = \alpha \langle vP \rangle \ge 0$. This is only possible if $\alpha = 0$.

Therefore, h is nonnegative, and by Proposition 1, increasing with respect to x and tends to zero as x tends to $+\infty$. Then h is identically equal to zero for all $x \ge 0$, for all v. ∎

Asymptotic behavior at ∞: Completion of the proof of Theorem 1. Without loss of generality, we renormalize the solution of (2.6) with respect to the constant K of the data. This is equivalent to treating the case $K = 1$. Therefore f solves

$$\begin{cases} \dfrac{v}{E} f_x + f_v + \dfrac{\sigma(v)}{E} f = \dfrac{1}{E} Q^+(f), \\ f(0,v) = k(v) \le P(v). \end{cases} \tag{3.10}$$

By (3.10), $0 \le f(x,v) \le P(v)$ for all v, and its first moment $j = \int_{\mathcal{R}} v f(x,v)\, dv$ is independent of x.

The strategy of the proof works as follows. First we shall prove a *key statement* in Theorem 5, which shows that if the first moment of the solution f of (3.10) is a proper fraction of the first moment of the homogeneous solution P, say by a factor $0 \le \lambda < 1$, then the spatial asymptotic behavior of f at infinity is given by exactly λP, which is the expected behavior for any solution of the initial value problem (3.10) at infinity. This result is equivalent to an a priori estimate, which means control on the spatial variation of the solutions by control on the variation of its first moment.

Second, we shall see that, in fact, the first moment of *any* solution to problem (3.10) is always a proper fraction of the first moment of the homogeneous solution of the problem; that is, $\lambda P(v)$ for some $0 \le \lambda < 1$.

Combining both results means that the spatial asymptotic behavior at infinite for f is actually $\lambda P(v)$. That is, the *quotient* between the first moments of the solution f and the homogeneous solution P and between the spatial asymptotic behavior solution of f and the homogeneous solution P, to problems (3.10) and (2.7), respectively, are both the same. In a sense, this is like a Harnack inequality for the kinetic problem.

In fact, these key estimates follow from Lemma 1, which states that if E is positive, then the first moment of the homogeneous solution P is positive. Thus we can make sense of a proper fraction of the first moment of the homogeneous state for a strong force scaling as well as all estimates that follow.

THEOREM 5. *If* $j = \int vf\,dv = \lambda \int vP \geq 0$, *with* $\lambda \in [0,1)$, *then*

$$\lim_{x\to+\infty} f(x,v) = \lambda P(v) = \frac{j}{\langle vP\rangle} P(v). \tag{3.11}$$

The proof of this theorem requires additional partial results that we write as lemmas and corollaries.

LEMMA 4 (initial control for the gain operator). *Assume* $\int vf\,dv = \langle vf\rangle = \lambda\langle vP\rangle > 0$ *for* $0 \leq \lambda < 1$. *Then*

$$Q^+(f) \leq \mu_0 Q^+(P) \quad \text{for all } x \geq 0, \tag{3.12}$$

where

$$0 < \mu_0 = 1 - \frac{s_0}{s_1}\frac{1-\lambda}{2v_0}\langle vP\rangle < 1, \tag{3.13}$$

where v_0 *satisfies*

$$0 < \int_{v_0}^{\infty} vP\,dv \leq \frac{1-\lambda}{2}\langle vP\rangle\,, \tag{3.14}$$

and the quotient $\frac{s_0}{s_1}$, *as defined in* (2.3), *measures the oscillation of the scattering rate function.*

Proof. From the existence result, it follows that $0 \leq f \leq P$. Since Q^+ is a positive linear operator,

$$Q^+(f)(v) = Q^+(P) - Q^+(P-f) \leq Q^+(P) - M(v)\int_0^{+\infty} s(v',v)\,(P-f)\,dv'. \tag{3.15}$$

Then, it is enough to prove that

$$M(v)\int_0^{+\infty} s(v',v)\,(P-f)\,dv' > \beta\; Q^+(P)(v) \quad \text{for some } \beta < 1. \tag{3.16}$$

In order to estimate (3.16), we use the hypothesis on the first moment of f and P, that is, if

$$\langle vf\rangle = \int_{-\infty}^{+\infty} vf\,dv = \lambda\int_{-\infty}^{+\infty} vP\,dv = \lambda\langle vP\rangle,$$

which yields the first moment flux estimate

$$\int_0^{+\infty} vf\,dv = -\int_{-\infty}^{0} vf\,dv + \lambda\langle vP\rangle \leq -\int_{-\infty}^{0} vP\,dv + \lambda\langle vP\rangle;$$

that subtracted from

$$\int_0^\infty vP = -\int_{-\infty}^0 vP + \langle vP\rangle$$

leads to the lower bound estimate for the first moment fraction of the difference between the stationary and homogeneous solution

$$\int_0^{+\infty} v(P-f)\,dv \geq (1-\lambda)\langle vP\rangle = \alpha. \tag{3.17}$$

The integrability of the first moment of the homogeneous solution P, and the fact that $\lambda < 1$, imply that there exists a $v_0 > 0$ such that

$$\int_{v_0}^{+\infty} vP\,dv \leq \left(\frac{1-\lambda}{2}\right)\langle vP\rangle = \frac{\alpha}{2}\,, \tag{3.18}$$

so that, since $vf > 0$ for $v \geq v_0 > 0$, also

$$\int_{v_0}^{+\infty} v(P-f)\,dv \leq \frac{\alpha}{2}\,. \tag{3.19}$$

Next, subtracting inequality (3.19) from inequality (3.17) leads to

$$\int_0^{v_0} v(P-f)\,dv \geq \frac{1-\lambda}{2}\langle vP\rangle. \tag{3.20}$$

Now, recalling that the scattering rate function $s = s(v', v)$ is bounded by $0 < s_0 \leq s(v', v) \leq s_1 < \infty$, multiplying and dividing the integrand by $s = s(v', v)$ yield a first moment flux fraction difference estimate by a fraction difference for the gain operator

$$\int_0^{v_0} v(P-f) = \int_0^{v_0} v\frac{s}{s}(P-f)\,dv \leq \frac{v_0}{s_0}\int_0^{v_0} s(v', v)(P-f)\,dv, \tag{3.21}$$

which combined with inequality (3.20) yields the following lower bound for the right-hand side of (3.21):

$$\int_0^{v_0} s(v', v)(P-f)\,dv \geq \frac{(1-\lambda)}{2v_0}s_0\langle vP\rangle = \frac{\alpha}{2}\frac{s_0}{v_0}\,. \tag{3.22}$$

In addition, since $\langle P\rangle = 1$, then $s_1 \geq \int_{-\infty}^{+\infty} s(v', v)P(v)\,dv \geq s_0$. Thus the right-hand side of (3.22) can be estimated as

$$\frac{\alpha}{2}\frac{s_0}{v_0} \geq \frac{\alpha}{2v_0}\frac{s_0}{s_1}\int_{-\infty}^{+\infty} s(v', v)P(v')\,dv' \;= \frac{\alpha}{2v_0}\frac{s_0}{s_1}Q^+(P), \tag{3.23}$$

where the fraction $\frac{s_0}{s_1} < 1$.

Since $P - f > 0$, inequalities (3.22) and (3.23) lead to

$$M(v)\int_0^{+\infty} s(v', v)(P-f)\,dv \geq M(v)\int_0^{v_0} s(v', v)(P-f)\,dv > \frac{\alpha}{2v_0}\frac{s_0}{s_1}Q^+(P)(v), \tag{3.24}$$

which yields the inequality (3.16) with $\beta = \frac{\alpha}{2v_0}\frac{s_0}{s_1}$.

Therefore (3.12) holds with

$$0 < \mu_0 = 1 - \frac{\alpha}{2v_0}\frac{s_0}{s_1} < 1, \qquad \alpha = (1-\lambda)\langle vP\rangle\,, \tag{3.25}$$

where v_0 is such that $\int_{v_0}^{+\infty} vP\,dv \le \frac{\alpha}{2}$. ∎

Remark. The choice of v_0 actually depends on the fact that the first moment of P is finite, that is, on the integrability properties of homogeneous solution P and its corresponding behavior at infinity, and not necessarily on the explicit form of P. This implies that these results can be extended to more general cases, as long as the first moment of P is strictly positive and the corresponding collision frequency is bounded below by a strictly positive constant and above by infinity.

LEMMA 5 (local control of f). *Let $x_k > 0$ such that*

$$Q^+(f) \le \mu_k Q^+(P) \tag{3.26}$$

for any $(x,v) \in D_k = \{(x,v),\, x \ge x_k,\, v \le \sqrt{2E(x-x_k)}\}$; then

$$f \le \mu_k P \quad \text{on } D_k. \tag{3.27}$$

Proof. Recall that $E > 0$. Now for any pair $(x',v') \in D_k$, then $x' \ge x_k$ and $x' = \frac{v'^2}{2E} + x''$, where $x'' > x_k$. Now let (x',v') be fixed (and so is x''), and consider the function $g(v) = f\big(\frac{v^2}{2E} + x'', v\big)$. The argument of the right-hand side of the previous equation lies in D_k, and we have

$$E\frac{\partial g}{\partial v} + \sigma(v)g = Q^+(f)\left(\frac{v^2}{2E} + x'', v\right), \tag{3.28}$$

so that $g(v) \to 0$ as $v \to -\infty$ (because $f(x,v) \le P$).

Now since, by assumption, $Q^+(f) \le \mu_k Q^+(P)$ in D_k, then subtracting the differential inequality from the homogeneous equation satisfied by P, multiplied by μ_0, the difference $g - \mu_k P$ satisfies the differential inequality with the condition in velocity at $-\infty$

$$\begin{aligned} &E\frac{\partial}{\partial v}(g - \mu_k P) + \sigma(v)(g - \mu_k P) \le 0,\\ &\lim_{v\to-\infty} g(v) - \mu_k P(v) = 0. \end{aligned} \tag{3.29}$$

Since $E > 0$, it implies $g \le \mu_k P$. In particular, taking $v = v'$, we get

$$f(x',v') \le \mu_k P(v') \ \text{ on } \ D_k. \tag{3.30}$$

The proof is completed. ∎

The strategy in order to show (3.11), the expected behavior at infinity for f, consists of constructing pairs (μ_{k+1}, x_{k+1}) for which the control of the gain operator of f by that of P is improved (see (3.26)), and so by Lemma 5, the control of f by P (see (3.27)) is also improved by the same factor μ_{k+1} in such a way that the limit of the sequence $\{\mu_k\}$ is equal to $\lambda \ge 0$, while the limit for $\{x_k\}$ tends to $+\infty$.

The construction of such sequences of pairs entails the following iterative procedure. First, construct iteratively the sequence (μ_k, x_k) starting from $x_0 = 0$ and μ_0 given by Lemma 4, for as long as $f \le \mu_k P$ and $0 \le \lambda < \mu_k$ for $k \ge 0$.

Second, find a pair (μ_{k+1}, x_{k+1}) such that (3.26) holds, that is, $x_{k+1} > x_k$ and $Q^+(f) \le \mu_{k+1} Q^+(P)$ on D_{k+1}, where $\mu_{k+1} \le \mu_k$, where the selection of (μ_{k+1}, x_{k+1}) depends on μ_k and λ in a control way so $\lim_{k\to\infty} \mu_k = \lambda$ as $x_k \to \infty$.

Finally, from Lemma 5, it follows that $f \le \mu_{k+1} P$ on D_{k+1}, for $0 \le \lambda < \mu_{k+1} < \mu_k < 1$.

This next lemma proves this second step.

LEMMA 6. *Let f satisfy the conditions of Lemmas 4 and 5 for a given pair (μ_k, x_k), and the corresponding D_k for $\lambda < \mu_k$, for $k \ge 0$. Then there exists a (μ_{k+1}, x_{k+1}) such that*

$$Q^+(f) \le \mu_{k+1} Q^+(P), \qquad 0 < \mu_{k+1} < \mu_k < 1 \ \text{in } D_k, \tag{3.31}$$

with $x_{k+1} \ge x_k$. Moreover, μ_{k+1} can be chosen so that

$$(\mu_k - \lambda) < C(\mu_k - \mu_{k+1})^2 \,, \qquad \text{with } C = 2\frac{s_1}{s_0}\frac{\langle v^3 P\rangle^3}{\langle v^P\rangle^{1/2}}. \tag{3.32}$$

Before proving Lemma 6, we state a corollary which follows immediately from Lemmas 5 and 6.

COROLLARY 1. *Let D_{k+1} be defined as in Lemma 5; then*

$$f \le \mu_{k+1} P, \qquad 0 < \mu_{k+1} < \mu_k < 1\,, \qquad \text{on } \ D_{k+1}. \tag{3.33}$$

Proof of Lemma 6. In order to prove (3.31) due to the linearity of the collision form, write

$$Q^+(f) = \mu_k Q^+(P) - Q^+(\mu_k P - f). \tag{3.34}$$

Since $f \le \mu_k P$ in D_k, then

$$M(v)\int_{-\infty}^{0} s(v,v')\,(\mu_k P(v') - f(x,v'))\,dv' \ge 0 \qquad \text{for } x \ge x_k,$$

with $x_k = \frac{v_k^2}{2E}$. Thus

$$Q^+(\mu_k P - f) \ge M(v)\int_0^{\infty} s(v,v')\,(\mu_k P(v') - f(x,v'))\,dv', \qquad x \ge x_k. \tag{3.35}$$

Our goal is to see that the right-hand side of (3.35) is bounded by a proper fraction of $\mu_k Q^+(P)$, that is,

$$M(v)\int_0^{\infty} s(v,v')\,(\mu_k P(v') - f(x,v'))\,dv' \ge \mu_{k+1} Q^+(P) \tag{3.36}$$

for $\mu_{k+1} < \mu_k$ and $x \ge x_{k+1} \ge x_k$, where x_{k+1} is to be determined.

Now, we know from Lemma 5 that $f \le \mu_k P$ on D_k and that, by assumption, $\langle vf\rangle = \lambda\langle vP\rangle$. Then, as in Lemma 4, since the set $\{x \ge x_k\} \times \{v \le 0\}$ is in D_k, this implies

$$-\int_{-\infty}^{0} vf\,dv \le -\mu_k \int_{-\infty}^{0} vP\,dv = -\mu_k\langle vP\rangle + \mu_k \int_0^{+\infty} vP$$

for $x \geq x_k$, which yields

$$\int_0^{+\infty} v(\mu_k P - f)\, dv \geq (\mu_k - \lambda)\langle vP\rangle = \alpha_k \quad \text{for} \quad x \geq x_k. \tag{3.37}$$

Next, we need to choose the set D_{k+1}, which means choosing v_{k+1} and the corresponding x_{k+1} such that we can control them, and $Q^+(f) \leq \mu_{k+1}, Q^+(P)$ for $x_{k+1} > x_k$ and for some $\mu_{k+1} < \mu_k$.

In order to see this fact, first since we can do this construction, as done in Lemma 4, for as long as $\mu_k > \lambda$, and by the integrability properties of the first moment of P, one can choose v_{k+1} such that

$$v_{k+1} \int_{v_{k+1}}^{+\infty} P\, dv \leq \int_{v_{k+1}}^{+\infty} vP\, dv \leq \frac{1}{2}(\mu_k - \lambda)\langle vP\rangle. \tag{3.38}$$

On the other hand, as a consequence of (2.14) and Lemma 1, the third moment of P is bounded, and

$$\int_{v_{k+1}}^{+\infty} vP\, dv \leq \frac{1}{v_{k+1}^2} \int_0^{\infty} v^3 P(v)\, dv \leq \frac{\langle v^3 P\rangle}{v_{k+1}^2} \quad \text{for all } v_{k+1} \geq 0\,.$$

Then, choose v_{k+1} large enough such that both (3.38) and

$$v_{k+1} \leq \frac{C}{\sqrt{\mu_k - \lambda}}, \quad \text{with} \quad C = \frac{\langle v^3 P\rangle}{\sqrt{2\langle vP\rangle}}, \tag{3.39}$$

are satisfied.

Hence, taking $x_{k+1} = \frac{v_{k+1}^2}{2E} + x_k$, it is clear that

$$\{(x,v) : x \geq x_{k+1},\ v \leq v_{k+1}\} \subset D_k, \quad \text{so that} \quad f(x,v) \leq \mu_k P(v). \tag{3.40}$$

Next, rewrite integral estimate (3.37) as

$$\int_0^{v_{k+1}} v(\mu_k P - f)\, dv + \int_{v_{k+1}}^{+\infty} v(\mu_k P - f)\, dv \geq (\mu_k - \lambda)\langle vP\rangle. \tag{3.41}$$

Also, since $\mu_k < 1$ and $f > 0$, from the estimate from below in (3.38) it follows that

$$\int_{v_{k+1}}^{\infty} (\mu_k P - f) \leq \int_{v_{k+1}}^{\infty} P \leq \frac{1}{2v_{k+1}}(\mu_k - \lambda)\langle vP\rangle; \tag{3.42}$$

then, combining (3.41) and (3.42) yields

$$v_{k+1} \int_0^{v_{k+1}} (\mu_k P - f) \geq \int_0^{v_{k+1}} v(\mu_k P - f) \geq \left(1 - \frac{1}{2v_{k+1}}\right)(\mu_k - \lambda)\langle vP\rangle. \tag{3.43}$$

Finally, this last estimate (3.43) leads to the one involving a fraction of the gain operator on the difference $\mu_k P - f$, as follows. First, recalling $0 < s_0 \leq s(v', v) \leq s_1$ and s_1 finite,

$$\int_0^{v_{k+1}} s(v, v')\,(\mu_k P(v') - f(x, v'))\, dv' \geq \left(1 - \frac{1}{2v_{k+1}}\right) \frac{s_0}{v_{k+1}}(\mu_k - \lambda)\langle vP\rangle. \tag{3.44}$$

Second, since $f \le P$ and $\mu_k < 1$, then

$$\begin{aligned}\int_{v_{k+1}}^{\infty} s(v',v)\,(\mu_k P(v') - f(x,v'))\,dv' &\ge -\frac{s_1}{v_{k+1}}(1-\mu_k)\int_{v_{k+1}}^{\infty} v'P(v')dv' \\ &\ge -\frac{s_1}{v_{k+1}}(1-\mu_k)\frac{1}{2v_{k+1}}(\mu_k-\lambda)\langle vP\rangle\end{aligned} \tag{3.45}$$

for any $x > x_{k+1}$.

Therefore, gathering (3.44) and (3.45), we obtain the following lower estimate for equation (3.35):

$$\begin{aligned}Q^+(\mu_k P - f) &\ge M(v)\int_0^{+\infty} s(v',v)\,(\mu_k P(v') - f(x,v'))\,dv' \\ &\ge M(v)\left[\Big(1-\frac{1}{2v_{k+1}}\Big)s_0 - \frac{1}{2v_{k+1}}s_1(1-\mu_k)\right]\frac{\langle vP\rangle}{v_{k+1}}(\mu_k-\lambda) \\ &= M(v)\left[s_0 - \frac{1}{2v_{k+1}}(s_0+s_1) + \frac{s_1}{2}\frac{\mu_k}{v_{k+1}}\right]\frac{\langle vP\rangle}{v_{k+1}}(\mu_k-\lambda) \\ &= M(v)\left(1 - \frac{1}{2v_{k+1}}\Big[\frac{s_0+s_1}{s_0}\Big] + \frac{1}{2v_{k+1}}\frac{s_1}{s_0}\mu_k\right)s_0\frac{\langle vP\rangle}{v_{k+1}}(\mu_k-\lambda).\end{aligned} \tag{3.46}$$

Now, we can choose v_{k+1} even larger than the choices in (3.38) and (3.40) such that $\frac{1}{2v_{k+1}}[\frac{s_1}{s_0}\mu_k - \frac{s_0+s_1}{s_0}] < \frac{1}{2}$, and thus (3.46) leads to

$$Q^+(\mu_k P - f) \ge \frac{s_0}{2}\frac{\langle vP\rangle}{v_{k+1}}(\mu_k-\lambda)M(v) \ge \frac{1}{2}\frac{s_0}{s_1}\frac{\langle vP\rangle}{v_{k+1}}(\mu_k-\lambda)Q^+(P), \tag{3.47}$$

which, after combination with (3.34), leads to

$$Q^+(f) \le \mu_{k+1}Q^+(P), \tag{3.48}$$

where

$$\mu_{k+1} = \left(\mu_k - \frac{1}{2}\frac{s_0}{s_1}\frac{\langle vP\rangle}{v_{k+1}}(\mu_k-\lambda)\right). \tag{3.49}$$

Finally, from (3.39), v_{k+1} is such that $v_{k+1} \le C(\mu_k-\lambda)^{-1/2}$; then combining this with (3.49), we get

$$(\mu_k-\lambda) < C(\mu_k-\mu_{k+1})^2 \quad \text{with } C = 2\frac{s_1}{s_0}\frac{\langle v^3P\rangle}{\langle v^P\rangle^{1/2}}. \tag{3.50}$$

Hence, (3.32) holds as well, and thus the proof of Lemma 6 is now completed. ∎

We can now complete the proof of Theorem 5.

Proof of Theorem 5. For as long as $\mu_k > \lambda$, proceed constructing the sequence $\{x_k\}$ as in Lemma 6. If $\mu_k \le \lambda$, set $x_{k+1} = x_k$. In particular, since $\mu_k - \mu_{k+1} \to 0$, as k is large, inequality (3.32) implies

$$\lim_{k\to\infty}\mu_k - \lambda = 0 \quad \text{with} \quad \lim_{k\to\infty} x_k = +\infty,$$

which implies

$$\limsup_{k\to\infty} f(x_k,v) \le \lim_{k\to\infty}\mu_k P(v) \le \lambda P(v). \tag{3.51}$$

Conversely, applying Lemma 6 and Corollary 1 to $P - f$, since we have assumed $0 \le \lambda < 1$, then $0 \le \langle vP - vf\rangle \le (1-\lambda)\langle vP\rangle$, and also

$$\limsup_{k\to\infty}(P-f)(x_x, v) \le (1-\lambda)P(v) \le 0,$$

or equivalently,

$$\liminf_{k\to\infty} f(x_k, v) \ge \lambda P(v). \tag{3.52}$$

Finally, from the construction of the sequence, for μ_k either larger or smaller than λ, (3.51) and (3.52) imply that

$$\lim_{x\to\infty} f(x,v) = \lambda P(v),$$

so (3.11) holds. The proof of Theorem 5 is now completed. ∎

Finally, in order to complete the proof of the Theorem 1, we define $n_\infty = \frac{\langle vf\rangle}{\langle vP\rangle}$. Then we need to show that n_∞ is always a nonnegative proper fraction, since this has been an assumption in Theorem 5.

THEOREM 6. *If* $\langle vf\rangle = n_\infty\langle vP\rangle$, *then*

$$0 \le n_\infty \le 1. \tag{3.53}$$

Proof. First, we recall from the existence construction, if the boundary data is $0 < k(v) \le P(v)$, for $v > 0$, then $0 < f < P$ for all $x \ge 0$ and all v.

Now, argue by contradiction. If $n_\infty < 0$, take $g = f - n_\infty P$. Clearly $\langle vg\rangle = \langle v(f - n_\infty P)\rangle = 0$.

Therefore, applying Theorem 5 to g with $\lambda = 0$,

$$\lim_{x\to\infty} g(x,v) \le 0$$

or equivalently,

$$\lim_{x\to\infty} f(x,v) \le n_\infty P < 0,$$

contradicting $f > 0$ for all (x,v), $x \ge 0$.

Similarly, if $n_\infty > 1$, then take $g = n_\infty P - f$. Then, $g(x,v) < n_\infty P(v)$ and $\langle vg\rangle = 0$. Hence,

$$\lim_{x\to\infty}(n_\infty P - f) \le 0, \tag{3.54}$$

or equivalently,

$$n_\infty P(v) \le \lim_{x\to\infty} f(x,v) \le P(v),$$

which implies $n_\infty \le 1$, contradicting the assumption. Then (3.53) holds, so Theorem 6 is proven. ∎

Completion of Theorem 1. If $n_\infty < 1$, then, from Theorems 5 and 6,

$$\lim_{x\to\infty} f(x,v) = Kn_\infty P(v),$$

where n_∞ is the fraction of the first moment of $\frac{f}{K}$ with respect to the first moment of P. And, if $n_\infty = 1$ or, equivalently, $\langle v(K\,P - f)\rangle = 0$, one gets, as in (3.54), $0 \le \lim_{x\to\infty}(K\,P - f) \le 0$, since, by the existence, also $0 \le K\,P - f$. Hence

$$\lim_{x\to\infty} f(x,v) = K\,P(v).$$

The proof of Theorem 1 is now completed. □

As a corollary to Theorem 1 we have the following.

COROLLARY 2. *Assume that $E > 0$ and that $k(v) = n_\infty P(v)$. Then the unique solution of* (2.6) *is $n_\infty P(v)$.*

3.3. The Milne problem for strong negative forces. The aim of this subsection is the proof of Theorem 2. In the negative electric field case, we have proven that the upper and lower solutions coincide. This means that *the* solution of (2.6) can be obtained by solving a truncated problem (3.6) with $k_1 = k$ and k_2 arbitrary, the limit as L tends to $+\infty$ being only dependent on k. For positive electric fields, this will not be the case, and the solution *does* depend on the boundary condition k_2.

Proof of Theorem 2. We first proceed with the construction of a solution with the given asymptotic behavior; that is, we construct a solution f of (2.6) which behaves like $n_\infty P_{\sigma,E}(v)$ as x tends to $+\infty$. It is natural to consider the truncated problem

$$\begin{cases} v\partial_x f_L + E\partial_v f_L = Q(f_L)(x,v), \\ f_L(0,v) = k(v) \text{ for } v > 0, \\ f_L(L,v) = n_\infty P(v) \text{ for } v < 0. \end{cases} \tag{3.55}$$

Since $0 \le k(v) \le KP(v)$, the maximum principle insures that $0 \le f_L \le K_2 P$, where $K_2 = \max(n_\infty, K)$. Therefore, up to the extraction of a subsequence, f_L converges in L^∞_{loc} weak star towards a solution f of (2.6). Of course, since the convergence is only local in x, we cannot say anything at the moment about the asymptotic behavior of f. This is the purpose of the next step.

To analyze the asymptotic behavior, we consider the following truncated solutions:

$$\begin{cases} v\partial_x f^1_L + E\partial_v f^1_L = Q(f^1_L)(x,v), \\ f^1_L(0,v) = 0 \text{ for } v > 0, \\ f^1_L(L,v) = n_\infty P(v) \text{ for } v < 0, \end{cases} \tag{3.56}$$

$$\begin{cases} v\partial_x f^2_L + E\partial_v f^2_L = Q(f^2_L)(x,v), \\ f^2_L(0,v) = KP(v) \text{ for } v > 0, \\ f^2_L(L,v) = n_\infty P(v) \text{ for } v < 0. \end{cases} \tag{3.57}$$

Obviously, $f^1_L \le f_L \le f^2_L$. Considering the limits f^1 and f^2 of f^1_L and f^2_L, we have

$$f^1 \le f \le f^2.$$

Moreover,

$$f^2 = KP + \left(1 - \frac{K}{n_\infty}\right) f^1. \tag{3.58}$$

Besides, Proposition 1 insures that f^1 is increasing with respect to x. This implies the existence of α such that

$$\lim_{x\to+\infty} f_1(x,v) = \alpha P_{\sigma,E}(v).$$

It is enough to prove that $\alpha = n_\infty$, because (3.58) implies that f_2 also converges towards $n_\infty P$. Since f is sandwiched between f_1 and f_2, this implies that

$$\lim_{x\to+\infty} f(x,v) = n_\infty P_{\sigma,E}(v).$$

Let us now prove that $\alpha = n_\infty$. To this aim, we invert the x-axis direction by setting

$$g_L^1(x,v) = f_L^1(L-x,-v).$$

This function satisfies the equation

$$\text{(3.59)} \qquad \begin{cases} v\partial_x g_L^1 - E\partial_v g_L^1 = \widehat{Q}(g_L^1)(x,v), \\ g_L^1(0,v) = n_\infty P_{\widehat{\sigma},-E} \text{ for } v>0, \\ g_L^1(L,v) = 0 \text{ for } v<0, \end{cases}$$

where

$$\widehat{Q}(g)(v) = \int \widehat{\sigma}(v,v')(Mf' - M'f)dv', \qquad \widehat{\sigma}(v,v') = \sigma(-v,-v'),$$

and where we have noticed that

$$P_{\widehat{\sigma},-E}(v) = P_{\sigma,E}(-v).$$

With this transformation, we have replaced the electric field E by $-E$, and we are back to the positive force case. We know from Corollary 2 that the limit of $g_L^1(x,v)$ as L tends to $+\infty$ is nothing but $n_\infty P_{\widehat{\sigma},-E}(v)$. Therefore, we can pass to the limit in the current and get

$$\lim_{L\to+\infty} \langle v g_L^1\rangle = n_\infty \langle v P_{\widehat{\sigma},-E}\rangle = -n_\infty \langle v P_{\sigma,E}\rangle.$$

On the other hand, $\langle v g_L\rangle = -\langle v f_L^1\rangle$, which leads to

$$\lim_{L\to+\infty} \langle v g_L^1\rangle = -\langle v f^1\rangle = -\alpha \langle v P_{\sigma,E}\rangle.$$

As a consequence, $\alpha = n_\infty$, which is the desired result.

The proof of uniqueness is identical to that of Lemma 2 where the truncated case is considered. The details are left to the reader.

The proof of Theorem 2 is completed. ∎

4. A numerical method. As explained in the introduction, for the purpose of finding boundary or transition conditions we are interested only in the asymptotic state n_∞ and in the reflected density $f(0,v)$, $v<0$, of (2.6). In this section we will describe an approximation procedure to compute these values in the case of a relaxation collision operator.

Part of the motivation for studying this problem is that a direct discretization method to solve the half space problem is, in general, costly. The idea behind the method presented here is to solve the macroscopic equations associated with (2.6) and its adjoint equation and to use a Chapman–Enskog–type expansion as an approximate solution, which in the case of relaxation is an exact calculation where diffusion and transport coefficients depend explicitly on the force field E, via the moments of the distribution P, as shown in (2.17).

4.1. Computation of the asymptotic states. We consider the half space equation (2.6) in the relaxation case:

$$v\partial_x f + E\partial_v f = Q(f) = \langle f\rangle M - f, \qquad f(0,v) = k(v), \quad v > 0, \tag{4.1}$$

with $x \in [0,\infty)$ and $E > 0$ a constant. Due to Theorem 1 we have a unique solution f with $\lim_{x\to\infty} f(x,v) = n_\infty P(v)$, where P is the solution of (2.7). Our aim is to determine an accurate and efficient approximation of n_∞.

In addition to (4.1) we consider the corresponding adjoint equation using the weighted inner product $\langle fgP^{-1}\rangle$. It is given by

$$-v\partial_x g - P\partial_v(EgP^{-1}) = Q(gP^{-1}M)M^{-1}P. \tag{4.2}$$

Boundary conditions are

$$g(0,v) = 0, \quad v < 0.$$

A change of variables $v \to -v$ gives the equivalent equation

$$v\partial_x g + E\tilde{P}\partial_v(g\tilde{P}^{-1}) = Q(g\tilde{P}^{-1}M)M^{-1}\tilde{P}, \qquad g(0,v) = 0, \quad v > 0, \tag{4.3}$$

with $\tilde{P}(v) = P(-v)$.

This system is a particular case of the nonhomogeneous problem

$$v\partial_x g + E\tilde{P}\partial_v(g\tilde{P}^{-1}) = Q(g\tilde{P}^{-1}M)M^{-1}\tilde{P}, \qquad g(0,v) = k(v), \quad v > 0. \tag{4.4}$$

For this problem, we can prove the following theorem.

THEOREM 7. (i) *If $E < 0$ and $|k(v)| \le K\tilde{P}(v)$ for some constant K, then (4.4) has a unique solution such that $g/\tilde{P} \in L^\infty(\mathcal{R}_x^+ \times \mathcal{R}_v)$. This solution satisfies $|g(x,v)| \le K\tilde{P}(v)$.*

(ii) *If $E > 0$ and $|k(v)| \le K\tilde{P}(v)$ for some constant K, then, for any $j \in \mathcal{R}$, there exists a unique solution g of (4.4) such that $g\tilde{P} \in L^\infty(\mathcal{R}_x^+ \times \mathcal{R}_v)$ and $\int_{\mathcal{R}} vg(x,v)\,dv = j$. This unique solution is also characterized by the condition*

$$\lim_{x\to+\infty} g(x,v) = \frac{j}{\int v\tilde{P}\,dv}\,\tilde{P},$$

and g determines $n_\infty = \langle vf\rangle/\langle vP\rangle$ by

$$n_\infty = \int_{v>0} vk(v)g(0,v)P^{-1}(v)dv, \tag{4.5}$$

which is approximated by

$$n_\infty = \frac{\langle vk\rangle_+}{\langle vP\rangle_+} + \frac{\langle vP\rangle_-}{\langle vP\rangle\left\langle \frac{v}{1+Ev}P\right\rangle_+}\left(\left\langle \frac{v}{1+Ev}\left(k - \frac{\langle vk\rangle_+}{\langle vP\rangle_+}P\right)\right\rangle_+\right). \tag{4.6}$$

Proof. The proof of this theorem follows the same strategy as the proof of Theorems 1 and 2. We shall not redo this proof since it relies on exactly the same strategy, and we will give only some hints.

The first brick of the proof is the study of the homogeneous-in-x problem. It is clear that $\tilde{P}$ is a solution of

$$E\tilde{P}\partial_v(g\tilde{P}^{-1}) = Q(g\tilde{P}^{-1}M)M^{-1}\tilde{P}. \tag{4.7}$$

Actually, any solution g such that $g/\tilde{P} \in L^\infty$ is a multiple of $\tilde{P}$. Indeed, it is enough to prove that a solution g of (4.7) such that $\int g\,dv = 0$ is nothing but the identically vanishing function: to this end, we consider a function h which solves $E\partial_v h + Q(h) = g$. Such a solution exists since $\int g\,dv = 0$ (see [30]). Multiplying (4.7) by $h/\tilde{P}$ and integrating leads to $\int g^2\tilde{P}^{-1}\,dv = 0$.

The second property to be noticed is that $\int v\tilde{P}\,dv$ and E have opposite signs. This is why the sign of E is inverted in Theorem 7. With these remarks, one can reproduce the proofs of subsections 3.1 and 3.3 as well as the proof of Theorem 4 (subsection 3.2). We conjecture that the results of subsection 3.2 can be translated to the adjoint problem (4.4). This would imply that, in the case $E < 0$, the unique solution converges as x tends to $+\infty$ towards a multiple of $\tilde{P}$.

Let us now solve (4.3) approximately by proceeding similarly to the Chapman–Enskog expansion method. We recall that $E > 0$ in this section, so that $\langle vg\rangle$ has to be prescribed. Since the problem is linear, the solution is given up to a multiplication factor, and we choose g such that

$$\langle vg\rangle = -1.$$

The first step of the approximate resolution of (4.3) is to introduce a diffusion approximation: introducing an artificial small parameter δ, we consider

$$v\partial_x g + E\tilde{P}\partial_v(g\tilde{P}^{-1}) = \frac{1}{\delta}Q(g\tilde{P}^{-1}M)M^{-1}\tilde{P}. \tag{4.8}$$

Using the series expansion

$$g = g^0 + \delta g^1 + \cdots$$

in (4.8) and collecting terms of equal order in δ gives to $O(1)$

$$Q(g^0\tilde{P}^{-1}M) = 0$$

or

$$g^0 = \rho\tilde{P}. \tag{4.9}$$

To $O(\delta)$ we have

$$g^1 = M^{-1}\tilde{P}Q^{-1}\left[M\tilde{P}^{-1}(v\partial_x g^0 + E\tilde{P}\partial_v(g^0\tilde{P}^{-1}))\right].$$

Using (4.9) and $Q(vM) = -vM$, one obtains

$$g^1 = -v\tilde{P}\partial_x\rho.$$

The solvability conditions for the $O(\delta^2)$-equations give

$$\partial_x \langle v g^1 M \tilde{P}^{-1} \rangle + E \langle M \partial_v (g^1 \tilde{P}^{-1}) \rangle = 0.$$

Using the special form of g^1, this yields

$$\partial_x^2 \langle v^2 M \rangle \rho + E \partial_x \langle M \rangle \rho = 0$$

or the following equation for ρ:

$$\partial_x^2 \rho + E \partial_x \rho = 0. \tag{4.10}$$

Equation (4.10) is the drift-diffusion equation associated to the kinetic half space equation (4.2). We note that this is in contrast to (4.1). In this case the associated drift-diffusion equation is $\partial_x^2 \rho - E \partial_x \rho = 0$. The solution of (4.10) can be determined exactly up to two parameters:

$$\rho(x) = A e^{-Ex} + B, \quad \text{where} \quad A, B \in \mathcal{R}.$$

Next we compute an approximation $\hat{g}$ of g solving the following equation:

$$\begin{aligned} v \partial_x \hat{g} + E \tilde{P} \partial_v (\hat{g} \tilde{P}^{-1}) &= \rho \tilde{P} - \hat{g}, \\ \hat{g}(0, v) &= 0, \quad v > 0. \end{aligned} \tag{4.11}$$

This equation has been obtained from (4.3) by substituting the first order approximation $\rho \tilde{P}$ for g in $\langle g \tilde{P}^{-1} M \rangle$ into (4.3), where ρ is determined from the drift-diffusion equation (4.10). Notice that $\langle v \hat{g} \rangle$ is no longer independent of x. Now (4.11) can be further simplified by using the above approximation also in those terms in (4.11) involving E. One obtains

$$\begin{aligned} v \partial_x \hat{g} &= \rho(x) \tilde{P} - \hat{g}, \\ \hat{g}(0, v) &= 0, \quad v > 0. \end{aligned} \tag{4.12}$$

The solution of (4.12) can be given explicitly. Assuming boundedness at infinity of the solution, we get

$$\hat{g}(x, v) \tilde{P}^{-1} = \begin{cases} \dfrac{A}{1 - Ev} (e^{-Ex} - e^{-\frac{x}{v}}) + B(1 - e^{-\frac{x}{v}}), & v > 0, \\ \dfrac{A}{1 - Ev} e^{-Ex} + B\,, & v < 0. \end{cases}$$

In particular, $g(\infty, v) = B \tilde{P}$ and

$$\hat{g}(0, v) \tilde{P}^{-1} = \begin{cases} 0, & v > 0, \\ \dfrac{A}{1 - Ev} + B\,, & v < 0. \end{cases}$$

We determine A and B by

$$\left\langle v \begin{Bmatrix} \hat{g}(\infty, v) \\ \hat{g}(0, v) \end{Bmatrix} \right\rangle = -1,$$

the closest analogue to $\langle v g \rangle = -1$. This yields

$$B = -\frac{1}{\langle v \tilde{P} \rangle} = \frac{1}{\langle v P \rangle}$$

and

$$A = \frac{1}{\langle vP\rangle}\frac{\langle vP\rangle_-}{\langle \frac{v}{1+Ev}P\rangle_+}.$$

Here and in the following we use the notation

$$\langle f\rangle_+ = \int_{v>0} f(v)dv, \qquad \langle f\rangle_- = \int_{v<0} f(v)dv.$$

We mention that the g- approximation can be iterated considering the equation for the remaining term $g - \hat{g}$ instead of (4.3) and proceeding as before. Now, one transforms backwards, $v \to -v$, to get the desired approximation of the solution g of (4.2).

The following observation is crucial for the whole scheme: if f is a solution of (4.1) and g one of (4.2),

$$\begin{aligned}
&\partial_x(\langle vf(x,v)g(x,v)P^{-1}\rangle)\\
&\quad = \langle v(\partial_x f)(x,v)g(x,v)P^{-1}\rangle + \langle v(\partial_x g)(x,v)f(x,v)P^{-1}\rangle\\
&\quad = \langle [Q(f) - E\partial_v f]gP^{-1}\rangle\\
&\qquad + \langle f[EgP^{-1}\partial_v P - E\partial_v g - Q(gP^{-1}M)M^{-1}P]P^{-1}\rangle.
\end{aligned}$$

Since

$$\begin{aligned}
\langle Q(f)gP^{-1}\rangle &= \langle Q(f)(gP^{-1}M)M^{-1}\rangle\\
&= \langle fQ(gP^{-1}M)M^{-1}PP^{-1}\rangle
\end{aligned}$$

and

$$\begin{aligned}
\langle E\partial_v f g P^{-1}\rangle &= -\langle Ef\partial_v(gP^{-1})\rangle\\
&= -E\langle f(\partial_v g)P^{-1}\rangle + E\langle fg\partial_v PP^{-2}\rangle,
\end{aligned}$$

we get

$$\partial_x(\langle vf(x,v)g(x,v)P^{-1}\rangle) = 0.$$

In other words, $\langle v\varphi gP^{-1}\rangle$ is an invariant in x. Using this invariant, we get

$$\langle vf(\infty,v)g(\infty,v)P^{-1}(v)\rangle = \langle vf(0,v)g(0,v)P^{-1}(v)\rangle,$$

and substituting gives

$$\langle vn_\infty g(\infty,v)\rangle = \int_{v>0} vk(v)g(0,v)P^{-1}(v)dv.$$

Or, with $\langle vg(x,v)\rangle = 1$,

$$n_\infty = \int_{v>0} vk(v)g(0,v)P^{-1}(v)dv,$$

and thus (4.5) holds. In addition, since

$$g(0,v) \sim \hat{g}(0,v),$$

$\hat{g}$ is given by

$$\hat{g}(0,v)P^{-1} = \begin{cases} \dfrac{A}{1+Ev} + B, & v > 0, \\ 0, & v < 0, \end{cases}$$

with A and B determined above.

Altogether, we obtain

$$n_\infty = \frac{\langle vk\rangle_+}{\langle vP\rangle_+} + \frac{\langle vP\rangle_-}{\langle vP\rangle \left\langle \frac{v}{1+Ev}P\right\rangle_+} \left(\left\langle \frac{v}{1+Ev}\left(k - \frac{\langle vk\rangle_+}{\langle vP\rangle_+}P\right)\right\rangle_+ \right), \tag{4.13}$$

and so (4.6) holds and the proof of the theorem is completed. □

Remark. If $k(v) = \lambda P(v)$, we obtain from the above formula the correct value $n_\infty = \lambda$.

Remark. For E tending to 0 we have

$$\frac{1}{1+Ev} \sim 1 - Ev \sim 1 - Ev + O(E^2).$$

Moreover, $P \to M$ as $E \to 0$. Thus, one obtains in the limit the same result as, for example, in [22], namely,

$$n_\infty = \frac{\langle vk\rangle_+}{\langle vM\rangle_+} + \left\langle v^2\left(k - \frac{\langle vk\rangle_+}{\langle vM\rangle_+}M\right)\right\rangle_+.$$

4.2. Computation of the Albedo operator. The outgoing density $f(0,v)$, $v < 0$, of (4.1) can be computed as follows. We proceed in a similar same way as before; however, now $f(\infty, v) = n_\infty$ is known. We start directly with (4.1).

The drift-diffusion equation for this equation is determined by the same procedure as above. One obtains $\partial_x^2\rho - E\partial_x\rho = 0$. Looking for solutions bounded at infinity, one obtains

$$\rho = B,$$

B a constant. Substituting as before $\rho P(v)$ for f in $E\partial_v f$ and $\langle f\rangle$ in (4.1) gives

$$v\partial_x\hat{f} = \rho P - \hat{f},$$
$$\hat{f}(0,v) = k(v),\ v > 0.$$

The solution is

$$\hat{f}(x,v) = \begin{cases} k(v)e^{-\frac{x}{v}} + B(1 - e^{-\frac{x}{v}})P(v), & v > 0, \\ BP(v), & v < 0. \end{cases}$$

In particular, one obtains

$$\hat{f}(\infty, v) = BP(v)$$

and therefore $B = n_\infty$. Moreover, $\hat{f}(0,v) = n_\infty P(v)$, $v < 0$. Again, considering the equation for the remainder term $f - \hat{f}$, one obtains a better approximation of the outgoing function.

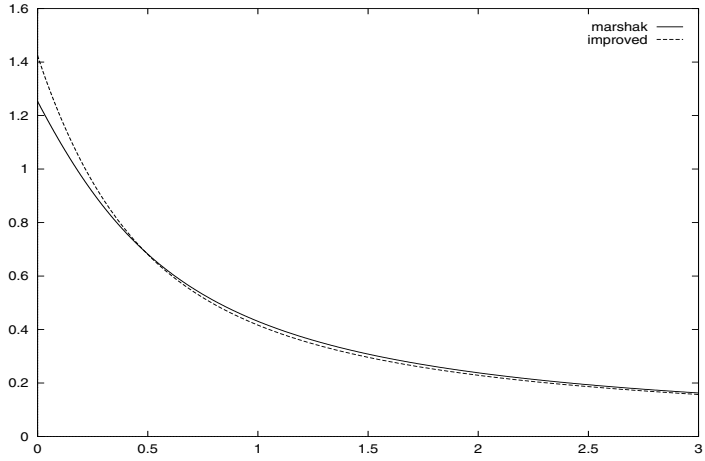


FIG. 1. *Asymptotic states for* $E \in [0, 3]$.

4.3. The Maxwell conditions. The following method was developed by Maxwell [27] and Marshak [26] (see also [6]) to derive approximate boundary conditions. In order to determine n_∞, one equalizes the half-fluxes at the boundary and at infinity, i.e.,

$$\int_{v>0} v\varphi(0,v)dv = \int_{v>0} v\varphi(\infty,v)dv,$$

which means in our context

$$n_\infty \int_{v>0} vP(v)dv = \int_{v>0} vk(v)dv$$

or

$$n_\infty = \frac{\langle vk\rangle_+}{\langle vP\rangle_+}. \tag{4.14}$$

This equality provides correct orders of magnitude in many situations. We observe that the value obtained by the procedure in section 4.1 obviously contains the term one obtains from the Marshak approximation (4.14). However, additionally, a correction term appears in (4.13). The Maxwell approximation of the outgoing distribution is simply

$$\varphi(0,v) = \varphi(\infty,v) = n_\infty P(v), \quad v < 0,$$

with n_∞ given by (4.14).

4.4. Numerical results. We used $k(v) = vM(v)$ to get in Figure 1 the asymptotic values for different values of the electric field $E > 0$. We computed these values by the approximations (4.14), labeled "marshak," and (4.13), labeled "variational." As E tends to 0 one obtains the same results as, for example, in [11, 22]: $n_\infty = 1.2533$ for the Maxwell–Marshak method and $n_\infty = 1.4245$ for the above approximation procedure, which is in case $E = 0$ equivalent to the so-called variational method; see, e.g., [25]. The true solution in this case is known: its numerical value is $n_\infty = 1.4371$.

REFERENCES

[1] M. A. Anile, J. A. Carrillo, I. M. Gamba, and C. W. Shu, *Approximation of the BTE by a relaxation-time operator: Simulations for a* 50*nm-channel Si diode*, VSLI Design J., 13 (2001), pp. 349–354.

[2] K. Aoki and Y. Sone, *Gas flows around the condensed phase with strong evaporation or condensation*, in Advances in Kinetic Theory and Continuum Mechanics, Proceedings of a Symposium in Honor of H. Cabannes, S. Gatignol, ed., Springer-Verlag, Berlin, 1991, pp. 43–54.

[3] M. D. Arthur and C. Cercignani, *Nonexistence of a steady rarefied supersonic flow in a half space*, Z. Angew. Math. Phys., 31 (1980), pp. 634–645.

[4] G. Baccarani, A. Gnudi, D. Ventura, and F. Odeh, *Two-dimensional MOSFET simulation by means of a multidimensional harmonic expansion of the Boltzmann transport equation*, Solid State Electron., 36 (1993), pp. 575–582.

[5] H. U. Baranger and J. W. Wilkins, *Ballistic structure in the electron distribution function of small semiconducting devices*, Phys. Rev. B, 36 (1987), pp. 1487–1502.

[6] C. Bardos, R. Santos, and R. Sentis, *Diffusion approximation and computation of the critical size*, in Numerical Solutions of Nonlinear Problems (Proceedings of the meeting in Rocquencourt, 1983), INRIA, Rocquencourt, France, 1984, pp. 1–39.

[7] N. Ben Abdallah and H. Chaker, *The high field asymptotics for degenerate semiconductors*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1253–1272.

[8] N. Ben Abdallah and P. Degond, *On a hierarchy of macroscopic models for semiconductors*, J. Math. Phys., 37 (1996), pp. 3306–3333.

[9] N. Ben Abdallah and J. Dolbeault, *Relative entropies for the Vlasov–Poisson system in bounded domains*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 867–872.

[10] J. A. Carrillo, I. M. Gamba, O. Muscato, and C. W. Shu, *Comparison of Monte Carlo and deterministic simulations of a silicon diode*, in Transport in Transition Regimes, IMA Vol. Math. Appl. 135, Springer-Verlag, New York, 2004, pp. 75–84.

[11] C. Cercignani, *The Boltzmann Equation and Its Applications*, Springer, New York, 1988.

[12] C. Cercignani, I. M. Gamba, and C. D. Levermore, *High field approximation to a Boltzmann–Poisson system and boundary conditions in a semiconductor*, Appl. Math. Lett., 10 (1997), pp. 111–118.

[13] C. Cercignani, I. M. Gamba, and C. D. Levermore, *A drift-collision balance for a Boltzmann–Poisson system in bounded domains*, SIAM J. Appl. Math., 61 (2001), pp. 1932–1958.

[14] F. Coron, *Computation of the asymptotic states for linear halfspace problems*, Transport. Theory Statist. Phys., 19 (1990), pp. 89–114; errata in 19 (1990) p. 581.

[15] F. Coron, F. Golse, and C. Sulem, *A classification of well-posed kinetic layer problems*, Comm. Pure Appl. Math., 41 (1988), pp. 409–439.

[16] P. Degond, *A model of near-wall conductivity and its application to plasma thrusters*, SIAM J. Appl. Math., 58 (1998), pp. 1138–1162.

[17] P. Degond, *An infinite system of diffusion equations arising in transport theory: The coupled spherical harmonic expansion model*, Math. Models Methods Appl. Sci., 11 (2001), pp. 903–932.

[18] P. Degond, N. Ben Abdallah, and S. Génieys, *An energy transport model for semiconductors derived from the Boltzamnn equation*, J. Statist. Phys., 84 (1996), pp. 205–231.

[19] G. Frosali, C. V. M. van der Mee, and S. L. Paveri-Fontana, *Conditions for runaway phenomena in the kinetic theory of particle swarms*, J. Math. Phys., 30 (1989), pp. 1177–1186.

[20] I. M. Gamba, J. A. Carrillo, and C. W. Shu, *Computational macroscopic approximations to the* 1*-D relaxation-time kinetic system for semiconductors*, Phys. D, 146 (2000), pp. 289–306.

[21] L. Garrigues, P. Degond, V. Latocha, and J. P. Boeuf, *Electron transport in stationary plasma thrusters*, Transport Theory Statist. Phys., 27 (1998), pp. 203–221.

[22] F. Golse and A. Klar, *A numerical method for computing asymptotic states and outgoing distributions for kinetic linear half-space problems*, J. Statist. Phys., 80 (1995), pp. 1033–1061.

[23] W. Greenberg, C. van der Mee, and V. Protopopescu, *Boundary Value Problems in Abstract Kinetic Theory*, Birkhäuser Boston, Cambridge, MA, 1987.

[24] A. Klar, *A numerical method for kinetic semiconductor equations in the drift-diffusion limit*, SIAM J. Sci. Comput., 20 (1999), pp. 1696–1712.

[25] S. K. Loyalka, *Approximate method in the kinetic theory*, Phys. Fluids, 14 (1971), pp. 2291–2294.

[26] R. E. Marshak, *The Milne problem for a large plane slab with constant source and anisotropic scattering*, Phys. Rev., 72 (1947), pp. 47–50.
[27] J. C. Maxwell, *The Scientific Papers of J. C. Maxwell*, Dover, New York, 1965.
[28] F. Poupaud, *Diffusion approximation of the linear semiconductor equation*, J. Asympt. Anal., 4 (1991), pp. 293–317.
[29] F. Poupaud, *Boundary value problems for the stationary Vlasov-Maxwell system*, Forum Math., 4 (1992), pp. 499–527.
[30] F. Poupaud, *Runaway phenomena and fluid approximation under high fields in semiconductor kinetic theory*, ZAMM Z. Angew. Math. Mech., 72 (1992), pp. 359–372.
[31] A. Saul, P. Dmitruk, and L. Reyna, *High electric field approximation in semiconductor devices*, Appl. Math. Lett., 5 (1992), pp. 99–102.
[32] C. Schmeiser and A. Zwirchmayr, *Elastic and drift-diffusion limits of electron-phonon interaction in semiconductors*, Math. Models Methods Appl. Sci., 8 (1998), pp. 37–54.
[33] C. E. Siewert and J. R. Thomas, *Strong evaporation into a half space* II, Z. Angew. Math. Phys., 33 (1982), pp. 202–218.
[34] S. A. Trugman and A. J. Taylor, *Analytic solution of the Boltzmann equation with applications to electron transport in inhomogeneous semiconductors*, Phys. Rev. B, 33 (1986), pp. 5575–5583.
[35] A. Yamnahakki, *Second order boundary conditions for the drift-diffusion equations of semiconductors*, Math. Models Methods Appl. Sci., 5 (1995), pp. 429–455.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1737–1748

DIFFUSIVE RELAXATION LIMIT OF MULTIDIMENSIONAL ISENTROPIC HYDRODYNAMICAL MODELS FOR SEMICONDUCTORS*

WEN-AN YONG†

Abstract. This work is concerned with multidimensional isentropic hydrodynamical models for semiconductors with short momentum relaxation time. With the help of the Maxwell iteration, we prove that, as the relaxation time tends to zero, periodic initial-value problems of a certain scaled hydrodynamical model have unique smooth solutions existing in the time interval where the classical drift-diffusion model has smooth solutions. Meanwhile, we justify a formal derivation of the latter from the former.

Key words. semiconductor models, diffusive relaxation limit, the Maxwell iteration, H^s-solutions, energy estimates

AMS subject classifications. 35L45, 35B25, 35M20

DOI. 10.1137/S0036139903427404

1. Introduction.

This work is concerned with the following scaled hydrodynamical model for semiconductors or plasmas:

$$\begin{aligned} &\partial_t n + \frac{1}{\epsilon}\operatorname{div}(nu) = 0,\\ &\partial_t(nu) + \frac{1}{\epsilon}\operatorname{div}(nu\otimes u) + \frac{1}{\epsilon}\nabla p(n) = \frac{n\nabla\phi}{\epsilon} - \frac{nu}{\epsilon^2},\\ &-\Delta\phi = \tilde{n}(x) - n. \end{aligned} \tag{1.1}$$

Here $n > 0, u$, and ϕ, as unknown functions of $(x,t) \in \mathbf{R}^d \times [0,+\infty)$ with $d \geq 1$, denote the electron density, velocity (d-vector), and electrostatic potential, respectively; $p = p(n)$ is a given strictly increasing function and denotes the pressure; $\tilde{n}(x)$ is the given background density of holes or ions; $\epsilon > 0$ is a parameter for the momentum relaxation time; and div, ∇, Δ, and $\otimes$ are the respective x-divergence operator, gradient operator, Laplacian, and symbol for the tensor products of two vectors. Note that the scaling

$$t = \epsilon\tilde{t}$$

converts (1.1) back into the original unipolar model in [1] with $\tilde{t}$ as its time variable. The scaled-time variable t was first introduced in [14] to study the relation between the hydrodynamical and drift-diffusion models [19, 18] (see (1.2) below).

It is known from [1, 16] that the hydrodynamical model describes some physical phenomena not accounted for in the classical drift-diffusion model. However, based on the previous results in [14, 6, 9] we expect that the two models give similar results

*Received by the editors May 8, 2003; accepted for publication (in revised form) October 24, 2003; published electronically July 2, 2004. This work was supported by the Deutsche Forschungsgemeinschaft through the Schwerpunktprogramm ANumE and SFB 359 at the University of Heidelberg and by the European TMR-Network "Hyperbolic and Kinetic Equations."
http://www.siam.org/journals/siap/64-5/42740.html
†Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany (yong.wen-an@iwr.uni-heidelberg.de).

when ϵ is small, which can be seen formally as follows. Applying the Maxwell iteration to the momentum equation in (1.1) gives

$$\begin{aligned} nu &= -\epsilon\nabla p(n) + \epsilon n\nabla\phi - \epsilon\,\mathrm{div}(nu\otimes u) - \epsilon^2\partial_t(nu)\\ &= -\epsilon\nabla p(n) + \epsilon n\nabla\phi + O(\epsilon^2). \end{aligned}$$

Substituting the truncation $nu = -\epsilon\nabla p(n) + \epsilon n\nabla\phi$ into the mass equation in (1.1), we arrive at the unipolar drift-diffusion model

$$\begin{aligned} \partial_t n &= \Delta p(n) - \mathrm{div}(n\nabla\phi),\\ -\Delta\phi &= \tilde{n}(x) - n. \end{aligned} \tag{1.2}$$

This is a parabolic-elliptic system, since $p(n)$ is strictly increasing.

The goal of this paper is to justify the above formal derivation of the drift-diffusion model for periodic IVPs (initial-value problems) with an emphasis on several space dimensions. For brevity, we deal with only the unipolar model (1.1). However, it is trivial to see that our arguments and results hold true for the bipolar model in [1, 16] and the hydrodynamical model for porous media discussed in [13].

In one space dimension, the above limit problem has been investigated by many authors in the compactness frameworks for nonsmooth solutions of conservation laws (see [14, 17, 6, 2, 3, 4]). However, the multidimensional limit problem is hardly studied in the literature, and [9, 8] are the only papers known to the author. In [9, 8], the authors considered (1.1) and the corresponding bipolar model with x in a bounded domain, assumed the existence of L^∞-solutions in an ϵ-independent time interval, and justified the relaxation limit in a compactness framework [13] for nonsmooth solutions.

In this paper, we revise the approach in [20] to study the multidimensional limit problem. Precisely, we assume that the drift-diffusion model (1.2) has a smooth solution (n,ϕ) with initial data $n(x,0) = n_0(x)$. Inspired by the Maxwell iteration above, we construct a formal approximation

$$n_\epsilon = n, \qquad n_\epsilon u_\epsilon = \epsilon n\nabla\phi - \epsilon\nabla p(n), \quad \phi_\epsilon = \phi \tag{1.3}$$

for the solution $(n^\epsilon, u^\epsilon, \phi^\epsilon)$ of (1.1) with initial data

$$n(x,0) = n_0(x), \qquad u(x,0) = \epsilon\nabla\phi(x,0) - \frac{\epsilon\nabla p(n_0)}{n_0}. \tag{1.4}$$

(Note that these initial data are in equilibrium.) Then we use energy methods to prove that $(n^\epsilon, u^\epsilon, \phi^\epsilon)$ exists in the finite time interval where n is well defined and can be expressed as

$$(n^\epsilon, u^\epsilon, \phi^\epsilon) = (n_\epsilon, u_\epsilon, \phi_\epsilon) + O(\epsilon^2) \tag{1.5}$$

in the Sobolev space $H^s(\mathbb{T}^d)$ with $s > d/2 + 1$. Furthermore, our conclusion implies that if the drift-diffusion model has a global smooth solution with n having a positive lower bound, then for any $T > 0$ there exists $\epsilon_0 > 0$ such that the hydrodynamical model has a unique smooth solution up to the time T when $\epsilon < \epsilon_0$. See Theorem 4.1 in section 4 for details.

Regarding the above result, we make the following remarks. First, (1.5) is more precise than the previous equations for weak solutions if the parabolic-elliptic system (1.2) has a smooth solution with n having a positive lower bound. For this assumption

about (1.2), the interested reader is referred to [5]. Second, it is not difficult to obtain results of the form (1.5) for more general periodic initial data by using the matched expansion method (see, e.g., [20, 10]) instead of the Maxwell iteration for the special data (1.4).

On the other hand, it does not seem easy to extend the above result for nonperiodic IVPs or initial-boundary value problems. In fact, when the initial data are not periodic, the Poisson equation for ϕ should be treated differently (see Proposition 2.1). And when initial-boundary value problems are concerned, new ideas are required to deal with the remaining (hyperbolic) part of the hydrodynamical model (1.1).

We note that, besides having the Poisson equation, the hydrodynamical models are essentially different from the hyperbolic systems studied in [10]. In fact, once the hydrodynamical models are rewritten as symmetrizable hyperbolic systems, the coefficients multiplying the spatial derivatives depend on the unknown in a stronger way than that in [10] (see Remark 2.1 in section 2). Because of this strong dependence, some key estimates in [10] have to be reconsidered, and our analysis depends heavily on the special nonlinear structure of the isentropic hydrodynamical models (see the proofs of Lemmas 4.2 and 4.3 in section 4).

Finally, let us mention that this work can be regarded as a contribution to the theory of diffusive limits for hyperbolic problems developed in [15]. Another limit problem for the hydrodynamical models is the large time behavior of the solutions [11], where stationary, instead of time-dependent, solutions of the drift-diffusion models are involved.

This paper is organized as follows. In section 2 we rewrite the hydrodynamical models as symmetrizable hyperbolic systems and review the convergence-stability lemma from [21]. Section 3 is devoted to the formal approximation (1.3). In section 4 we prove the validity of the formal approximation and conclude the existence of the solution to (1.1) in the time interval where n is well defined.

Notation. $|U|$ denotes some norm of a vector or matrix U. $L^2 = L^2(\mathbb{T}^d)$ is the space of square integrable (vector- or matrix-valued) functions on the d-dimensional unit torus $\Omega = (0,1]^d$. For a nonnegative integer s, $H^s = H^s(\mathbb{T}^d)$ is defined as the space of functions whose distribution derivatives of order $\le s$ are all in L^2. We use $\|U\|_s$ to denote the standard norm of $U \in H^s$, and $\|U\| \equiv \|U\|_0$. When A is a function of another variable t as well as x, we write $\|A(\cdot,t)\|_s$ to recall that the norm is taken with respect to x while t is viewed as a parameter. In addition, we denote by $C([0,T],\mathbf{X})$ (resp., $C^1([0,T],\mathbf{X})$) the space of continuous (resp., continuously differentiable) functions on $[0,T]$ with values in a Banach space $\mathbf{X}$.

2. Preliminaries. In this section, we write the hydrodynamical models as symmetrizable hyperbolic systems and review the convergence-stability lemma from [21]. To begin with, we recall the following elementary fact, which can be easily proven by using Fourier series.

PROPOSITION 2.1. *$\nabla\Delta^{-1}$ is a bounded linear operator on $L^2(\mathbb{T}^d)$.*

It is this proposition that requires the initial data to be periodic.

Now we write (1.1) as a symmetric hyperbolic system. To do this, we introduce the *enthalpy* $h = h(n) > 0$ defined for $n > 0$ and satisfying

$$h'(n) = \frac{p'(n)}{n}.$$

Since $p(n)$ is strictly increasing, so is $h(n)$. Thus, $h(n)$ has an inverse $n = n(h)$. Set

$$q(h) = p'(n(h)).$$

Then, for smooth solutions, (1.1) is equivalent to

$$\begin{gathered} q(h)^{-1}\left(\partial_t h+\frac{1}{\epsilon}u\cdot\nabla h\right)+\frac{1}{\epsilon}\operatorname{div}u=0,\\ \partial_t u+\frac{1}{\epsilon}u\cdot\nabla u+\frac{1}{\epsilon}\nabla h=\frac{\nabla\phi}{\epsilon}-\frac{u}{\epsilon^2},\\ -\Delta\phi=\tilde{n}(x)-n(h) \end{gathered} \tag{2.1}$$

or

$$\partial_t\begin{pmatrix}h\\u\end{pmatrix}+\frac{1}{\epsilon}\sum_{j=1}^{d}A_j(h,u)\partial_{x_j}\begin{pmatrix}h\\u\end{pmatrix}=\frac{1}{\epsilon^2}\begin{pmatrix}0\\\epsilon\nabla\Delta^{-1}(n(h)-\tilde{n})-u\end{pmatrix}. \tag{2.2}$$

Here the coefficients have the following structure:

$$\begin{gathered} A_j(h,u)=A_0^{-1}(h)C_j+u_jI_{d+1},\\ A_0(h)=\operatorname{diag}\left(\frac{1}{q(h)},I_d\right),\\ \text{each } C_j \text{ is a constant symmetric matrix,}\\ \text{and the first element } C_j^{11} \text{ in the first row of } C_j \text{ is zero,} \end{gathered} \tag{2.3}$$

where I_k denotes the unit matrix of order k and u_j is the jth component of u. Thus, (2.2) is a symmetrizable hyperbolic system with A_0 the symmetrizer.

Remark 2.1. Although (2.2) is of the form of the systems studied in [10], it is essentially different from them. In fact, a crucial assumption in [10] is that the coefficients A_j and the symmetrizer A_0 depend on the unknown $W\equiv(h,u)$ only through ϵW, that is, $A_j=A_j(\epsilon W)$ and $A_0=A_0(\epsilon W)$. This assumption is obviously not satisfied by our present system (2.2).

Thanks to Proposition 2.1, the local-in-time existence theory for periodic IVPs of first-order symmetrizable hyperbolic systems can be well applied to (2.2). Moreover, we recall the *convergence-stability lemma* in [21] for general singular limit problems of IVPs for quasi-linear first-order symmetrizable hyperbolic systems depending (singularly) on parameters in several space variables:

$$\begin{gathered} U_t+\sum_{j=1}^{d}A_j(U,\epsilon)U_{x_j}=Q(U,\epsilon),\\ U(x,0)=\bar{U}(x,\epsilon). \end{gathered} \tag{2.4}$$

Here ϵ represents a parameter in a topological space, $A_j(U,\epsilon)$ $(j=1,2,\dots,d)$ and $Q(U,\epsilon)$ are sufficiently smooth functions of $U\in G\subset\mathbf{R}^n$, and $\bar{U}(x,\epsilon)$ is a given initial-value function. For simplicity, we assume that $\bar{U}(x,\epsilon)$ is periodic in x with period $(1,1,\dots,1)\in\mathbf{R}^d$.

Assume $\bar{U}(x,\epsilon)\in G_0\subset\subset G$ for all (x,ϵ) and $\bar{U}(\cdot,\epsilon)\in H^s$ with $s>d/2+1$ an integer. Fix ϵ. According to the local existence theory for IVPs of symmetrizable hyperbolic systems (see Theorem 2.1 in [12]), there is a time interval $[0,T]$ so that (2.4) has a unique H^s-solution

$$U^\epsilon\in C([0,T],H^s).$$

Define

$$T_\epsilon = \sup\{T > 0 : U^\epsilon \in C([0,T], H^s)\}. \tag{2.5}$$

Namely, $[0, T_\epsilon)$ is the maximal time interval of H^s existence. Note that T_ϵ depends on G and may tend to zero as ϵ goes to a certain singular point, say 0.

In order to show that $\underline{\lim}_{\epsilon\to 0} T_\epsilon > 0$, which means the stability (see [7, 12]), we make the following assumption.

Convergence assumption. There exists $T_* > 0$ and $U_\epsilon \in L^\infty([0,T_*], H^s)$ for each ϵ, satisfying

$$\bigcup_{x,t,\epsilon} \{U_\epsilon(x,t)\} \subset\subset G,$$

such that for $t \in [0, \min\{T_*, T_\epsilon\})$,

$$\sup_{x,t} |U^\epsilon(x,t) - U_\epsilon(x,t)| = o(1),$$
$$\sup_t \|U^\epsilon(\cdot,t) - U_\epsilon(\cdot,t)\|_s = O(1)$$

as ϵ tends to the singular point.

With such a convergence assumption, we are in a position to state the following fact established in [21].

LEMMA 2.2. *Suppose $\bar{U}(x,\epsilon) \in G_0 \subset\subset G$ for all (x,ϵ), $\bar{U}(\cdot,\epsilon) \in H^s$ with an integer $s > d/2 + 1$, and that the convergence assumption holds. Let $[0, T_\epsilon)$ be the maximal time interval such that* (2.4) *has a unique H^s-solution $U^\epsilon \in C([0,T_\epsilon), H^s)$. Then*

$$T_\epsilon > T_*$$

for all ϵ in a neighborhood of the singular point.

Thanks to Lemma 2.2, our task is reduced to finding a $U_\epsilon(x,t)$ such that the convergence assumption holds. Below, we will use this lemma with G replaced by its compact subsets.

3. Formal approximations. In this section we propose a construction of the approximation U_ϵ in the convergence assumption for the hydrodynamical model (2.2). Let n solve the IVP of the unipolar drift-diffusion model (1.2) or

$$\begin{aligned} \partial_t n &= \Delta p(n) - \operatorname{div}(n\nabla\Delta^{-1}(n-\tilde{n})), \\ n(x,0) &= n_0(x). \end{aligned} \tag{3.1}$$

Inspired by the Maxwell iteration, we take

$$\begin{aligned} n_\epsilon &= n, \\ u_\epsilon &= \epsilon\nabla\Delta^{-1}(n-\tilde{n}) - \frac{\epsilon\nabla p(n)}{n}. \end{aligned} \tag{3.2}$$

Define

$$\begin{aligned} R = \frac{\partial_t u_\epsilon + u_\epsilon\cdot\nabla u_\epsilon/\epsilon}{\epsilon} &= \partial_t(\nabla\Delta^{-1}(n-\tilde{n}) - \nabla h) \\ &+(\nabla\Delta^{-1}(n-\tilde{n}) - \nabla h)\cdot\nabla(\nabla\Delta^{-1}(n-\tilde{n}) - \nabla h). \end{aligned} \tag{3.3}$$

Then we have

$$\partial_t n_\epsilon + \frac{1}{\epsilon}\operatorname{div}(n_\epsilon u_\epsilon) = 0,$$

$$\partial_t(n_\epsilon u_\epsilon) + \frac{1}{\epsilon}\operatorname{div}(n_\epsilon u_\epsilon \otimes u_\epsilon) + \frac{1}{\epsilon}\nabla p(n_\epsilon) = \frac{n_\epsilon \nabla\Delta^{-1}(n_\epsilon - \tilde{n})}{\epsilon} - \frac{n_\epsilon u_\epsilon}{\epsilon^2} + \epsilon n_\epsilon R,$$

or

$$\begin{aligned} q(h_\epsilon)^{-1}\left(\partial_t h_\epsilon + \frac{1}{\epsilon}u_\epsilon\cdot\nabla h_\epsilon\right) + \frac{1}{\epsilon}\operatorname{div} u_\epsilon &= 0, \\ \partial_t u_\epsilon + \frac{1}{\epsilon}u_\epsilon\cdot\nabla u_\epsilon + \frac{1}{\epsilon}\nabla h_\epsilon &= \frac{\nabla\Delta^{-1}(n_\epsilon - \tilde{n})}{\epsilon} - \frac{u_\epsilon}{\epsilon^2} + \epsilon R. \end{aligned} \tag{3.4}$$

Regarding (n_ϵ, u_ϵ), we have the following regularity result.

LEMMA 3.1. *Let $s > d/2$ be an integer. Assume $p \in C^\infty(0,\infty)$ and $p'(n) > 0$. If $n \in C([0,T_*], H^s) \cap C^1([0,T_*], H^{s-1})$ has a positive lower bound, then so does $h = h(n)$. Moreover, if $\tilde{n} \in H^{s-1}$, then $u_\epsilon \in C([0,T_*], H^{s-1}) \cap C^1([0,T_*], H^{s-2})$ and $R \in C([0,T_*], H^{s-2})$ in case $s > d/2 + 1$.*

The proof of this lemma is based on the well-known calculus inequalities in Sobolev spaces, which we state here for further reference and for the convenience of the reader.

LEMMA 3.2 (see, e.g., [12]). *Let s, s_1, and s_2 be three nonnegative integers and $s_0 = [d/2] + 1$.*

1. *If $s_3 = \min\{s_1, s_2, s_1 + s_2 - s_0\} \geq 0$, then $H^{s_1}H^{s_2} \subset H^{s_3}$. Here the inclusion symbol $\subset$ implies the continuity of the embedding.*

2. *Suppose $s \geq s_0 + 1$, $A \in H^s$, and $U \in H^{s-1}$. Then for all multi-indices α with $|\alpha| \leq s$, $\partial^\alpha(AU) - A\partial^\alpha U \in L^2$ and*

$$\|\partial^\alpha(AU) - A\partial^\alpha U\| \leq C_s\|A\|_s\|U\|_{|\alpha|-1}.$$

3. *Suppose $s \geq s_0$, $A \in C_b^s(G)$, and $V \in H^s(\Omega, G)$. Then $A(V(\cdot)) \in H^s$ and*

$$\|A(V(\cdot))\|_s \leq C_s|A|_s(1 + \|V\|_s^s).$$

Here and below, C_s denotes a generic constant depending only on s and d, and $|A|_s$ stands for $\sup_{\{U\in G, |\alpha|\leq s\}} |\partial_U^\alpha A(U)|$.

4. The main result. Having constructed the formal approximation (n_ϵ, u_ϵ) for the periodic IVP of the hydrodynamical model (2.2), we prove here the validity of the approximation under some regularity assumptions on the given data and an existence result for the IVP. The main result of this paper is stated as follows.

THEOREM 4.1. *Let $s > d/2 + 1$ be an integer. Suppose $p \in C^\infty(0,+\infty)$, $p'(n) > 0$, $\tilde{n} \in H^s(\mathbf{R}^d)$, and that the drift-diffusion model (3.1) has a solution $n \in C([0,T_*], H^{s+2}) \cap C^1([0,T_*], H^{s+1})$ with a positive lower bound.*

Then, for ϵ sufficiently small, the hydrodynamical model (2.2) with periodic initial data

$$\begin{aligned} h(x,0) &= h(n(x,0)), \\ u(x,0) &= \epsilon\nabla\Delta^{-1}(n(x,0) - \tilde{n}) - \frac{\epsilon\nabla p(n(x,0))}{n(x,0)} \end{aligned} \tag{4.1}$$

has a unique solution $(n^\epsilon, u^\epsilon) \in C([0,T_*], H^s)$, and there exists a constant $K > 0$, independent of ϵ but dependent on $T_* < \infty$, such that

$$\sup_{t\in[0,T_*]} \|(n^\epsilon - n_\epsilon, u^\epsilon - u_\epsilon)(\cdot,t)\|_s \leq K\epsilon^2. \tag{4.2}$$

Proof. Since $n \in C([0,T_*],H^{s+1})$ has a positive lower bound, there are two positive numbers a and b such that $h(n)$ takes values in $[2a,b]$. Denote by $[0,T_\epsilon)$ the maximal time interval where the symmetrizable hyperbolic system (2.2) with the initial data (4.1) has a unique H^s-solution (h^ϵ,u^ϵ) with values in $(a,2b)\times\mathbf{R}^d \equiv G$. Thanks to Lemma 2.2, it suffices to prove the error estimate in (4.2) for $t \in [0,\min\{T_*,T_\epsilon\})$.

To this end, we set

$$E = \begin{pmatrix} h_\epsilon - h^\epsilon \\ u_\epsilon - u^\epsilon \end{pmatrix} \equiv \begin{pmatrix} E^h \\ E^u \end{pmatrix}.$$

From the equations in (2.2) and (3.4) it follows that the error E satisfies

$$E_t + \frac{1}{\epsilon}\sum_{j=1}^d A_j(h^\epsilon,u^\epsilon)E_{x_j} = \frac{-1}{\epsilon^2}\begin{pmatrix} 0 \\ E^u \end{pmatrix} + \frac{1}{\epsilon}\begin{pmatrix} 0 \\ \nabla\Delta^{-1}(n(h_\epsilon)-n(h^\epsilon)) + \epsilon^2 R \end{pmatrix} + \frac{1}{\epsilon}\sum_{j=1}^d [A_j(h^\epsilon,u^\epsilon) - A_j(h_\epsilon,u_\epsilon)]\begin{pmatrix} h_{\epsilon x_j} \\ u_{\epsilon x_j} \end{pmatrix}.$$

We differentiate this equation with ∂^α (in x) for a multi-index α satisfying $|\alpha| \le s$ to get

$$E_{\alpha t} + \frac{1}{\epsilon}\sum_{j=1}^d A_j(h^\epsilon,u^\epsilon)E_{\alpha x_j} = \frac{-1}{\epsilon^2}\begin{pmatrix} 0 \\ E^u_\alpha \end{pmatrix} + F_1^\alpha + F_2^\alpha, \tag{4.3}$$

where

$$\begin{aligned} F_1^\alpha &= \frac{1}{\epsilon}\begin{pmatrix} 0 \\ \nabla\Delta^{-1}(n(h_\epsilon)-n(h^\epsilon)) + \epsilon^2 R \end{pmatrix}_\alpha, \\ F_2^\alpha &= \frac{1}{\epsilon}\sum_{j=1}^d \left([A_j(h^\epsilon,u^\epsilon) - A_j(h_\epsilon,u_\epsilon)]\begin{pmatrix} h_{\epsilon x_j} \\ u_{\epsilon x_j} \end{pmatrix}\right)_\alpha \\ &\quad + \frac{1}{\epsilon}\sum_{j=1}^d (A_j(h^\epsilon,u^\epsilon)E_{\alpha x_j} - (A_j(h^\epsilon,u^\epsilon)E_{x_j})_\alpha) \\ &\equiv f_1^\alpha + f_2^\alpha. \end{aligned}$$

For the sake of clarity, we divide the following arguments into lemmas.

LEMMA 4.2. *Under the conditions of Theorem* 4.1*, we have*

$$\frac{d}{dt}\int_\Omega e(E_\alpha)dx + \frac{2}{\epsilon^2}\|E^u_\alpha\|^2 \le 2\|E^u_\alpha\|\|F_1^\alpha\| + \frac{C}{\epsilon}\int_\Omega |\operatorname{div} u^\epsilon||E_\alpha|^2 dx + C\|E_\alpha\|\|F_2^\alpha\|.$$

Here $e(E_\alpha) = E_\alpha^* A_0(h^\epsilon,u^\epsilon)E_\alpha$, and C is a generic constant depending only on the range $[a,2b]$ of h^ϵ.

Proof. Since matrices $A_0^\epsilon \equiv A_0(h^\epsilon)$ and $A_0^\epsilon A_j^\epsilon \equiv A_0^\epsilon A_j(h^\epsilon,u^\epsilon)$ are symmetric, we

multiply (4.3) by $E_\alpha^* A_0^\epsilon$ to obtain

$$
\begin{aligned}
&e(E_\alpha)_t + \frac{1}{\epsilon}\sum_{j=1}^{d}(E_\alpha^* A_0^\epsilon A_j^\epsilon E_\alpha)_{x_j} \\
&\quad = \frac{-2}{\epsilon^2}\operatorname{Re} E_\alpha^* A_0^\epsilon \begin{pmatrix} 0 \\ E_\alpha^u \end{pmatrix} + 2\operatorname{Re} E_\alpha^* A_0^\epsilon F_1^\alpha \\
&\qquad + 2\operatorname{Re} E_\alpha^* A_0^\epsilon F_2^\alpha + E_\alpha^* \left\{ \frac{\partial}{\partial t} A_0^\epsilon + \frac{1}{\epsilon}\sum_{j=1}^{d} \frac{\partial}{\partial x_j}(A_0^\epsilon A_j^\epsilon) \right\} E_\alpha \\
&\quad \equiv I_1^\alpha + I_2^\alpha + I_3^\alpha + I_4^\alpha.
\end{aligned} \tag{4.4}
$$

Recall from (2.3) that

$$
A_0^\epsilon = \operatorname{diag}\left(\frac{1}{q(h^\epsilon)}, I_d \right).
$$

It is obvious that

$$
I_1^\alpha = -\frac{2}{\epsilon^2}|E_\alpha^u|^2 \quad \text{and} \quad I_2^\alpha \le 2|E_\alpha^u||F_1^\alpha|.
$$

On the other hand, since h^ϵ takes values in the compact set $[a, 2b]$, I_3^α is simply estimated as

$$
I_3^\alpha \le C|E_\alpha||F_2^\alpha|.
$$

Moreover, we use the relations in (2.3) and the h-equation in (2.1) to compute

$$
\begin{aligned}
&\partial_t A_0(h^\epsilon) + \frac{1}{\epsilon}\sum_{j=1}^{d} \partial_{x_j}(A_0(h^\epsilon)A_j(h^\epsilon, u^\epsilon)) \\
&\quad = A_0'(h^\epsilon)\left(\partial_t h^\epsilon + \frac{1}{\epsilon}\sum_{j=1}^{d} u_j^\epsilon \partial_{x_j} h^\epsilon \right) + \frac{1}{\epsilon}\sum_{j=1}^{d} \partial_{x_j} u_j A_0(h^\epsilon) \\
&\quad = \frac{\operatorname{div} u^\epsilon}{\epsilon}(A_0(h^\epsilon) - q(h^\epsilon)A_0'(h^\epsilon)).
\end{aligned}
$$

Thus, we have

$$
I_4^\alpha \le \frac{C|E_\alpha|^2|\operatorname{div} u^\epsilon|}{\epsilon}.
$$

Now we integrate (4.3) with respect to x over Ω and use the periodicity of the data to conclude the lemma. □

For the right-hand side of the inequality in Lemma 4.2, we have the following claim.

LEMMA 4.3. *Set*

$$
D = D(t) = \frac{\|E(\cdot, t)\|_s}{\epsilon}.
$$

Then, for $\epsilon < 1$,

$$
\begin{aligned}
|\operatorname{div} u^\epsilon| &\le C\epsilon + C\epsilon D,\\
\|E_\alpha^u\|\|F_1^\alpha\| &\le \frac{\|E_\alpha^u\|^2}{4\epsilon^2} + C\epsilon^4 + C(1+D^{2s})\|E^h\|_{|\alpha|}^2,\\
\|F_2^\alpha\| &\le \frac{C(1+D^s)\|E^u\|_{|\alpha|}}{\epsilon} + C(1+D^s)\|E\|_{|\alpha|}.
\end{aligned}
$$

Proof. Recall that

$$
u_\epsilon = \epsilon\nabla\Delta^{-1}(n-\tilde{n}) - \frac{\epsilon\nabla p(n)}{n}. \tag{4.5}
$$

Thus, for $s > s_0 = [d/2]+1$, we use the well-known embedding inequality to obtain

$$
|\operatorname{div} u^\epsilon| \le C\|\operatorname{div} u^\epsilon\|_{s_0} \le C\|\operatorname{div} u_\epsilon\|_{s_0} + C\|\operatorname{div}(u^\epsilon - u_\epsilon)\|_{s_0} \le C\epsilon + \epsilon D.
$$

Next we estimate $\|E_\alpha^u\|\|F_1^\alpha\|$. Since

$$
n(h_\epsilon) - n(h^\epsilon) = E^h \int_0^1 n'(h^\epsilon + \sigma E^h)d\sigma
$$

and the convexity of $[a, 2b]$ gives

$$
h^\epsilon(x,t) + \sigma E^h(x,t) = (1-\sigma)h^\epsilon + \sigma h_\epsilon \in [a, 2b]
$$

for all $(x,t,\sigma) \in \Omega \times [0, \min\{T_\epsilon, T_*\}) \times [0,1]$ and $\epsilon > 0$, it follows from Lemma 3.2 that

$$
\begin{aligned}
\|(n(h^\epsilon + E^h) - n(h^\epsilon))_\alpha\| &\le C\|E^h\|_{|\alpha|}\left\|\int_0^1 n'(h^\epsilon + \sigma E^h)d\sigma\right\|_s\\
&\le C\|E^h\|_{|\alpha|}\int_0^1 \|n'(h^\epsilon + \sigma E^h)\|_s d\sigma\\
&\le C\|E^h\|_{|\alpha|}\int_0^1 (1 + \|h^\epsilon + \sigma E^h\|_s^s)d\sigma\\
&\le C\|E^h\|_{|\alpha|}(1 + \|E^h\|_s^s) \le C\|E^h\|_{|\alpha|}(1+D^s).
\end{aligned} \tag{4.6}
$$

Here the last inequality has used $h^\epsilon + \sigma E^h = (\sigma-1)E^h + h_\epsilon$ and the boundedness of $\|h_\epsilon\|_s = \|h(n)\|_s$ indicated in Lemma 3.1. Thus, we deduce from Proposition 2.1 that

$$
\begin{aligned}
\|E_\alpha^u\|\|F_1^\alpha\| &\le \frac{\|E_\alpha^u\|}{\epsilon}(\epsilon^2\|R_\alpha\| + \|\nabla\Delta^{-1}(n(h_\epsilon) - n(h^\epsilon))_\alpha\|)\\
&\le \frac{\|E_\alpha^u\|}{\epsilon}(C\epsilon^2 + C\|(n(h_\epsilon) - n(h^\epsilon))_\alpha\|)\\
&\le \frac{\|E_\alpha^u\|^2}{4\epsilon^2} + C\epsilon^4 + C(1+D^{2s})\|E^h\|_{|\alpha|}^2.
\end{aligned}
$$

Now we turn to estimate $\|F_2^\alpha\| \le \|f_1^\alpha\| + \|f_2^\alpha\|$ with the help of Lemma 3.2. For f_1^α, from (2.3) we have

$$
A_j^\epsilon - A_j(h_\epsilon, u_\epsilon) = (u_j^\epsilon - u_{j\epsilon})I_{d+1} + ((A_0^\epsilon)^{-1} - A_0^{-1}(h_\epsilon))C_j.
$$

Since $C_j^{11} = 0$ as in (2.3) and $(A_0^\epsilon)^{-1} - A_0^{-1}(h_\epsilon) = \mathrm{diag}(q(h^\epsilon) - q(h_\epsilon), 0)$, it is clear that

$$((A_0^\epsilon)^{-1} - A_0^{-1}(h_\epsilon))C_j(h_\epsilon, u_\epsilon)^T = (q(h^\epsilon) - q(h_\epsilon))O(|u_\epsilon|).$$

Note that

$$\|q(h_\epsilon) - q(h^\epsilon)\|_{|\alpha|} \le C\|E^h\|_{|\alpha|}(1 + D^s)$$

can be proved in the same fashion as that used for (4.6). Thus, we use Lemma 3.2, the boundedness of $\|(h_\epsilon, u_\epsilon)(\cdot, t)\|_{s+1}$ indicated in Lemma 3.1, and (4.5) to conclude that

$$\begin{aligned}\epsilon\|f_1^\alpha\| &\le C\sum_j \|(h_\epsilon, u_\epsilon)_{x_j}\|_s\|u_j^\epsilon - u_{j\epsilon}\|_{|\alpha|} + C\sum_j \|q(h^\epsilon) - q(h_\epsilon)\|_{|\alpha|}\|u_{\epsilon x_j}\|_s \\ &\le C\|E^u\|_{|\alpha|} + C\epsilon(1 + D^s)\|E^h\|_{|\alpha|}.\end{aligned}$$

In the similar spirit, we estimate f_2^α. Since

$$\epsilon f_2^\alpha = \sum_{j=1}^d (u_j^\epsilon E_{\alpha x_j} - (u_j^\epsilon E_{x_j})_\alpha) + \sum_{j=1}^d ((A_0^\epsilon)^{-1}C_j E_{\alpha x_j} - ((A_0^\epsilon)^{-1}C_j E_{x_j})_\alpha)$$

due to (2.3), f_2^α can be bounded as

$$\begin{aligned}\epsilon\|f_2^\alpha\| &\le C\sum_{j=1}^d \|u_j^\epsilon\|_s\|E_{x_j}\|_{|\alpha|-1} + C\sum_{j=1}^d \|q(h^\epsilon)\|_s\|E_{x_j}^u\|_{|\alpha|-1} \\ &\le C(\|u^\epsilon - u_\epsilon\|_s + \|u_\epsilon\|_s)\|E\|_{|\alpha|} + C(1 + \|h^\epsilon\|_s^s)\|E^u\|_{|\alpha|} \\ &\le C\epsilon(1 + D)\|E\|_{|\alpha|} + C(1 + D^s)\|E^u\|_{|\alpha|}.\end{aligned}$$

Hence the estimate on $\|F_2^\alpha\|$ is obtained by putting the above together. This completes the proof. □

Substituting the estimates in Lemma 4.3 into the inequality in Lemma 4.2 yields

$$\frac{d}{dt}\int_\Omega e(E_\alpha)dx + \frac{1}{\epsilon^2}\|E_\alpha^u\|^2 \le C\epsilon^4 + C(1 + D^{2s})\|E\|_{|\alpha|}^2. \tag{4.7}$$

Note that $C^{-1}|E_\alpha|^2 \le e(E_\alpha) \le C|E_\alpha|^2$. We integrate (4.7) from 0 to T with $[0, T] \subset [0, \min\{T_\epsilon, T_*\})$ to obtain

$$\|E_\alpha(T)\|^2 + \frac{1}{\epsilon^2}\int_0^T \|E_\alpha^u(t)\|^2 dt \le CT\epsilon^4 + \int_0^T C(1 + D^{2s})\|E(t)\|_{|\alpha|}^2 dt.$$

Here we have used the fact that the initial data are in equilibrium. Summing up the last inequality over all α satisfying $|\alpha| \le s$, we get

$$\|E(T)\|_s^2 + \frac{1}{\epsilon^2}\int_0^T \|E^u(t)\|_s^2 dt \le CT_*\epsilon^4 + C\int_0^T (1 + D^{2s})\|E(t)\|_s^2 dt. \tag{4.8}$$

We apply Gronwall's lemma to (4.8) to get

$$\|E(T)\|_s^2 \le CT_*\epsilon^4 \exp\left[C\int_0^T \left(1 + D^{2s}\right) dt\right]. \tag{4.9}$$

Since $\|E\|_s = \epsilon D$, it follows from (4.9) that

$$D(T)^2 \le CT_*\epsilon^2 \exp\left[C\int_0^T \left(1 + D^{2s}\right) dt\right] \equiv \Phi(T). \tag{4.10}$$

Thus,

$$\Phi'(t) = C(1 + D^{2s})\Phi(t) \le C\Phi(t) + C\Phi^{s+1}(t).$$

Applying the nonlinear Gronwall-type inequality in [20] to the last inequality yields

$$\Phi(t) \le e^{CT*}$$

for $t \in [0, \min\{T_\epsilon, T_*\})$ if we choose ϵ so small that

$$\Phi(0) = CT_*\epsilon^2 < e^{-CT*}.$$

Because of (4.10), there exists a constant c, independent of ϵ, such that

$$D(T) \le c \tag{4.11}$$

for any $T \in [0, \min\{T_\epsilon, T_*\})$. Finally, the theorem is concluded from (4.9) with (4.11). This completes the proof. □

REFERENCES

[1] K. Bløtekjær, *Transport equations for electrons in two-valley semiconductors*, IEEE Trans. Electron. Devices, 17 (1970), pp. 38–47.

[2] I. Gasser and R. Natalini, *The energy transport and the drift diffusion equations as relaxation limits of the hydrodynamic model for semiconductors*, Quart. Appl. Math., 57 (1999), pp. 269–282.

[3] L. Hsiao and K. Zhang, *The relaxation of the hydrodynamical model for semiconductors to the drift-diffusion equations*, J. Differential Equations, 165 (2000), pp. 315–354.

[4] L. Hsiao and K. Zhang, *The global weak solution and relaxation limits of the initial-boundary value problem to the bipolar hydrodynamical model for semiconductors*, Math. Models Methods Appl. Sci., 10 (2000), pp. 1333–1361.

[5] A. Jüngel, *Quasi-Hydrodynamic Semiconductor Equations*, Birkhäuser, Basel, 2001.

[6] A. Jüngel and Y.-J. Peng, *A hierarchy of hydrodynamic models for plasmas zero-relaxation-time limits*, Comm. Partial Differential Equations, 24 (1999), pp. 1007–1033.

[7] S. Klainerman and A. Majda, *Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math., 34 (1981), pp. 481–524.

[8] C. Lattanzio, *On the 3-D bipolar isentropic Euler-Poisson model for semiconductors and the drift-diffusion limit*, Math. Models Methods Appl. Sci., 10 (2000), pp. 351–360.

[9] C. Lattanzio and P. Marcati, *The relaxation to the drift-diffusion system for the 3-D isentropic Euler-Poisson model for semiconductors*, Discrete Contin. Dynam. Systems, 5 (1999), pp. 449–455.

[10] C. Lattanzio and W.-A. Yong, *Hyperbolic-parabolic singular limits for first-order nonlinear systems*, Comm. Partial Differential Equations, 26 (2001), pp. 939–964.

[11] T. Luo, R. Natalini, and Z. Xin, *Large time behavior of the solutions to a hydrodynamic model for semiconductors*, SIAM J. Appl. Math., 59 (1998), pp. 810–830.

[12] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer-Verlag, New York, 1984.

[13] P. Marcati and A. Milani, *The one-dimensional Darcy's law as the limit of a compressible Euler flow*, J. Differential Equations, 84 (1990), pp. 129–147.

[14] P. Marcati and R. Natalini, *Weak solutions to a hydrodynamic model for semiconductors and relaxation to the drift-diffusion equation*, Arch. Ration. Mech. Anal., 129 (1995), pp. 129–145.

[15] P. Marcati and B. Rubino, *Hyperbolic to parabolic relaxation theory for quasilinear first order systems*, J. Differential Equations, 162 (2000), pp. 359–399.
[16] P. A. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, Vienna, 1990.
[17] R. Natalini, *The bipolar hydrodynamic model for semiconductors and the drift-diffusion equations*, J. Math. Anal. Appl., 198 (1996), pp. 262–281.
[18] F. Poupaud, *Diffusion approximation of the linear semiconductor Boltzmann equation: Analysis of boundary layers*, Asymptotic Anal., 4 (1991), pp. 293–317.
[19] W. van Roosbroeck, *Theory of flow of electrons and holes in germanium and other semiconductors*, Bell System Tech. J., 29 (1950), pp. 560–607.
[20] W.-A. Yong, *Singular perturbations of first-order hyperbolic systems with stiff source terms*, J. Differential Equations, 155 (1999), pp. 89–132.
[21] W.-A. Yong, *Basic aspects of hyperbolic relaxation systems*, in Advances in the Theory of Shock Waves, H. Freistühler and A. Szepessy, eds., Progr. Nonlinear Differential Equations Appl. 47, Birkhäuser Boston, Boston, 2001, pp. 259–305.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1749–1774

SINGULAR INTEGRALS, IMAGE SMOOTHNESS, AND THE RECOVERY OF TEXTURE IN IMAGE DEBLURRING*

ALFRED S. CARASSO†

Abstract. Total variation (TV) image deblurring is a PDE-based technique that preserves edges, but often eliminates vital small-scale information, or *texture*. This phenomenon reflects the fact that most natural images are not of bounded variation. The present paper reconsiders the image deblurring problem in Lipschitz spaces $\Lambda(\alpha, p, q)$, wherein a wide class of nonsmooth images can be accommodated. A new and fast FFT-based deblurring method is developed that can recover texture in cases where TV deblurring fails completely. Singular integrals, such as the Poisson kernel, are used to create an effective new image analysis tool that can calibrate the lack of smoothness in an image. It is found that a rich class of images $\in \Lambda(\alpha, 1, \infty) \cap \Lambda(\beta, 2, \infty)$, with $0.2 < \alpha,\ \beta < 0.7$. The Poisson kernel is then used to regularize the deblurring problem by appropriately constraining its solutions in $\Lambda(\alpha, 2, \infty)$ spaces, leading to new L^2 error bounds that substantially improve on the Tikhonov–Miller method. This so-called *Poisson Singular Integral* or *PSI* method is only one of an infinite variety of singular integral deblurring methods that can be constructed. The method is found to be well-behaved in both the L^1 and L^2 norms, producing results closely matching those obtained in the theoretically optimal, but practically unrealizable, case of true Wiener filtering. Deblurring experiments on synthetically defocused images illustrate the PSI method's very significant improvements over both the total variation and Tikhonov–Miller methods. In addition, successful reconstructions with *inexact* prior Lipschitz space information, highlight the robustness and practicality of the PSI method.

Key words. image deblurring, total variation, nonsmooth images, loss of texture, Lipschitz spaces, Besov spaces, semigroup approximations, singular integrals, Poisson kernel, Gaussian kernel, recovery of texture, Tikhonov–Miller method, true Wiener filtering, Poisson Singular Integral method, PSI method

AMS subject classifications. 94A08, 65R30, 65T60, 47D06

DOI. 10.1137/S0036139903428306

1. Introduction.

The space $BV(R^2)$ of functions of bounded variation, normed by the "total variation" seminorm $\int_{R^2} |\nabla f| dxdy$, plays an important role in much recent work in image analysis. See, e.g., [11], [13], [14], [15], [16], [20], [21], [26], [33], and [39]. In particular, highly successful applications of the total variation approach to image denoising have been well-documented. In contrast, total variation image deblurring is generally not well-behaved, and often results in unacceptable loss of fine scale information. This phenomenon is now believed traceable to an improper choice of function space [24]. The present paper reconsiders the image deblurring problem in Lipschitz spaces $\Lambda(\alpha, p, q)$, wherein a wide class of nonsmooth images can be accommodated. A new and fast FFT-based deblurring technique is developed that can demonstrably recover texture in cases where total variation deblurring fails completely. The approximation properties of certain singular integral operators are intimately linked to such Lipschitz spaces [3], [4], [36]. Here, these properties are

*Received by the editors May 20, 2003; accepted for publication (in revised form) December 16, 2003; published electronically July 2, 2004.
http://www.siam.org/journals/siap/64-5/42830.html

†Mathematical and Computational Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 (alfred.carasso@nist.gov).

exploited in two distinct ways. In the first part of the paper, singular kernels are used to create an effective new FFT-based image analysis tool that can calibrate the lack of smoothness in an image. This tool can be used in contexts unrelated to deblurring, e.g., as a sharpness analysis tool in performance evaluation of imaging systems or image reconstruction software [38], or as a tool for detecting and quantifying "fine-structure" content in images. In the second part of the paper, singular integrals are used as regularization tools in the deblurring problem. Specifically, we show how to stabilize ill-posedness by using the Poisson kernel to impose a priori constraints, in appropriate $\Lambda(\alpha, p, \infty)$ spaces, on the desired nonsmooth deblurred image. This so-called *Poisson Singular Integral* or *PSI* method, is only one of an infinite variety of singular integral deblurring methods that can be constructed. In particular, Gaussian kernels may also be used, leading to the *Gaussian Singular Integral* or *GSI* method. Restricting attention to the case of defocus blurs, we derive L^2 error bounds for the PSI method for images in $\Lambda(\alpha, 2, \infty)$, and demonstrate robust recovery of fine structure in synthetically blurred images. Similar results hold for other types of blurs.

Extensive numerical experiments with known exact solutions indicate that the PSI method is remarkably well-behaved. In both the L^1 and L^2 norms, relative errors in the PSI method are found to closely approximate those obtained in the theoretically optimal, but practically unrealizable, case of *true* Wiener filtering. The latter method requires prior knowledge of the exact power spectra of both the noise and unknown desired sharp image, i.e., a total of $8N^2$ prior data values for a $2N \times 2N$ image. Since the PSI method requires only 4 prior data values, its ability to closely track Wiener filtering is especially noteworthy. The availability of reliable fast deblurring methods is of major significance in the case of nonsmooth images. The true value of the Lipschitz exponent α in the desired sharp image is usually not known in advance, although a plausible range of values for α can often be deduced. In section 10, we document the practicality and robustness of the PSI method, by showing that good quality reconstructions can often be obtained with *inexact*, but plausible, Lipschitz space information. Such initial restorations can then be refined interactively. Here, fast algorithms enable *simultaneous* computation and display of large numbers of trial deblurred images, resulting from multiple choices for α and/or some of the regularization parameters. We stress that the PSI method is exclusively intended for deblurring and is not intended for denoising.

2. Lack of smoothness of images. In [27], a new analytical framework for image processing is introduced, whereby a given image $f(x, y)$ is conceptualized as being the sum of three components, $f(x, y) = u(x, y) + v(x, y) + w(x, y)$. Loosely speaking, $u(x, y)$ contains the edges and the other high-priority information that is sufficient for object recognition, $v(x, y)$ contains the fine-scale details and other low-priority information that is often not necessary for recognition, and $w(x, y)$ represents noise. The $v(x, y)$ component is called *texture*. One example of $v(x, y)$ might be the hair in a photograph of a person's face. Another example of $v(x, y)$ might be the heat-shield tiles in an image of the Columbia space shuttle. The ability to resolve individual hairs is generally not necessary for identification. In several image processing tasks, such as compression, segmentation, or face recognition, this texture component can often be neglected. However, there are numerous other situations, such as in the ill-fated Columbia episode, where $v(x, y)$ may be of paramount interest. It is shown in [27] that only the $u(x, y)$ component can generally be expected to lie in $BV(R^2)$. In [24], it is proved that most natural images are *not* of bounded variation, because the texture component $v(x, y)$ generally has infinite total variation.

Denoising and deblurring are two basic image processing tasks where total variation restoration has been extensively applied. Such restoration can be accomplished most effectively by solving an initial value problem for an appropriate nonlinear anisotropic diffusion equation, using the stepwise marching scheme described in [26]. In deblurring, one typically starts with a degraded image $g(x,y)$ which differs from the desired true image $f(x,y)$ in that the $u(x,y)$ component is blurred but recognizable, the $v(x,y)$ component is seriously attenuated and often not recognizable, and the $w(x,y)$ component is usually small. Reconstructing $v(x,y)$ while keeping $w(x,y)$ small, is the prime objective in numerous medical, astronomical, industrial, and scientific contexts [9], [10]. However, while total variation deblurring sharpens $u(x,y)$ and keeps $w(x,y)$ small, the texture component $v(x,y)$ is often eliminated due to the "staircase effect" [13], [20], [29], [30], [39]. This is in accordance with the analyses in [24], [27].

Let $x = (x_1, x_2) \in R^2$. Postulating $f(x) \in BV(R^2)$ means that $f(x)$ is constrained to satisfy,

$$\int_{\mathbf{R^2}} |f(x+h) - f(x)|dx \leq Const\ |h|. \tag{1}$$

However, from the standpoint of modeling texture, it is advantageous to consider functions $f(x)$ satisfying weaker constraints, such as

$$\left\{\int_{\mathbf{R^2}} |f(x+h) - f(x)|^p dx\right\}^{1/p} \leq Const\ |h|^\alpha, \qquad 0 < \alpha < 1. \tag{2}$$

Such an f lies in $\Lambda(\alpha, p, \infty)$. With $0 < \alpha < 1,\ 1 \leq p < \infty$, the Lipschitz (Besov) spaces $\Lambda(\alpha, p, q)$ [36], [37], consist of the class of functions $f(x) \in L^p(R^2)$ with finite seminorm $\parallel f \parallel_{\alpha pq}$, where

$$\parallel f \parallel_{\alpha pq} = \left\{\int_{R^2} \left(|h|^{-\alpha} \parallel f(x+h) - f(x) \parallel_p\right)^q dh/|h|^2\right\}^{1/q}, \qquad 1 \leq q < \infty, \tag{3}$$

$$\parallel f \parallel_{\alpha p\infty} = \sup_{h \in R^2} \left\{|h|^{-\alpha} \parallel f(x+h) - f(x) \parallel_p\right\}, \qquad q = \infty. \tag{4}$$

For given p and q, functions with larger values of α are better behaved, or "smoother", than functions with smaller values of α, and functions in $\Lambda(\alpha, p, q_1)$ are smoother than functions in $\Lambda(\alpha, p, q_2)$ if $q_1 < q_2$. In fact, the following continuous embedding results are proved in [37, Theorem 9]

$$\Lambda(\alpha_2, p, q_1) \subset \Lambda(\alpha_1, p, q_2), \qquad 0 < \alpha_1 \leq \alpha_2 < 1;\ 1 \leq q_1 \leq q_2 \leq \infty. \tag{5}$$

Also, in R^2,

$$\Lambda(\alpha, p, q) \subset \Lambda(\beta, r, q), \qquad \alpha - 2/p = \beta - 2/r,\ \ p \leq r. \tag{6}$$

Let $r = 2$, let the pair (α, p) satisfy $2/(1+\alpha) < p \leq 2$, and let $\beta = 1 + \alpha - 2/p$. Then, $0 < \beta \leq \alpha$, and it follows from (5) and (6) that

$$\Lambda(\alpha, p, q) \subset \Lambda(\beta, 2, q) \subset \Lambda(\beta, 2, \infty) \subset L^2(R^2). \tag{7}$$

This result will be important in what follows.

For given fixed p with $1 < p < \infty$ and $q = 2p/(2+\alpha p)$, a class of Lipschitz spaces $\Lambda(\alpha, q, q) \subset L^p(R^2)$ is considered in [18], [19], and shown to contain common types of images. A method for empirically estimating image smoothness is developed in [18], [19], based on analyzing the behavior of *lossy wavelet compression* of the image $f(x, y)$. In [12], the spaces $\Lambda(\alpha, q, q) \subset L^2(R^2),\ q = 2/(1+\alpha)$, are advocated as being particularly appropriate for accommodating a rich variety of real images in an L^2 setting. Lossy wavelet compression is again used to estimate image smoothness, and values of α in the range $0.4 < \alpha < 0.75$ are reported in [12] for a class of 24 test images $\in \Lambda(\alpha, \frac{2}{1+\alpha}, \frac{2}{1+\alpha})$. Such α values are an indication of true image smoothness only when the image is largely *noise free.* If the noise component $w(x, y)$ is not sufficiently small, artificially low values of α must be expected. A basic limitation of the above wavelet compression approach is the restriction on q, which precludes consideration of the larger spaces $\Lambda(\alpha, p, \infty) \supset \Lambda(\alpha, p, q)$.

The present independent method of estimating image smoothness rests on an entirely different analytical basis, and requires neither wavelet expansions nor image compression. Instead, the method uses fast FFT algorithms to convolve the image with a specific type of kernel, and then analyzes how well this convolved image approximates the original image as the kernel approaches the Dirac δ-function. This simple direct approach permits consideration of the spaces $\Lambda(\alpha, p, \infty),\ 1 \le p < \infty$. The results obtained here are compatible with those obtained in [18], [19], [12], and [24]. We indeed find that most natural images are not of bounded variation, and that a rich variety of images $\in \Lambda(\alpha, 1, \infty)$ with $0.2 < \alpha < 0.7$.

Remark 1. We deal with high resolution images $f(x, y)$ of size 512×512 or 1024×1024 pixels. Such an $f(x, y)$ may be viewed as a piecewise constant or trigonometric polynomial approximation to the original intensity field $f^\infty(x, y)$, or as some other kind of finite dimensional representation of the infinite dimensional object f^∞. All norms are equivalent on a finite dimensional space. Hence, even if $f^\infty(x, y)$ is not of bounded variation, the discrete total variation norm for $f(x, y)$ is always finite, though it may be very large. To estimate smoothness properties of $f^\infty(x, y)$ by examination of the finite dimensional representation $f(x, y)$ requires some sagacity. In [18, section 4B, section 5B], the authors stress that in their method of estimating the value of α by monitoring the rate of convergence as a function of the number $\mathcal{N}$ of nonzero wavelet coefficients, one must restrict attention to *low values* of $\mathcal{N}$. At high values of $\mathcal{N}$, the fact that $f(x, y)$ is actually piecewise constant causes the error to decrease much too rapidly, resulting in an *artificially high* reading for α that diverges from true behavior in $f^\infty(x, y)$. This same finite dimensionality pitfall occurs in the present approach, but wears a different guise. See Remark 2 and the discussion surrounding Figures 1 and 2 below.

We shall use the spaces $\Lambda(\alpha, 1, \infty)$ and $\Lambda(\alpha, 2, \infty)$ for examining and classifying image smoothness. However, deblurring applications will be limited to the spaces $\Lambda(\beta, 2, \infty) \subset L^2(R^2)$, wherein all spaces $\Lambda(\alpha, p, q),\ 2/(1+\alpha) < p \le 2,\ 1 \le q \le \infty$, are continuously embedded. The spaces $\Lambda(\alpha, 2, \infty)$ will be shown to contain a rich and significant class of images.

3. The spaces $\Lambda(\alpha, p, q)$ and the Poisson singular integral. Define the Fourier transform $\hat{h}(\xi, \eta)$ of $h(x, y) \in L^1(R^2)$ by

$$\mathcal{F}\{h\} = \hat{h}(\xi, \eta) \equiv \int_{R^2} h(x, y) e^{-2\pi i(\xi x + \eta y)} dx dy. \tag{8}$$

For each fixed $t > 0$, consider the Poisson kernel in R^2

$$\psi(x, y, t) = \frac{t}{2\pi(x^2 + y^2 + t^2)^{3/2}}, \qquad (x, y) \in R^2. \tag{9}$$

We have

$$\hat{\psi}(\xi, \eta, t) = e^{-t\rho}, \qquad \rho = \sqrt{\xi^2 + \eta^2}. \tag{10}$$

For each $t > 0$, define the linear operator U^t on $L^p(R^2),\ 1 \le p < \infty$, by

$$U^t f = \int_{R^2} \psi(u, v, t) f(x - u, y - v) du dv. \tag{11}$$

It can be shown that $\lim_{t \downarrow 0} \parallel U^t f - f \parallel_p = 0$. Moreover, defining U^0 to be the identity operator, we have that for $s, t \ge 0,\ U^t U^s = U^{t+s}$. In fact, $\{U^t\}_{t \ge 0}$ is a holomorphic contraction semigroup on $L^p(R^2)$. See [3]. We may write $U^t = e^{-tA}$, where $-A$ is the infinitesimal generator of U^t. Here, A corresponds to the fractional differential operator $(-\Delta)^{1/2}$. Note that for $t > 0,\ U^t$ maps $L^p(R^2)$ into $D(A)$, so that $AU^t f$ is well-defined for arbitrary $f \in L^p$. In general, this is not the case for nonholomorphic semigroups. The Gauss singular integral, where the two-dimensional Gaussian kernel is used in lieu of ψ in (9), defines an analogous holomorphic semigroup W^t, with $A = -\Delta$. Many such singular integral semigroups S^t exist. A very rich variety can be constructed by *subordination* [7], [40]. For small $t > 0,\ S^t$ behaves as an approximate identity on L^p. There is a large literature on how well $S^t f$ approximates f as $t \downarrow 0$. See [3], [4], [5], [35], [36], [37], and the references therein. As $t \downarrow 0$, we have $\parallel S^t f - f \parallel_p = o(1)$ for arbitrary $f \in L^p$, $\parallel S^t f - f \parallel_p = O(t)$ if and only if $f \in D(A)$, and $\parallel S^t f - f \parallel_p = o(t)$ if and only if $S^t f = f$ for all $t \ge 0$. Thus, the optimal rate is always $O(t)$. Of particular interest in this paper is the case of *nonoptimal* approximation, where $f \notin D(A)$ yet retains sufficient smoothness that $\parallel S^t f - f \parallel_p = O(t^\alpha),\ 0 < \alpha < 1$, as $t \downarrow 0$. While complete theories exist for a wide class of singular kernels, the simplest such theory revolves around the Poisson semigroup U^t in (11). We have from [37, Theorem 4], the following result.

THEOREM 1. *Let $U^t,\ t > 0$, be the Poisson integral operator in* (11)*, and let $0 < \alpha < 1,\ 1 \le p, q < \infty$. Then, $f \in \Lambda(\alpha, p, q)$ if and only if*

$$\int_0^\infty \left(t^{-\alpha} \parallel U^t f - f \parallel_p\right)^q dt/t \ < \ \infty. \tag{12}$$

For $q = \infty$, we have $f \in \Lambda(\alpha, p, \infty)$ if and only if

$$\sup_{t>0} \ t^{-\alpha} \parallel U^t f - f \parallel_p \ < \ \infty. \tag{13}$$

Using the embedding results in (7) together with (13) leads to the following corollary.

THEOREM 2 (corollary). *Let $f \in \Lambda(\alpha, p, q)$, with $2/(1 + \alpha) < p \le 2$, and let $\beta = 1 + \alpha - 2/p$. Then, in the L^2 norm*

$$\sup_{t>0} \ t^{-\beta} \parallel U^t f - f \parallel_2 \ < \ \infty. \tag{14}$$

4. Periodized problems, the Poisson summation formula, and FFT algorithms. The above results can be used to fashion a practical image analysis tool. Theoretically, given any image $f(x,y)$ in $L^1(R^2)$, one could use the Fourier transform (8) to form

$$\mathcal{F}\left\{U^t f\right\} = e^{-t\rho}\hat{f}(\xi,\eta), \qquad \rho = \sqrt{\xi^2+\eta^2}, \tag{15}$$

for sequences of positive t-values tending to zero. Inverse transformation is always possible on account of the factor $e^{-t\rho}$, and this can be used to produce an infinite sequence of positive numbers $\mu_n = \{\| U^{t_n} f - f \|_1 / \| f \|_1\}$ with $t_n \downarrow 0$. If every such sequence (t_n, μ_n), ultimately lies below the curve $\mu(t) = C\ t^\alpha,\ 0 < t \leq \bar{t}$, for suitably chosen constants $C > 0$ and $0 < \alpha < 1$, then $\| U^t f - f \|_1 \leq C \| f \|_1\ t^\alpha$, as $t \downarrow 0$, and $f(x,y) \in \Lambda(\alpha, 1, \infty)$ by Theorem 1. However, this does not lead to a practical procedure.

On the other hand, Theorems 1 and 2 remain valid in the periodic case [36], [37]. Here, the image $f(x,y)$ and the kernel $\psi(x,y,t)$ in (9) are now periodized [5], [6]. Let Ω denote the unit square $-1/2 < x, y \leq 1/2$ in R^2. The image $f(x,y)$ is now viewed as originally defined on Ω from which it is extended by periodicity to all of R^2. Let

$$\hat{f}(\xi,\eta) = \int_\Omega f(x,y) e^{-2\pi i(\xi x + \eta y)} dxdy. \tag{16}$$

Define the periodized Poisson kernel $\psi^*(x,y,t)$ by

$$\psi^*(x,y,t) = \sum_{k,m=-\infty}^{\infty} \psi(x+k, y+m, t), \qquad t > 0, \quad (x,y) \in R^2, \tag{17}$$

and let

$$U^t f = \int_\Omega \psi^*(u,v,t) f(x-u, y-v) du dv, \qquad t > 0. \tag{18}$$

The Poisson summation formula, [1], [5], [6], [22], [37], can be used to show that the periodized Poisson kernel has a complex Fourier series with Fourier coefficients again given by (10), but where ξ, η are now *integers* running from $-\infty$ to $+\infty$. Moreover,

$$U^t f = \sum_{\xi,\eta=-\infty}^{\infty} e^{-t\rho}\hat{f}(\xi,\eta) e^{2\pi i(x\xi + y\eta)}, \qquad t > 0, \quad \rho = \sqrt{\xi^2+\eta^2}. \tag{19}$$

Again the factor $e^{-t\rho}$ assures uniform convergence of the Fourier series in (19). Let

$$f_N(x,y) = \sum_{\xi,\eta=-N}^{N} e^{-t\rho}\hat{f}(\xi,\eta) e^{2\pi i(x\xi + y\eta)}, \qquad t > 0, \quad \rho = \sqrt{\xi^2+\eta^2}. \tag{20}$$

Since $L^p(\Omega) \subset L^1(\Omega)$, $p > 1$, we may apply this approach to any $f \in L^p$, and $\| U^t f - f_N \|_p$ can be made arbitrarily small by choosing N large enough in (20). Next, given the $2J \times 2J$ digitized image $f(x,y)$ with $J > N$, the discrete Fourier transform [2] is now the appropriate numerical tool for analyzing this periodized problem. One can use FFT algorithms to form the Fourier coefficients $\hat{f}(\xi,\eta)$, $-J \leq \xi, \eta \leq J$, and then apply the filter $(e^{-t\rho} - 1)$ as in (15). An inverse FFT then yields an accurate approximation to $U^t f - f$ at each of the $2J \times 2J$ pixels, for each small $t > 0$. We

may then examine the discrete L^p relative error in Poisson approximation as $t \downarrow 0$, and locate constants C and α such that $\| U^t f - f \|_p \leq C \| f \|_p t^\alpha,\ 0 < t \leq \bar{t}$. In summary, we have constructed an accurate numerical procedure, based on correct mathematical analysis, for assessing membership in any $\Lambda(\alpha, p, \infty)$ space. Equally important, the values of C and α constitute a priori information that will be useful in stabilizing the ill-posed deblurring problem.

Remark 2. Analogously to the case of lossy wavelet compression discussed in Remark 1, there is a finite dimensionality pitfall in the above singular integral methodology that *necessitates the exclusion of very small values of* $t > 0$. Let $f^\infty(x, y)$ be the original image intensity field as in Remark 1, and assume that $f^\infty(x, y) \in \Lambda(0.5, p, \infty)$, so that $\| U^t f^\infty - f^\infty \|_p = O(\sqrt{t})$ as $t \downarrow 0$, by Theorem 1. Let $f(x, y)$ be the $2J \times 2J$ digitized image corresponding to $f^\infty(x, y)$. We shall show that at very small values of $t > 0$, the behavior of $\| U^t f - f \|_p$ *diverges* from true behavior in $f^\infty(x, y)$, resulting in a *false reading* for α. Let $S^t = e^{-tA}$ be any contraction semigroup on $L^p(R^2)$. As already pointed out, if $f \in D(A)$, $\| S^t f - f \|_p = O(t)$ as $t \downarrow 0$. This follows from

$$S^t f - f = \int_0^t \frac{d}{du}(S^u f) du = -\int_0^t S^u A f du, \tag{21}$$

so that $\| S^t f - f \|_p \leq t \| Af \|_p$, for all $t > 0$. In addition, $\| S^t f - f \|_p \approx t \| Af \|_p$, for all sufficiently small $t > 0$, because $S^u Af \approx Af$ for all sufficiently small u. In the above Poisson semigroup U^t, the unbounded operator A is defined as follows in Fourier space

$$\mathcal{F}\{Af\} = \rho \hat{f}(\xi, \eta), \qquad \rho = \sqrt{\xi^2 + \eta^2}. \tag{22}$$

Since the digitized $2J \times 2J$ image $f(x, y)$ is a trigonometric polynomial, it is always $\in D(A)$ and $\| Af \|_p$ is always finite, although it may be very large. Consequently, with a possibly large positive constant K, we always have $\| U^t f - f \|_p \leq Kt$ for all $t > 0$, as well as actual linear behavior $\| U^t f - f \|_p \approx Kt$ for all sufficiently small t, irrespective of the behavior of $\| U^t f^\infty - f^\infty \|_p$ at these same values of t. This phenomenon is well-illustrated in Figures 1 and 2 below.

5. Application to real images. The following examples illustrate the use of the Poisson singular integral approach. Our first example, in Figure 1, is the 512×512 Mandrill image highlighted in [24] as an example of an image $\notin BV(R^2)$. The above FFT procedure was used to obtain the L^1 and L^2 relative errors in Poisson approximation

$$\mu(t) = \| U^t f - f \|_p \,/\, \| f \|_p, \qquad p = 1, 2, \tag{23}$$

at 300 values of t given by $t_n = 0.5(0.95)^n,\ n = 1, 300$. For the L^1 norm, a plot of $\mu(t)$ versus t on a log-log scale produced the solid curve A in Figure 1. Least squares fitting was used to find the two distinct majorizing dashed straight lines Γ and Σ. For each dashed line, the y-axis intercept value obtained by least squares was slightly increased so as to make each line lie visibly above the solid curve A; however, the slope of each line remains the same as that obtained from least squares. The line Γ, defined by $\log \mu(t) = 3.2 + 0.994 \log t$, accurately captures the misleading linear trend in (23) for very small values of t, while being grossly inaccurate at larger values of t. It was obtained by excluding data corresponding to $\log t > -7$ from the least squares fit. The line Γ implies that $\| U^t f - f \|_1 < 24.53 \| f \|_1 t^{0.994}$

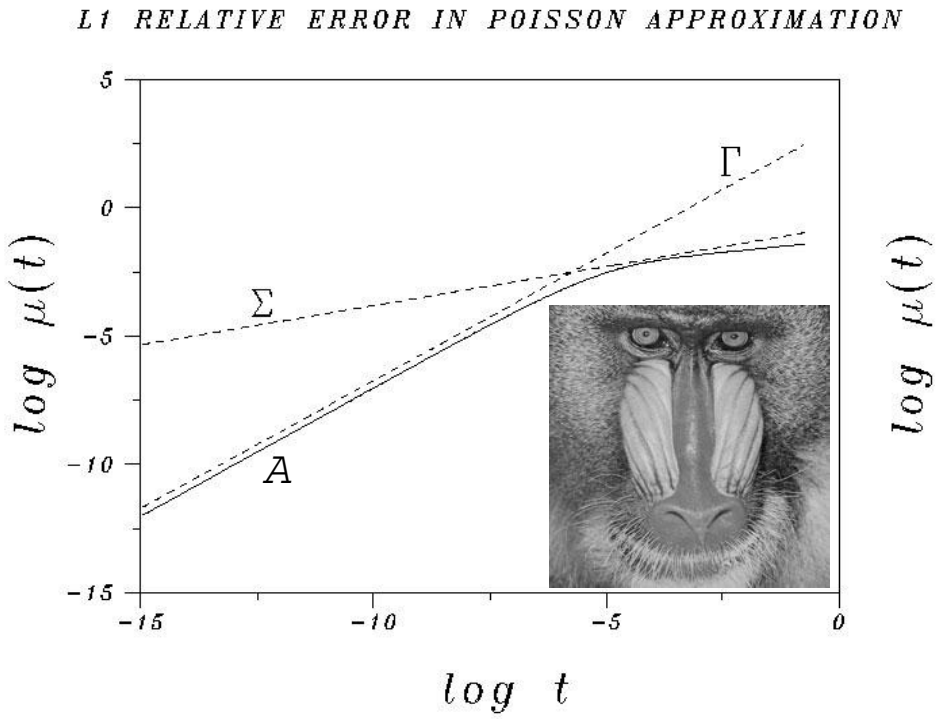


FIG. 1. 512×512 *Mandrill image was identified in* [24] *as not in* $BV(R^2)$. *This is confirmed in above graphical use of Theorem* 1, *using FFT techniques discussed in section* 4. *Solid curve A is a plot of* $\mu(t) = \| U^t f - f \|_1 / \| f \|_1$ *versus* t, *on a log-log scale. Majorizing dashed straight line* Γ, *defined by* $\log \mu(t) = 3.2 + 0.994 \log t$, *accurately captures linear behavior in* (23) *for very small values of* t, *but is grossly inaccurate at larger values of* t. *Linear behavior at very small* t *is misleading, and falsely implies that image is of bounded variation. (See Remarks* 1 *and* 2*). Majorizing dashed straight line* Σ, *defined by* $\log \mu(t) = -0.75 + 0.306 \log t$, *accurately reflects behavior for* $-6 \le \log t \le -1$, *while being grossly inaccurate at very small* t. *Behavior along* Σ *is taken to be true behavior in Mandrill image, implying image* $\in \Lambda(0.306, 1, \infty)$ *with* $\| U^t f - f \|_1 \le 0.472 \| f \|_1 t^{0.306}$, $0 < t \le 0.1$.

for all $t > 0$. As emphasized in Remark 2, this correct statement primarily reflects the fact that the 512×512 Mandrill image lies in a finite dimensional space; it *does not* describe the smoothness properties of the intensity field $f^{\infty}(x, y)$ that gave rise to the digitized Mandrill image. The majorizing dashed straight line Σ, defined by $\log \mu(t) = -0.75 + 0.306 \log t$, accurately reflects behavior of (23) for $-6 \le \log t \le -1$, while being grossly inaccurate at very small values of t. The line Σ was obtained by excluding all data corresponding to $\log t < -6$ from the least squares fit. Note that this still leaves over 100 data points remaining. The behavior along Σ indicates that $\| U^t f - f \|_1 \le 0.472 \| f \|_1 t^{0.306}$, $0 < t \le 0.1$, and this is taken to be the true behavior in the Mandrill image. From (13), this implies that the Mandrill image $\in \Lambda(0.306, 1, \infty)$, and hence, is not of bounded variation. The behavior in the L^2 norm is strikingly similar, and indicates the image $\in \Lambda(0.271, 2, \infty)$. Estimates of α in any other discrete L^p norm can be obtained similarly. All α estimates shown in this paper were obtained using the above procedure of constructing the line Σ in log-log plots of $\mu(t)$, after excluding all data corresponding to $\log t < -6$. As in [18,

section 5B], we have occasionally found contradictory examples where the value of α in the L^2 norm was greater than that in the L^1 norm. When that happened, a new Σ line was constructed for the L^2 trace, based on excluding data corresponding to $\log t < -5$. It is recommended that data for very small values of t always be included in log-log plots of $\mu(t)$, so as to enable clear identification of the spurious linear trend, prior to rejecting that part of the data.

Our second example, in Figure 2(A), is a 1024×1024 Whirlpool galaxy image, taken at the National Optical Astronomy Observatory, (NOAO/AURA/NSF), by T. Rector and M. Ramirez. As in the case of Figure 1, Poisson integral approximation in L^1 was used to obtain the solid curve A, and the line Σ_A was constructed using least squares. This procedure was repeated for the L^2 norm. The results indicate that Figure 2(A) satisfies $\| U^t f - f \|_1 \leq 0.6 \| f \|_1 t^{0.530}$, $0 < t \leq 0.1$, and that Figure 2(A) $\in \Lambda(0.530, 1, \infty) \cap \Lambda(0.462, 2, \infty)$. Interestingly, if we sharpen Figure 2(A) using the *APEX method* [9], we obtain the image in Figure 2(B). This enhanced image displays significant fine scale detail not readily visible in the original image, and strongly resembles a Whirlpool galaxy plate taken by Milton Humason in 1950 using the 200 inch Mt. Palomar telescope. See [34, plate 26]. Here, L^1 Poisson analysis produced the solid curve B and the majorizing line Σ_B. We find that Figure 2(B) $\in \Lambda(0.239, 1, \infty) \cap \Lambda(0.230, 2, \infty)$, and thus has substantially *lower* values of α than does Figure 2(A). This result is highly plausible. Presumably, any low-pass blurring process that may have affected Figure 2(A) would have attenuated fine scale features, and thereby *increased* the values of α. The result also indicates that APEX processing of image (A) produced relatively more sharpening in the L^1 norm than in the L^2 norm.

The nine images in Figure 3 and Table 1 form an interesting collection that includes natural as well as man made objects, exhibiting a wide range of sizes. The last row contains a nanoscale electron microscopy micrograph, a galactic scale object, and a cosmological scale structure. Along with the three images in Figures 1 and 2, this paper has applied the Poisson integral method to 12 high resolution images, and we have found that, in either $\Lambda(\alpha, 1, \infty)$ or $\Lambda(\alpha, 2, \infty)$, the values of α lie in the range $0.2 < \alpha < 0.7$. This range of values is compatible with that found in [12], [18], [19], using an entirely different method. Moreover, while $\Lambda(\alpha, 2, \infty)$ are smaller spaces than are $\Lambda(\alpha, 1, \infty)$, they are evidently wide enough to contain each of these 12 images, albeit with smaller values of α. Notice also that the values of the constant C in Table 1 are confined to a very narrow range in both L^1 and L^2.

Remark 3. Following [18], [19], the values of C and α reported in Table 1 and elsewhere in this paper, are given to two and three decimal places. Such precision must be viewed with skepticism. These values are based on the use of the Poisson operator U^t, together with the particular sequence $t_n = 0.5(0.95)^n$, $n = 1, 300$. Also, the specific interval $-6 \leq \log t \leq 0$ was chosen for the least squares fit to $\log \mu(t)$. However, different sequences t_n tending to zero might be used, as well as slightly larger or slightly smaller t-intervals for the least squares fit. In addition, the Gaussian operator G^t may be used in place of U^t, in which case the image Lipschitz exponent $\alpha = 2\delta$, where δ is the slope of the corresponding Σ line. It is found that such variations in the basic methodology result in slightly different values for C and α. For this reason, all reported values of C and α should probably be *rounded to one decimal place*. It may not be feasible to determine true image Lipschitz space parameters to higher place accuracy.

The PSI deblurring method to be described in section 7 below requires prior knowledge of the values of C and α in the desired unknown deblurred image. In

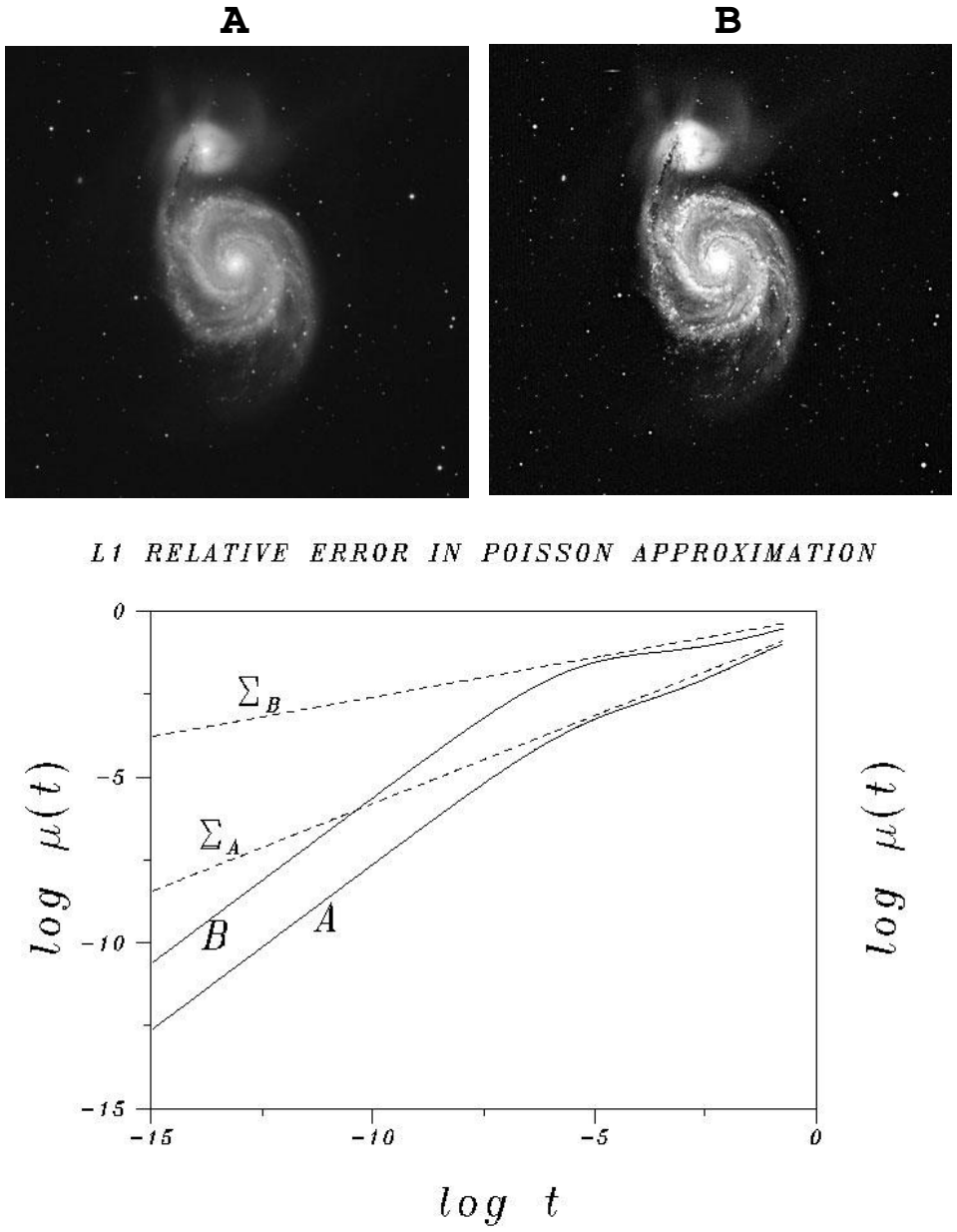


FIG. 2. *Whirlpool galaxy M*51. *Original and enhanced images have noticeably different* L^1 *Poisson traces* $\mu(t) = \| U^t f - f \|_1 / \| f \|_1$, *reflecting sharply distinct Lipschitz exponents.* (A) *Original* 1024×1024 *image taken by T. Rector and M. Ramirez, National Optical Astronomy Observatory, (NOAO/AURA/NSF).* L^1 *Poisson relative error* $\mu(t)$, *shown in solid trace A, is majorized by dashed straight line* Σ_A *defined by* $\log \mu(t) = -0.5 + 0.530 \log t$. *This implies that image* (A) $\in \Lambda(0.530, 1, \infty)$. (B) *Blind deconvolution of* (A) *using APEX method* [9], *brings out significant fine scale detail, and results in solid trace* B, *majorized by dashed straight line* Σ_B *defined by* $\log \mu(t) = -0.2 + 0.239 \log t$. *This indicates that deblurred image* (B) $\in \Lambda(0.239, 1, \infty)$. *Image* (B) *strongly resembles* [34, plate 26] *taken by Milton Humason using* 200 *inch Mt. Palomar telescope. For* $\log\ t < -7$, *solid traces* A *and* B *have identical slopes of* 0.994. *This confirms the observation in Remark* 2 *that behavior at very small* t *is artificial and disconnected from true image smoothness.*

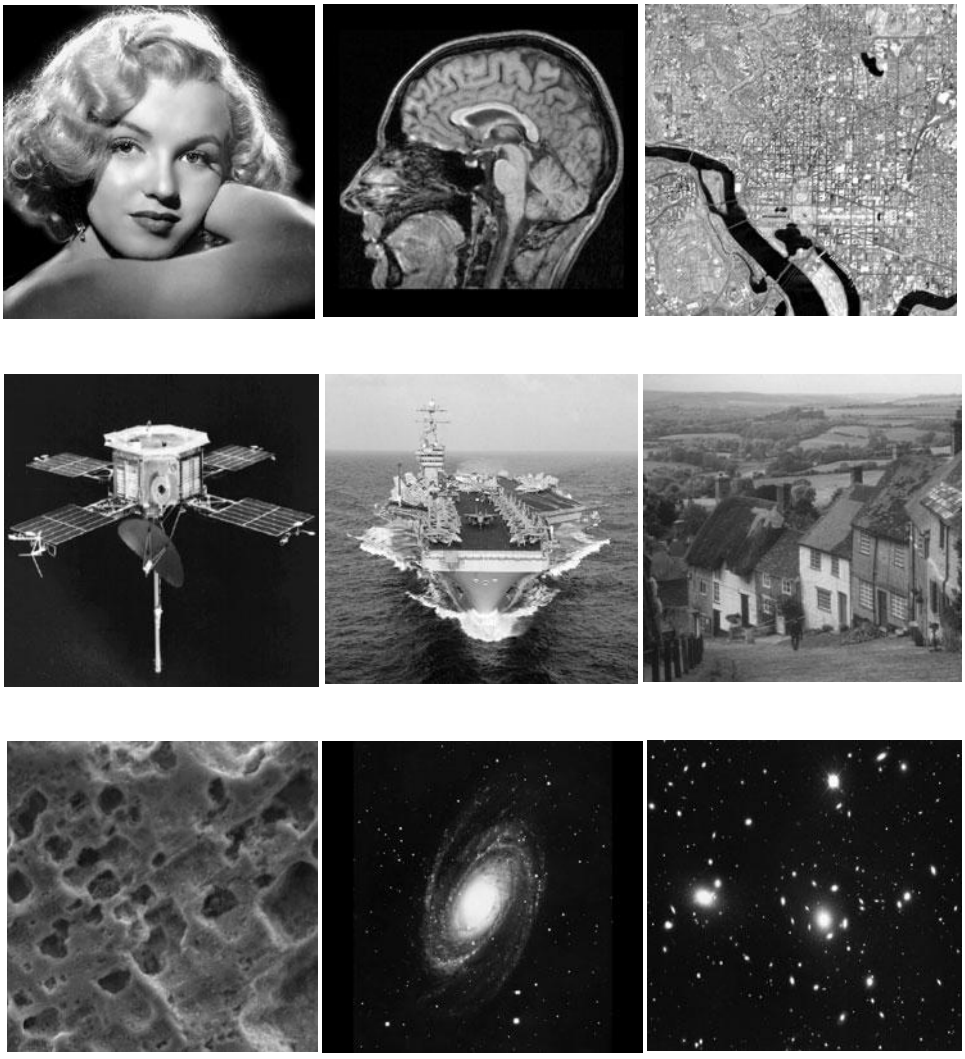


FIG. 3. *A significant class of high-resolution* 8*-bit images have Lipschitz exponents* α *in the range* $0.2 < \alpha < 0.7$*, in either* L^1 *or* L^2*, and are not of bounded variation.*

general, these values will not be known. However, as shown in Figure 2, it is reasonable to assume that α in the deblurred image will be *lower* than in the given blurred image, provided that image is relatively noise free. Inspection of Table 1, or of other more extensive tables pertaining to the types of images under consideration, may indicate plausible initial estimates for (C, α). As will be shown in section 10, the PSI method is sufficiently robust as to produce good quality reconstructions, even with *inexact* Lipschitz data. Moreover, given a fast algorithm, because of the narrow range of values involved in both C and α, it is feasible to refine such reconstructions by simultaneous computation and display of multiple trial restorations, based on neighboring values of (C, α). Efficient exploration in parameter space is usually the key to the successful solution of inverse problems, when such problems can be solved.

TABLE 1

Values of $(C,\ \alpha)$ *in* $\| U^t f - f \|_p \le C \| f \|_p t^\alpha,\ 0 < t \le 0.1,\ p = 1, 2$, *for each image* $f(x, y)$ *in Figure* 3, *when* U^t *is Poisson operator in* (19). *The* (C, α) *values shown below may be more meaningful if rounded to one decimal place. See Remark* 3.

Image	*Size*	$(C,\ \alpha)\ \in \Lambda(\alpha, 1, \infty)$	$(C,\ \alpha)\ \in \Lambda(\alpha, 2, \infty)$
Marilyn Monroe	512^2	$C = 0.77,\ \ \alpha = 0.565$	$C = 0.68,\ \ \alpha = 0.474$
Sagittal brain MRI	512^2	$C = 1.28,\ \ \alpha = 0.590$	$C = 1.02,\ \ \alpha = 0.520$
Washington DC Landsat	512^2	$C = 0.45,\ \ \alpha = 0.341$	$C = 0.55,\ \ \alpha = 0.340$
Mariner 5 spacecraft	512^2	$C = 0.90,\ \ \alpha = 0.448$	$C = 0.99,\ \ \alpha = 0.417$
USS Eisenhower	512^2	$C = 0.47,\ \ \alpha = 0.420$	$C = 0.50,\ \ \alpha = 0.362$
English village	512^2	$C = 0.49,\ \ \alpha = 0.472$	$C = 0.55,\ \ \alpha = 0.439$
Nanoscale micrograph	1024^2	$C = 0.45,\ \ \alpha = 0.415$	$C = 0.55,\ \ \alpha = 0.415$
Spiral galaxy M81	1024^2	$C = 0.68,\ \ \alpha = 0.365$	$C = 0.78,\ \ \alpha = 0.327$
Cluster of galaxies	1024^2	$C = 0.65,\ \ \alpha = 0.222$	$C = 0.97,\ \ \alpha = 0.216$

6. Image deblurring in $L^2(R^2)$. We now consider the image deconvolution problem $Pf = g$ with a known *shift-invariant* point spread function (psf) $p(x, y)$,

$$Pf \equiv p(x, y) \otimes f(x, y) = g(x, y), \qquad g(x, y) = g_e(x, y) + n(x, y). \tag{24}$$

Here, $\otimes$ denotes convolution, $g(x, y)$ is the given recorded noisy blurred image, $g_e(x, y)$ is the hypothetical exact blurred image that would have been recorded in the absence of any noise, and $n(x, y)$, presumed small, represents the cumulative effects of all noise processes and other errors affecting final acquisition of the digitized array $g(x, y)$. The noise may be *multiplicative*. Neither $g_e(x, y)$ nor $n(x, y)$ are known, only their sum $g(x, y)$. Denoting the unknown exact sharp image by $f_e(x, y)$, we have

$$Pf_e = p(x, y) \otimes f_e(x, y) = g_e(x, y). \tag{25}$$

Given only (24), we seek a solution $f(x, y)$ in (24) such that $Pf \approx g$, *and* such that $\| f - f_e \|_2$ is small. To achieve this goal, some a priori information about f_e and n is always necessary. Most real images $f_e(x, y)$ contain fine scale features, sharp edges, and other kinds of nondifferentiable singularities. Deblurring techniques that impose stabilizing constraints in the form of prescribed bounds on partial derivatives of $f(x, y)$ in (24), are generally inapplicable, although they are often used. Penalties for such use include smoothing out of sharp features, and possible loss of vital diagnostic information. Indeed, the desire to accurately reconstruct edges and other sharp singularities was the principal reason for developing total variation methods. In fact, several deblurring methods actually exist that do not require prescribed bounds on derivatives [8].

A wide variety of blurs can be used as illustrative examples in (24). Here, we consider the case of uniform defocus blur, where the psf is proportional to the characteristic function of a disc of radius R. This is the so-called "pillbox" model [17], [25], [16], [31]. If $R > 0$ is the radius of the "circle of confusion", the psf for defocus blur is given by

$$p(x, y) = \begin{cases} (\pi R^2)^{-1}, & x^2 + y^2 \le R^2, \\ 0, & x^2 + y^2 > R^2. \end{cases} \tag{26}$$

This has a Fourier transform given by the "sombrero function" [23, p. 72]

$$\hat{p}(\xi, \eta) = 2J_1(R\rho)/(R\rho), \qquad \rho = \sqrt{\xi^2 + \eta^2}, \tag{27}$$

where $J_1(x)$ is the Bessel function of the first kind of order 1. In our numerical experiments below on $2N \times 2N$ images, the expression (27) is used to blur images by Fourier domain multiplication with a preselected $R > 0$, and (ξ, η) are integers with $-N \le \xi, \eta \le N$. Rather than interpret R as a radius, we simply observe that the severity of such a blur is determined by the number of zeroes[1] in $|\hat{p}(\rho)|$ on $0 < \rho \le N$.

6.1. True Wiener filtering and the Tikhonov–Miller method. Wiener filtering [32, p. 356], is an important example of a method that does not impose differentiability constraints. It assumes instead that the power spectra $|\hat{n}(\xi,\eta)|$ and $|\hat{f}_e(\xi,\eta)|$ of each of $n(x,y)$ and $f_e(x,y)$ are known. When this is the case, Wiener filtering produces a solution $f^w(x,y)$ in (24) defined as follows in Fourier space

$$\hat{f}^w(\xi,\eta) = \frac{\overline{\hat{p}}(\xi,\eta)\hat{g}(\xi,\eta)}{|\hat{p}(\xi,\eta)|^2 + |\hat{n}(\xi,\eta)|^2/|\hat{f}_e(\xi,\eta)|^2}, \tag{28}$$

where $\overline{z}$ denotes the complex conjugate of z. Under some additional conditions, it can be shown that $f^w(x,y)$ is an approximate solution of $Pf = g$ that *minimizes* the error $\| f - f_e \|_2$ over all $f \in L^2$. In practice, the power spectra $|\hat{n}(\xi,\eta)|$ and $|\hat{f}_e(\xi,\eta)|$ are very seldom known in advance, and true Wiener filtering is almost never realizable. However, the solution (28) is of considerable theoretical interest because of its optimality property. Note that numerous ad hoc versions of (28) exist, in which more readily available quantities are substituted in place of the required, but unavailable, true power spectra. Such versions are sometimes called Wiener filtering by an abuse of terminology. However, these substitute versions *do not* satisfy the Wiener optimality criterion, nor do they elicit the same degree of theoretical interest.

One of the best-known rigorously analyzable and feasible versions of Wiener filtering is the Tikhonov–Miller method [28], now considered canonical in image deblurring [25]. Significantly, this method makes no a priori assumptions regarding the statistical character of the data noise. For nondifferentiable images, Tikhonov–Miller restoration requires the following a priori information: an upper bound $\epsilon > 0$ for the L^2 norm of the noise $n(x,y)$ in the blurred image $g(x,y)$, and an upper bound M for the L^2 norm of the unblurred image f_e

$$\| n \|_2 = \| Pf_e - g \|_2 \le \epsilon, \qquad \| f_e \|_2 \le M, \qquad \epsilon/M \ll 1. \tag{29}$$

It is assumed that ϵ and M are compatible with the existence of an $f_e(x,y) \in L^2$ satisfying (29). Tikhonov–Miller restoration is defined as the unique function $f^\tau(x,y)$ such that

$$f^\tau(x,y) = Arg \min_{f \in L^2(R^2)} \left\{ \| Pf - g \|_2^2 + (\epsilon/M)^2 \| f \|_2^2 \right\}. \tag{30}$$

As will be seen from Theorem 3, where the Tikhonov–Miller method corresponds to the special case $\Gamma_{\overline{t}} = 0$, this minimum problem has a unique solution satisfying

$$Q_\tau f^\tau = P^* g, \qquad Q_\tau = P^* P + (\epsilon/M)^2 I. \tag{31}$$

Moreover, there holds the following best-possible error bound for Tikhonov–Miller reconstruction

$$\| f^\tau - f_e \|_2 \le \epsilon \sqrt{2} \, \| Q_\tau^{-1/2} \|_2, \tag{32}$$

[1] The first five positive zeroes of $J_1(x)$ are 3.83171, 7.01559, 10.17347, 13.32369, and 16.47063.

where

$$\| Q_\tau^{-1/2} \|_2 = \sup_{\xi,\eta} \left\{ |\hat{p}(\xi,\eta)|^2 + (\epsilon/M)^2 \right\}^{-1/2}. \tag{33}$$

Given the psf $p(x,y)$, together with the a priori information ϵ, M, one can always find the maximum value in the $2N \times 2N$ array on the right of (33). As in (28) we may implement (31) in Fourier space. We have

$$\hat{f}^\tau(\xi,\eta) = \frac{\overline{\hat{p}}(\xi,\eta)\hat{g}(\xi,\eta)}{|\hat{p}(\xi,\eta)|^2 + (\epsilon/M)^2}. \tag{34}$$

Moreover, from (29) and Parseval's relation

$$\int_{R^2} |\hat{n}(\xi,\eta)|^2 d\xi d\eta \le \epsilon^2, \qquad \int_{R^2} |\hat{f}_e(\xi,\eta)|^2 d\xi d\eta \le M^2. \tag{35}$$

Therefore, the Tikhonov–Miller method can be viewed as an approximate version of true Wiener filtering where the unavailable pointwise values of the spectra in (28) are replaced by more readily available integrals of these quantities. However, it may be anticipated that since true Wiener filtering requires prior knowledge in the form of $8N^2$ numbers for a $2N \times 2N$ image, whereas the Tikhonov–Miller method requires only 2, less accurate results must generally be expected from the latter method.

7. The Poisson Singular Integral (PSI) method for images $\in \Lambda(\alpha, 2, \infty)$. The preceding discussion was necessary to set the stage for the PSI method. Here, in addition to the a priori constraints (29), the behavior of $\| U^t f_e - f_e \|_2$ on $0 \le t \le \bar{t}$ is assumed known, as in the case of Table 1. The constants $C_{\bar{t}}$ and α are now used to place a further constraint on $f_e(x,y)$. For any $f \in L^2(R^2)$, we have on Fourier transforming $f - U^t f$ and using (10),

$$\mathcal{F}\left\{ f - U^t f \right\} = \left(1 - e^{-t\rho}\right) \hat{f}(\xi,\eta), \qquad \rho = \sqrt{\xi^2 + \eta^2}. \tag{36}$$

Therefore, from Parseval's theorem,

$$\int_0^t \| U^s f - f \|_2^2 \, ds = \int_0^t ds \int_{R^2} \left(1 - e^{-s\rho}\right)^2 |\hat{f}(\xi,\eta)|^2 d\xi d\eta. \tag{37}$$

For fixed $t > 0$, define $\hat{z}(\xi,\eta,t) \ge 0$ by

$$\hat{z}(\xi,\eta,t) = \left\{ \int_0^t \left(1 - e^{-s\rho}\right)^2 ds \right\}^{1/2} = \left\{ t + \frac{4e^{-t\rho} - e^{-2t\rho} - 3}{2\rho} \right\}^{1/2}. \tag{38}$$

It follows directly from the integral definition in (38) that for any fixed $t > 0$, $\hat{z}(\rho,t)$ is a strictly increasing function of ρ, and that $\hat{z}(0,0,t) = 0$. For fixed $t > 0$, define the linear operator $Z(t)$ in $L^2(R^2)$ by

$$Z(t)f = \int_{R^2} \hat{z}(\xi,\eta,t)\hat{f}(\xi,\eta) e^{2\pi i(\xi x + \eta y)} d\xi d\eta. \tag{39}$$

Then, from (37),

$$\int_0^t \| U^s f - f \|_2^2 \, ds = \| Z(t)f \|_2^2. \tag{40}$$

For any $f_e \in \Lambda(\alpha, 2, \infty)$, $0 < \alpha < 1$, we have $\| U^s f_e - f_e \|_2 \leq C_{\bar{t}} \| f_e \|_2 s^\alpha$, $0 \leq s \leq \bar{t}$, where $C_{\bar{t}}$ is a positive constant depending on $\bar{t}$, f_e and α. Therefore, with $\| f_e \|_2 \leq M$,

$$\| Z(\bar{t}) f_e \|_2^2 \leq \frac{C_{\bar{t}}^2 M^2 \bar{t}^{1+2\alpha}}{1+2\alpha}. \tag{41}$$

Define

$$\Gamma_{\bar{t}} = \left\{ \frac{1+2\alpha}{C_{\bar{t}}^2 \, \bar{t}^{1+2\alpha}} \right\}^{1/2}. \tag{42}$$

The exact image $f_e(x, y)$ satisfies the following a priori constraints:

$$\| Pf_e - g \|_2 \leq \epsilon, \qquad (\epsilon/M) \| f_e \|_2 \leq \epsilon, \qquad (\epsilon/M) \, \Gamma_{\bar{t}} \| Z(\bar{t}) f_e \|_2 \leq \epsilon. \tag{43}$$

Fix $\bar{t} > 0$, and consider the minimization problem

$$f^\psi(x, y) = Arg \min_{f \in L^2(R^2)} \left\{ \| Pf - g \|_2^2 + (\epsilon/M)^2 \left(\| f \|_2^2 + \Gamma_{\bar{t}}^2 \| Z(\bar{t}) f \|_2^2 \right) \right\}. \tag{44}$$

As will be seen in Theorem 3 below, this minimum problem has a unique solution satisfying

$$Q_\psi f^\psi = P^* g, \qquad Q_\psi = P^* P + (\epsilon/M)^2 \left\{ I + \Gamma_{\bar{t}}^2 Z(\bar{t})^* Z(\bar{t}) \right\}. \tag{45}$$

The function $f^\psi(x, y)$ in (44) is defined to be the PSI deblurred image. Moreover, there holds the following error bound for PSI deblurring

$$\| f^\psi - f_e \|_2 \leq \epsilon \sqrt{3} \, \| Q_\psi^{-1/2} \|_2, \tag{46}$$

where

$$\| Q_\psi^{-1/2} \|_2 = \sup_{\xi, \eta} \left\{ |\hat{p}(\xi, \eta)|^2 + (\epsilon/M)^2 \left(1 + \Gamma_{\bar{t}}^2 |\hat{z}(\xi, \eta, \bar{t})|^2 \right) \right\}^{-1/2}. \tag{47}$$

Given the psf $p(x, y)$, together with the a priori information ϵ, M, and $\Gamma_{\bar{t}}$, one can always find the maximum value in the $2N \times 2N$ array on the right of (47). Again, as in (28) and (34), f^ψ can be found explicitly in Fourier space. We have

$$\hat{f}^\psi(\xi, \eta) = \frac{\overline{\hat{p}}(\xi, \eta) \hat{g}(\xi, \eta)}{|\hat{p}(\xi, \eta)|^2 + (\epsilon/M)^2 \{ 1 + \Gamma_{\bar{t}}^2 |\hat{z}(\xi, \eta, \bar{t})|^2 \}}. \tag{48}$$

Equations (45)–(48) should be compared with equations (31)–(34). Tikhonov–Miller deblurring can then be seen as an extreme case of PSI deblurring, the case where $f_e(x, y)$ is presumed no smoother than the most general L^2 function, so that $\| U^t f_e - f_e \|_2 = o(1)$ as $t \downarrow 0$. This corresponds to $C_{\bar{t}} = \infty$ in (41), and hence, $\Gamma_{\bar{t}} = 0$ in (48).

THEOREM 3. *Fix $\bar{t} > 0$ and let the exact image $f_e(x, y)$ satisfy the a priori constraints* (43). *Let $f^\psi(x, y)$ minimize* (44), *and let Q_ψ be the positive self-adjoint operator on $L^2(R^2)$ given by*

$$Q_\psi = P^* P + (\epsilon/M)^2 \left\{ I + \Gamma_{\bar{t}}^2 Z(\bar{t})^* Z(\bar{t}) \right\}. \tag{49}$$

Then f^ψ is the unique solution of $Q_\psi f^\psi = P^* g$, and f^ψ satisfies

$$\begin{aligned} &\| Pf^\psi - g \|_2^2 + (\epsilon/M)^2 \left\{ \| f^\psi \|_2^2 + \Gamma_{\bar{t}}^2 \| Z(\bar{t}) f^\psi \|_2^2 \right\} \le 3\epsilon^2, \\ &\| P(f^\psi - f_e) \|_2^2 + (\epsilon/M)^2 \left\{ \| f^\psi - f_e \|_2^2 + \Gamma_{\bar{t}}^2 \| Z(\bar{t})(f^\psi - f_e) \|_2^2 \right\} \le 3\epsilon^2. \end{aligned} \tag{50}$$

This implies the L^2 error bound

$$\| f^\psi - f_e \|_2 \le \epsilon \sqrt{3} \| Q_\psi^{-1/2} \|_2, \tag{51}$$

where

$$\| Q_\psi^{-1/2} \|_2 = \sup_{\xi,\eta} \left\{ |\hat{p}(\xi,\eta)|^2 + (\epsilon/M)^2 \left(1 + \Gamma_{\bar{t}}^2 |\hat{z}(\xi,\eta,\bar{t})|^2 \right) \right\}^{-1/2}. \tag{52}$$

Proof. Let $\mathcal{H}$ denote the Hilbert space direct sum $L^2(R^2) \bigoplus L^2(R^2) \bigoplus L^2(R^2)$ with elements $[u,v,w]$, scalar product $([u_1,v_1,w_1],[u_2,v_2,w_2]) \equiv \langle u_1,u_2\rangle + \langle v_1,v_2\rangle + \langle w_1,w_2\rangle$, and norm $|||\ |||$. Let $\tilde{P} : L^2(R^2) \mapsto \mathcal{H}$ be defined by $\tilde{P}f = [Pf,\ \omega f,\ \omega\Gamma_{\bar{t}} Z(\bar{t}) f]$, where $\omega = (\epsilon/M)$, and let $\tilde{g} = [g,0,0]$. We seek to minimize $|||\tilde{P}f - \tilde{g}|||$ over all $f \in L^2(R^2)$. The normal equation $\tilde{P}^*\tilde{P}f^\psi = \tilde{P}^*\tilde{g}$ gives $Q_\psi f^\psi = P^* g$ with Q as in (49). By hypothesis $|||\tilde{P}f_e - \tilde{g}|||^2 \le 3\epsilon^2$. The minimizing element f^ψ is such that $\tilde{P}f^\psi$ is the orthogonal projection in $\mathcal{H}$ of $\tilde{g}$ on the range of $\tilde{P}$. By the Pythagorean theorem

$$|||\tilde{P}f^\psi - \tilde{g}|||^2 + |||\tilde{P}(f_e - f^\psi)|||^2 = |||\tilde{P}f_e - \tilde{g}|||^2 \le 3\epsilon^2. \tag{53}$$

This proves (50). We now establish (51). From (49), (50),

$$\| Q_\psi^{1/2}(f_e - f^\psi) \|_2^2 = \langle Q_\psi(f_e - f^\psi), (f_e - f^\psi)\rangle = |||\tilde{P}(f_e - f^\psi)|||^2 \le 3\epsilon^2. \tag{54}$$

Hence,

$$\| f_e - f^\psi \|_2 = \| Q_\psi^{-1/2} Q_\psi^{1/2}(f_e - f^\psi) \|_2 \le \epsilon\sqrt{3} \| Q_\psi^{-1/2} \|_2 . \tag{55}$$ ∎

8. A preliminary deblurring experiment. In the following controlled experiment, knowledge of the exact solution $f_e(x,y)$ is used to derive *exact* values for all parameters that constitute a priori information in each of the above three methods. Such exact knowledge is not available in practice. The experiment is primarily of theoretical interest. It is designed to illustrate major differences in behavior, and to properly locate the PSI method in relation to Wiener filtering and the Tikhonov–Miller method. The PSI method with *inexact* information is discussed in section 10.

The 8-bit 512×512 Marilyn Monroe image $f_e(x,y)$ in Figure 3 was synthetically defocused by Fourier domain multiplication with the expression in (27) using $R = 0.06$. This produced the exact blurred image $g_e(x,y)$. Multiplicative noise $n(x,y)$ was then added to g_e as follows. Each pixel value $g_e(x,y)$ was perturbed by adding to it the quantity $n(x,y) = 0.03\sigma(x,y)g_e(x,y)$, where $\sigma(x,y)$ is an array of uniformly distributed random numbers in the range $[-1,1]$. We term this process "adding 3% noise". With varying percentages, we shall use the same process in all our experiments. Note that no noise is thereby added at points where $g_e(x,y) = 0$. The resulting $g(x,y) = g_e(x,y) + n(x,y)$ is shown in Figure 4(A). We find $\| n \|_2 = \epsilon = 2.247$, and $\| f_e \|_1 = 107.59$, $\| f_e \|_2 = M = 131.13$. Therefore $\epsilon/M = 0.01713$. From Table

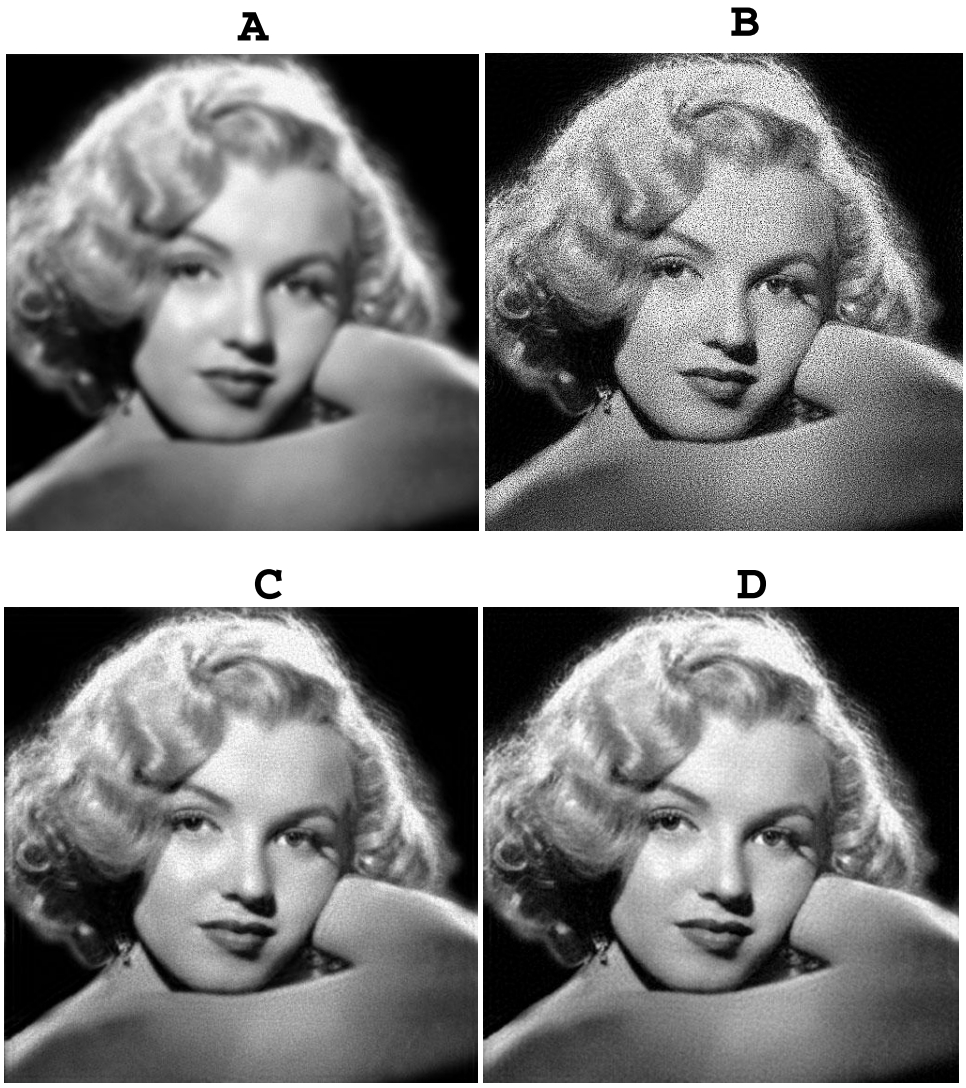


FIG. 4. *Instructive deblurring experiment with exact a priori information highlights significant differences in behavior in above three FFT-based methods.* (A) *Defocused Marilyn Monroe image with* $R = 0.06$ *and* 3% *multiplicative noise.* (B) *Tikhonov–Miller method with exact parameters* ϵ *and* M*, brings out significant noise.* (C) *PSI method with exact parameters* $\epsilon,\ M,\ \alpha = 0.474,\ C_{\bar{t}} = 0.68$. (D) *True Wiener filtering with exact power spectra* $|\hat{n}(\xi,\eta)|,\ |\hat{f}_e(\xi,\eta)|$. *Realizable PSI deblurring closely matches unrealizable true Wiener filtering.*

TABLE 2
Behavior in defocused Marilyn Monroe image in Figure 4.

Deblurring method	L^1 *relative error*	L^2 *relative error*
Tikhonov–Miller (B)	29.82%	34.17%
Poisson Singular Integral (C)	6.89%	9.04%
True Wiener filtering (D)	6.03%	7.88%

1, we have $\| U^t f_e - f_e \|_2 \leq 0.68 \| f \|_2 t^{0.474}$, $0 < t \leq 0.1$. With $\bar{t} = 0.1$, (42) gives $\Gamma_{\bar{t}} = 19.33$. Next, using FFT algorithms, we obtain the exact power spectra $|\hat{n}(\xi,\eta)|$, $|\hat{f}_e(\xi,\eta)|$. We are now ready to compare these three FFT-based procedures under optimal conditions for each method.

The Tikhonov–Miller reconstruction is shown in Figure 4(B). Significantly, this reconstruction is quite noisy, despite the use of *exact* values for ϵ and M. While the regularizing information in (29) prevents *explosive* noise amplification, it is obviously insufficient to prevent serious noise contamination. This is generally the case in the Tikhonov–Miller method. The PSI restoration is shown in Figure 4(C). Here, the additional information that $f_e \in \Lambda(\alpha, 2, \infty)$, together with the values of the constants $C_{\bar{t}}$ and α, were evidently decisive in eliminating noise. The Wiener filtered solution, shown in Figure 4(D), appears only slightly better than the PSI solution. However, the very major difference between true Wiener filtering and the approximate version known as the Tikhonov–Miller method, is another significant result brought out by this deblurring experiment.

It is instructive to study the L^1 and L^2 relative error pattern shown in Table 2. It is widely assumed in practice that the L^2 minimum error property of true Wiener filtering remains valid for the more feasible, approximate versions of such filtering. This is emphatically not the case. The Tikhonov–Miller relative errors are more than *four times larger* than the true Wiener errors. On the other hand, relative errors in the PSI method are only slightly larger than those for true Wiener filtering. Put another way, the PSI method appears to be a feasible procedure that can very substantially improve upon the Tikhonov–Miller method.

Insight into how this improvement comes about can be gained by an analysis of the respective error bounds for each method. Notice that each of the denominators on the right-hand sides of (34) and (48) are radially symmetric functions of (ξ,η), while this is not the case in (28). These denominators play a dual role. They define the actual regularization procedures in (34) and (48), and they define the resulting error bounds in (33) and (47). Because of the radial symmetry, a one-dimensional picture tells the whole story. Define the respective Tikhonov–Miller and PSI *error bound functions* $\theta_\tau(\xi)$, $\theta_\psi(\xi)$ as follows

$$
\begin{aligned}
\theta_\tau(\xi) &= \left\{ |\hat{p}(\xi,0)|^2 + (\epsilon/M)^2 \right\}^{-1/2}, \\
\theta_\psi(\xi) &= \left\{ |\hat{p}(\xi,0)|^2 + (\epsilon/M)^2 (1 + \Gamma_{\bar{t}}^2 |\hat{z}(\xi,0,\bar{t})|^2) \right\}^{-1/2}.
\end{aligned} \tag{56}
$$

In Figure 5, we plot $\theta_\tau(\xi)$ and $\theta_\psi(\xi)$ as determined by the actual parameter values that entered the deblurring experiment in Figure 4. The significant differences in these two curves translate into fundamental differences in the Fourier space regularization that defines the corresponding procedures. From (27), we see that $\theta_\tau(\xi)$ has a maximum of $M/\epsilon = 58.36$, at every point $\xi > 0$ where $J_1(0.06\ \xi) = 0$. There are 4 such points on $0 < \xi \leq 256$. The curve $\theta_\psi(\xi)$ also develops maxima at these same points, but these maxima are about *five times smaller* than those in $\theta_\tau(\xi)$, owing to the additional term involving $\hat{z}(\xi,0,\bar{t})$. Since the error estimate in each method is proportional to the maximum along the corresponding curve, it is natural to find substantially smaller errors in Figure 4(C) than in Figure 4(B).

9. Comparing total variation deblurring with PSI deblurring. The use of initial value PDE methods in image processing and computer vision has mushroomed into an important new branch of applied mathematics. The basic idea originates in

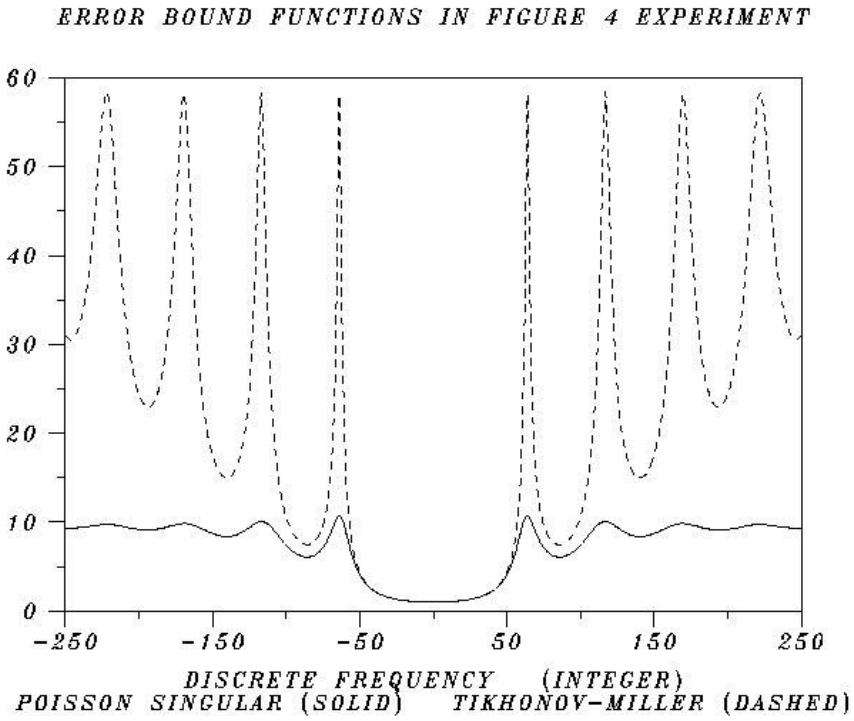


FIG. 5. *Plot of error bound functions $\theta_\tau(\xi)$ (dashed curve), and $\theta_\psi(\xi)$ (solid curve), as defined in* (56)*, for the deblurring experiment in Figure* 4*. Maximum value in θ_τ is more than five times larger than in θ_ψ. Qualitative difference in behavior in these two curves implies significant difference in Fourier domain regularization in the PSI and Tikhonov–Miller methods. Difference in maximum values explains large difference in L^2 relative errors in Figures* 4(B) *and* 4(C).

gradient descent methods for minimizing appropriate energy functionals. Instructive surveys of this general set of ideas may be found in [11] and [39].

The total variation approach introduced in [33] is one of the most popular PDE methods, and it is primarily designed to recover edges in the original image. Given the deconvolution problem $Pf = g$ as in (24), TV deblurring presupposes the exact sharp image $f_e(x, y) \in BV(R^2)$, and it produces an image $f^{tv}(x, y)$ defined by

$$f^{tv}(x,y) = Arg \min_{f \in BV(R^2)} \left\{ (\lambda/2) \parallel Pf - g \parallel_2^2 + \int_{R^2} |\nabla f| dxdy \right\}. \tag{57}$$

This means that $f^{tv}(x, y)$ is the solution of

$$P^* P f^{tv} - \lambda^{-1} \nabla . \left(\frac{\nabla f^{tv}}{|\nabla f^{tv}|} \right) = P^* g. \tag{58}$$

Here, $\lambda > 0$ is a regularization parameter that can be tuned. Provided the noise level in $g(x, y)$ is small, larger values of λ produce sharper images. Too large a value of λ leads to computational instability. Unlike the cases in (31) and (45), (58) is a nonlinear deconvolution problem that cannot be solved explicitly in Fourier space. In fact, considerable effort is generally required to obtain f^{tv} for large size imagery. In pure denoising applications, where $P = I$, this effort is usually warranted by the

quality of the resulting restoration. Recently, a new time dependent evolutionary approach to (58) has been developed [26], whereby $f^{tv}(x,y)$ is obtained as the steady state solution to the following nonlinear anisotropic diffusion problem

$$\text{(59)} \qquad \begin{cases} u_t = -\lambda|\nabla u|\ P^*(Pu-g) + |\nabla u|\ \nabla.\left(\nabla u/\{\sqrt{|\nabla u|^2+\beta}\}\right), \\ u(x,y,0) = g(x,y), \end{cases}$$

where the given blurred image $g(x,y)$ is used as the initial value. In addition, $u(x,y,t)$ satisfies homogeneous Neumann conditions at the boundary of the unit square Ω. In (59), $\beta > 0$ is a small constant designed to prevent division by zero. In [26, section 5], an efficient new explicit finite difference scheme for (59) is proposed. This scheme has improved stability and edge-enhancing properties, and converges rapidly to the desired steady state solution. Accordingly, we shall use that method in our total variation deblurring experiments.

This paper has drawn attention to the fact that most images are not smooth. The PSI method is predicated on locating $f_e(x,y)$ in the correct Lipschitz space, while TV deblurring assumes $f_e(x,y) \in BV(R^2)$. It may be argued that such refined smoothness measures are primarily applicable to $f^{\infty}(x,y)$, the original intensity field that gave rise to the digitized finite dimensional object $f_e(x,y)$, but may not be meaningful for $f_e(x,y)$ itself. Indeed, since all norms are equivalent in finite dimensional space, it remains to be seen whether such abstruse function space notions are ultimately of any computational significance.

Our first experiment involves a slightly defocused image with very little noise. The original sharp USS Eisenhower image is shown in Figure 6(A). Fourier space multiplication with (27) using $R = 0.03$, followed by the addition of 0.1% multiplicative noise, produced the blurred Figure 6(B). Because of the low noise level, we chose β small and λ large in (59), as recommended in [26]. With $\beta = 0.0001,\ \Delta t = 0.1(\Delta x)^2$ and $\lambda = 300$, we obtained Figure 6(C) at $T = 100\Delta t$. Higher values of λ were computationally unstable. Moreover, the resulting TV image did not improve if more time steps were taken. Figure 6(D) is the PSI deblurred image using exact values for $\epsilon,\ M$, and using $\alpha = 0.362,\ C_{\bar{t}} = 0.50$, from Table 1. Zooming on selected parts of the image in Figures 6(E) and 6(F), clearly shows significant loss of structural detail in the TV image, as compared with PSI deblurring. For completeness, the L^1 relative errors in this experiment were as follows: true Wiener filtering (not shown) 1.67%, PSI method 2.18%, TV deblurring 6.83%, and Tikhonov–Miller (not shown) 4.71%.

In our second experiment, the sharp English village image in Figure 7(A) was moderately defocused using $R = 0.06$, and 0.1% multiplicative noise was again added to form Figure 7(B). With β and Δt as in Figure 6(C), it was possible to choose $\lambda = 500$, and obtain Figure 7(C) at $T = 100\Delta t$. Again, no improvement was noted with more time steps. Figure 7(D) is the PSI deblurred image using the exact values for $\epsilon,\ M$, together with $\alpha = 0.439,\ C_{\bar{t}} = 0.55$, from Table 1. Because of the stronger blur, more information is now lost in TV deblurring. Zooming in on the first three houses in Figures 7(E) and 7(F), we see that the windows and roof shingles have been virtually eliminated in the TV image. The L^1 relative errors in this experiment were as follows: true Wiener filtering (not shown) 1.98%, PSI method 3.05%, TV deblurring 6.70%, and Tikhonov–Miller (not shown), 7.42%.

10. The PSI method with inexact Lipschitz data. The controlled experiments in sections 8 and 9 were designed to illustrate important theoretical points, and involved use of the PSI method with exact prior Lipschitz space data. In fact,

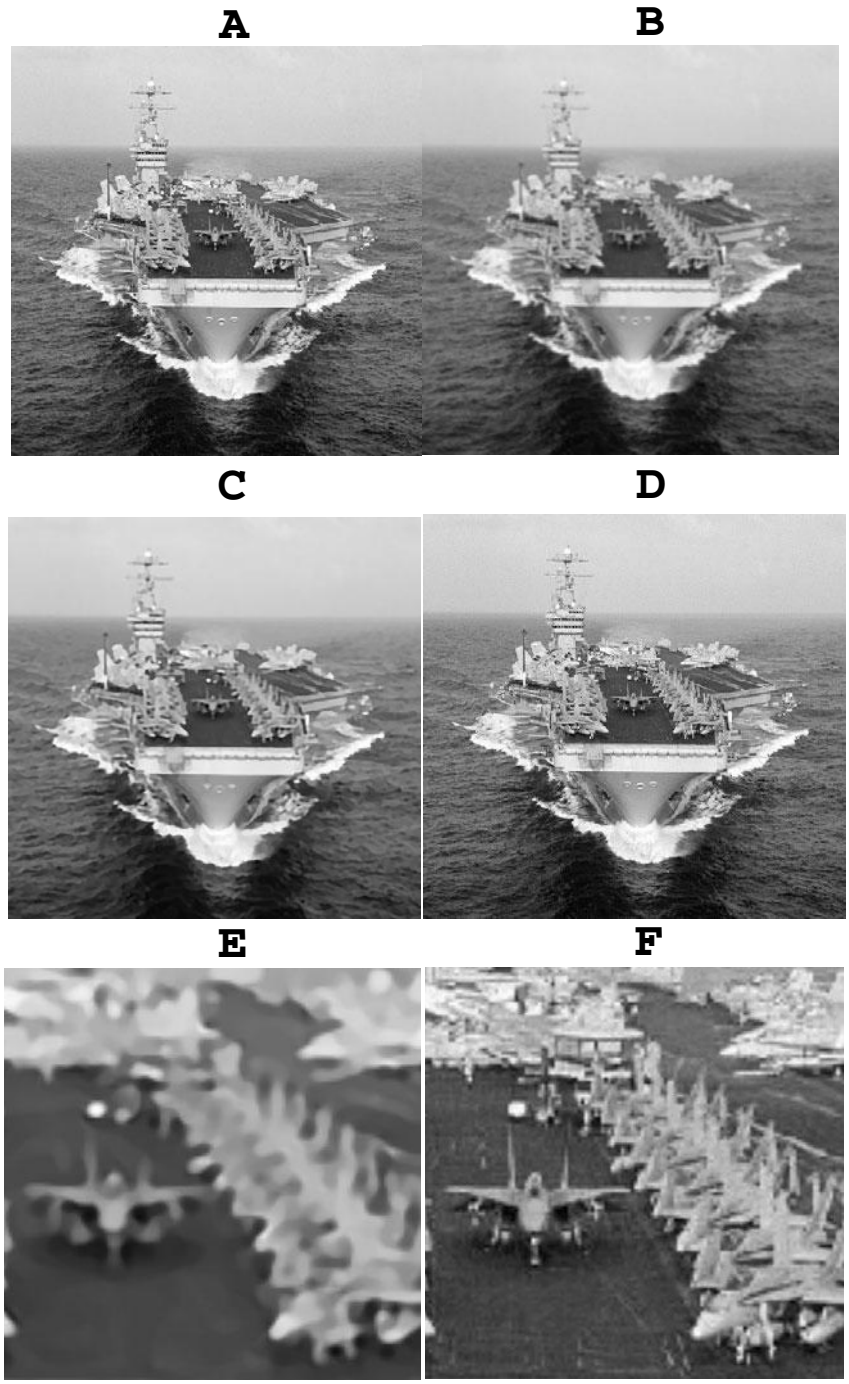


FIG. 6. *Comparison of total variation and PSI deblurring on mildly blurred image. Zooming on selected parts of the image enables meaningful comparisons between the two methods.* (A) *Original sharp USS Eisenhower image.* (B) *Mildly defocused image with* $R = 0.03$ *and* 0.1% *multiplicative noise.* (C) *Total variation deblurring by applying finite difference scheme in* [26, section 5] *to* (59), *with* $\beta = 0.0001,\ \lambda = 300,\ \Delta t = 0.1(\Delta x)^2,\ T = 100\Delta t$. (D) *PSI deblurring using exact a priori parameters,* $\epsilon,\ M,\ \alpha = 0.362,\ C_{\overline{t}} = 0.50$. (E) *Zooming in TV deblurred image reveals significant loss of structural detail.* (F) *Zooming on same region in PSI image.*

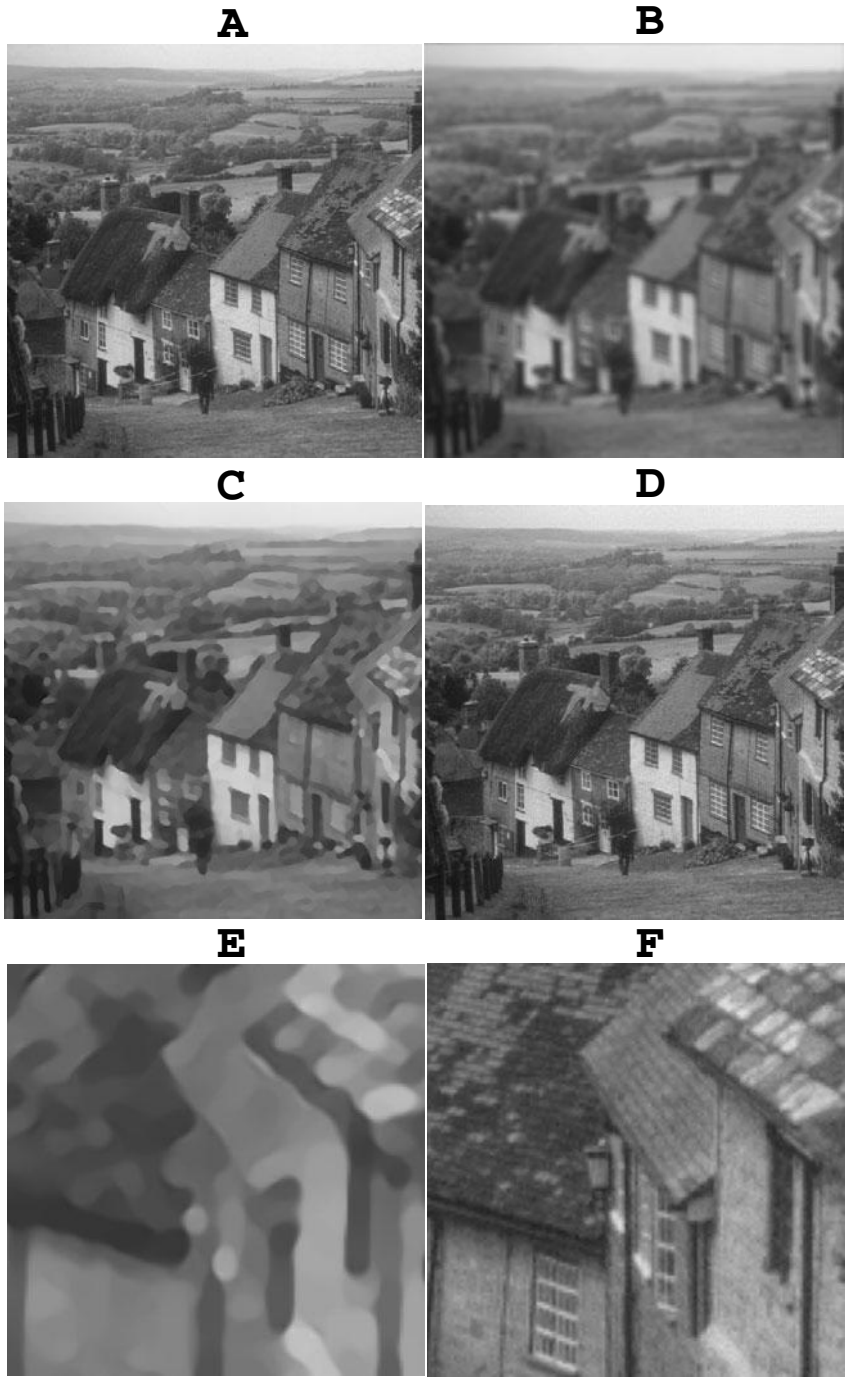


FIG. 7. *Comparison of total variation and PSI methods on moderately blurred image. Zooming now reveals unacceptable loss of content in TV deblurring.* (A) *Original sharp English village image.* (B) *Moderately defocused image with* $R = 0.06$ *and* 0.1% *multiplicative noise.* (C) *Total variation deblurring using scheme in* [26, section 5] *with* $\beta = 0.0001,\ \lambda = 500,\ \Delta t = 0.1(\Delta x)^2,\ T = 100\Delta t$. (D) *PSI method with exact a priori parameters,* $\epsilon,\ M,\ \alpha = 0.439,\ C_{\overline{t}} = 0.55$. (E) *Zooming in image* (C) *reveals loss of windows and roof shingles.* (F) *Zooming on same region in PSI image.*

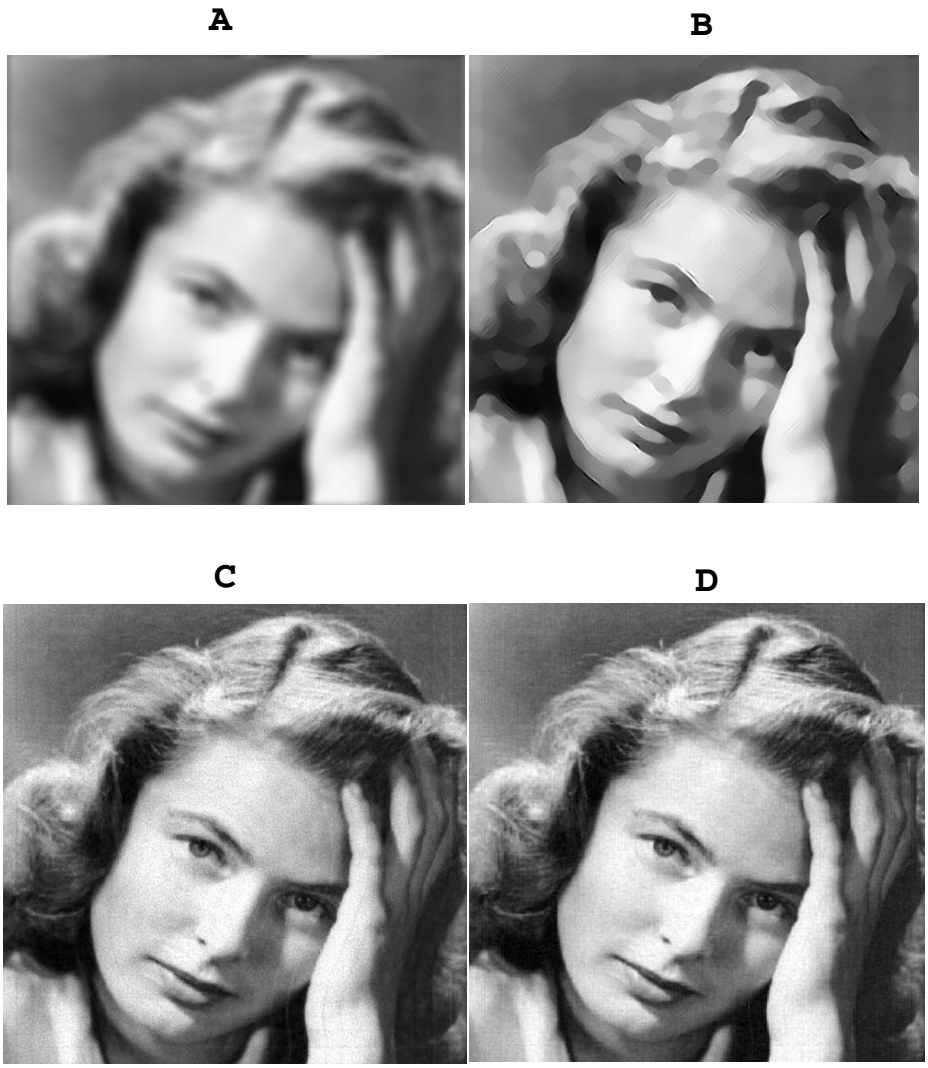


FIG. 8. *Use of plausible guess. Robust PSI method can produce useful reconstructions, even with inexact Lipschitz data.* (A) *Moderately defocused Ingrid Bergman image with* $R = 0.12$ *and* 0.5% *multiplicative noise.* (B) *Total variation deblurring using scheme in* [26, section 5]*, with* $\lambda = 400,\ \beta = 0.0001,\ \Delta t = 0.1(\Delta x)^2$*, and* $T = 150\Delta t$*, produces lifeless, mannequin-like appearance.* (C) *PSI method using plausible guess* $\alpha = 0.5,\ C_{\bar{t}} = 0.5$. (D) *True Wiener filtering with exact power spectra* $|\hat{n}(\xi,\eta)|,\ |\hat{f}_e(\xi,\eta)|$*. Fast PSI deblurring, with fictitious Lipschitz data, produces good first approximation to unrealizable optimal Wiener image.*

TABLE 3
Behavior in moderately defocused Ingrid Bergman image in Figure 8.

Deblurring method	L^1 *relative error*	L^2 *relative error*
Tikhonov–Miller (not shown)	24.15%	27.33%
Total Variation (B)	5.79%	8.45%
PSI with plausible guess (C)	4.75%	5.76%
True Wiener filtering (D)	3.53%	4.40%

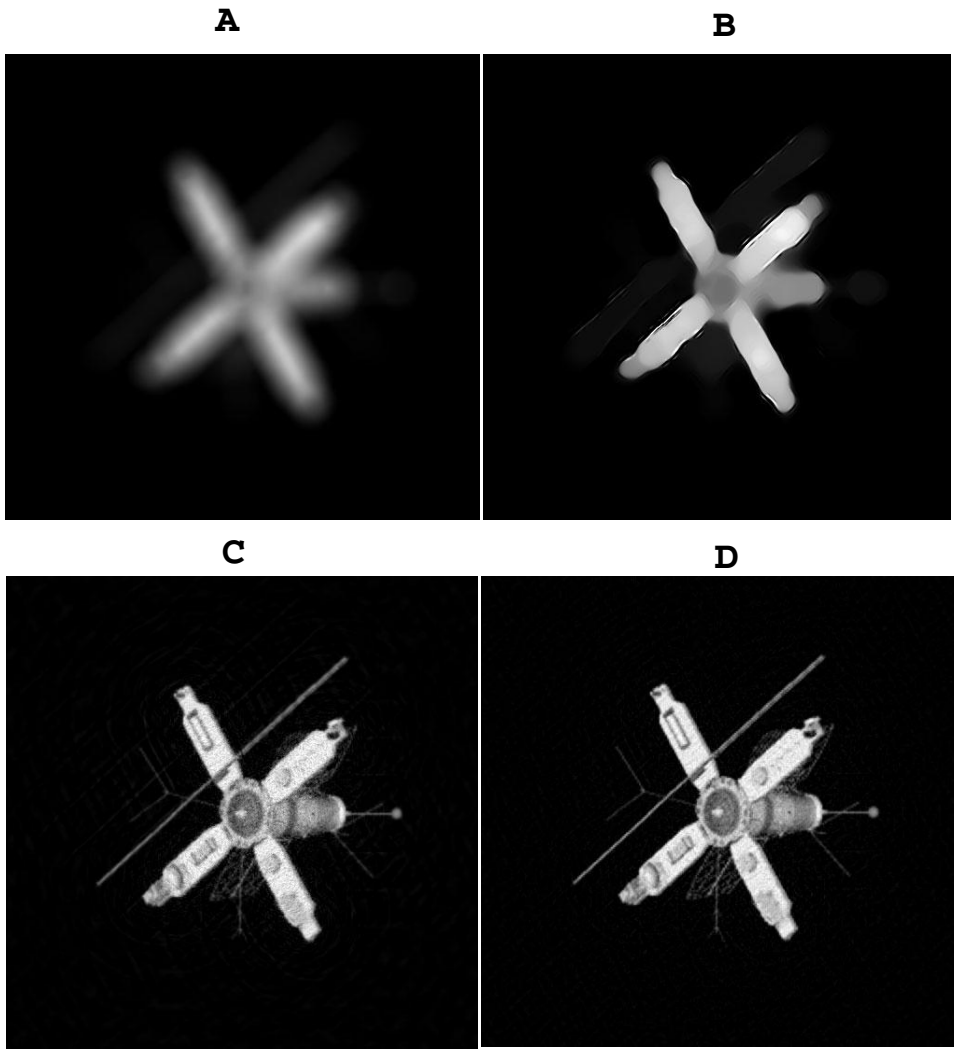


FIG. 9. *Use of substitute information. Robust PSI method produces remarkably good reconstruction using Lipschitz space data corresponding to image of "similar" object.* (A) *Strongly defocused USAF satellite image with* $R = 0.25$ *and* 0.5% *multiplicative noise.* (B) *Total variation deblurring using scheme in* [26, section 5]*, with* $\lambda = 400,\ \beta = 0.0001,\ \Delta t = 0.1(\Delta x)^2$*, and* $T = 150\Delta t$*, results in severe loss of texture.* (C) *PSI method using substitute Lipschitz data* $\alpha = 0.417,\ C_{\bar{t}} = 0.99$*, obtained from Mariner* 5 *image in Figure* 3. (D) *True Wiener filtering with exact power spectra* $|\hat{n}(\xi,\eta)|,\ |\hat{f}_e(\xi,\eta)|$*. Fast PSI deblurring, using substitute data, closely matches unrealizable optimal Wiener image.*

TABLE 4
Behavior in strongly defocused USAF satellite image in Figure 9.

Deblurring method	L^1 *relative error*	L^2 *relative error*
Tikhonov–Miller (Not shown)	37.95%	33.56%
Total Variation (B)	32.91%	31.82%
PSI based on similar object (C)	20.69%	16.83%
True Wiener filtering (D)	17.50%	13.04%

the PSI method is a robust method of great practical significance that can produce useful reconstructions even with *inexact* Lipschitz data. Inspection of Table 1 shows that the values of (C, α) are typically confined to a narrow range. Indeed, a plausible guess for (C, α) might be $(0.5, 0.5)$ in many cases. In other situations, a sharp image of a similar object might provide useful values for (C, α). Such initial reconstructions can then be interactively refined through *fast* simultaneous computation and display of multiple trial PSI images, corresponding to neighboring (C, α) values. We now give two examples that show how good such *initial* reconstructions can be.

Figure 8(A) is an 8-bit 512×512 synthetically defocused Ingrid Bergman image, obtained using (27) with $R = 0.12$, followed by adding 0.5% multiplicative noise. Total variation deblurring using the scheme in [26, section 5], with $\lambda = 400,\ \beta = 0.0001$, and $T = 150\Delta t$, produces the lifeless, mannequin-like appearance shown in Figure 8(B). However, the PSI method with the plausible guess $\alpha = 0.5,\ C_{\frac{1}{t}} = 0.5$, produces Figure 8(C). This initial reconstruction is already in good qualitative agreement with the true Wiener image in Figure 8(D). The L^1 and L^2 relative errors in Table 3 indicate that the PSI method, even with such inexact data, significantly improves upon the Tikhonov–Miller and total variation methods.

Our final experiment involves the strongly defocused USAF satellite image in Figure 9(A), where $R = 0.25$ and 0.5% multiplicative noise was added. As may be expected in such a severely blurred image, total variation deblurring, shown in Figure 9(B), results in severe loss of structural detail. Here, the Mariner 5 image shown in Figure 3 may be considered a "similar" object, and the corresponding $\Lambda(\alpha, 2, \infty)$ information in Table 1, $C = 0.99,\ \alpha = 0.417$, may be used in the PSI method. Remarkably, this produces the reconstruction shown in Figure 9(C). While faint honeycomb artifacts are visible against the dark background in Figure 9(C), this initial PSI image is an excellent approximation to the optimal true Wiener image shown in Figure 9(D). Relative errors in this experiment are shown in Table 4.

In Figure 4, the PSI method's improvement over the Tikhonov–Miller method can be traced to the fact that the constraints in (29) allowed the solution to be *too rough*. In Figures 6 through 9, PSI's improvement over the total variation method stems from the fact that the minimum principle (57) forces the solution to be *too smooth*. Apparently, the use of singular integrals to calibrate image smoothness, together with the direct use of that information in constraining the solutions of the deblurring problem, constitutes an important new idea in image deconvolution.

REFERENCES

[1] S. Bochner, *Harmonic Analysis and the Theory of Probability*, University of California Press, Berkeley, CA, 1955.

[2] E. O. Brigham, *The Fast Fourier Transform*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[3] P. L. Butzer and H. Berens, *Semi-Groups of Operators and Approximation*, Springer-Verlag, New York, 1967.

[4] P. L. Butzer and R. J. Nessel, *Favard classes for n-dimensional singular integrals*, Bull. Amer. Math. Soc., 72 (1966), pp. 493–498.

[5] P. L. Butzer and R. J. Nessel, *Fourier Analysis and Approximation*, Academic Press, New York, 1971.

[6] A. P. Calderón and A. Zygmund, *Singular integrals and periodic functions*, Studia Math., 14 (1954), pp. 249–271.

[7] A. S. Carasso and T. Kato, *On subordinated holomorphic semigroups*, Trans. Amer. Math. Soc., 327 (1991), pp. 867–877.

[8] A. S. Carasso, *Linear and nonlinear image deblurring: A documented study*, SIAM J. Numer. Anal., 36 (1999), pp. 1659–1689.

[9] A. S. Carasso, *The APEX method in image sharpening and the use of low exponent Lévy stable laws*, SIAM J. Appl. Math., 63 (2002), pp. 593–618.

[10] A. S. Carasso, D. S. Bright, and A. E. Vladár, *The APEX method and real-time blind deconvolution of scanning electron microscope imagery*, Optical Engineering, 41 (2002), pp. 2499–2514.
[11] V. Caselles, J. M. Morel, G. Sapiro, and A. Tannenbaum, eds., *Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing*, IEEE Trans. Image Process., 7, 1998.
[12] A. Chambolle, R. A. DeVore, N. Lee, and B. J. Lucier, *Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage*, IEEE Trans. Image Process., 7 (1998), pp. 319–335.
[13] A. Chambolle and P. L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.
[14] T. Chan, A. Marquina, and P. Mulet, *High-order total variation-based image restoration*, SIAM J. Sci. Comput., 22 (2000), pp. 503–516.
[15] T. F. Chan and J. Shen, *Mathematical models for local nontexture inpaintings*, SIAM J. Appl. Math., 62 (2002), pp. 1019–1043.
[16] T. F. Chan and C. K. Wong, *Total variation blind deconvolution*, IEEE Trans. Image Process., 7 (1998), pp. 370–375.
[17] S. Chaudhuri and A. N. Rajagopalan, *Depth from Defocus: A Real Aperture Imaging Approach*, Springer-Verlag, New York, 1999.
[18] R. A. DeVore, B. Jawerth, and B. J. Lucier, *Image compression through wavelet transform coding*, IEEE Trans. Inform. Theory, 38 (1992), pp. 719–746.
[19] R. A. DeVore and B. J. Lucier, *Classifying the smoothness of images: Theory and application to wavelet image processing*, in Proceedings of the 1994 IEEE International Conference on Image Processing, Austin TX, Vol. 2, IEEE Press, Los Alamitos, CA, 1994, pp. 6–10.
[20] D. C. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1198.
[21] S. Durand and J. Froment, *Reconstruction of wavelet coefficients using total variation minimization*, SIAM J. Sci. Comput., 24 (2003), pp. 1754–1767.
[22] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed., Vol. 2, Wiley, New York, 1971.
[23] J. D. Gaskill, *Linear Systems, Fourier Transforms, and Optics*, Wiley, New York, 1978.
[24] Y. Gousseau and J.-M. Morel, *Are natural images of bounded variation?*, SIAM J. Math. Anal., 33 (2001), pp. 634–648.
[25] R. L. Lagendijk and J. Biemond, *Iterative Identification and Restoration of Images*, Kluwer Academic Publishers, Norwell, MA, 1991.
[26] A. Marquina and S. Osher, *Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal*, SIAM J. Sci. Comput., 22 (2000), pp. 387–405.
[27] Y. Meyer, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*, Univ. Lecture Ser. 22, AMS, Providence RI, 2001.
[28] K. Miller, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52–74.
[29] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM J. Appl. Math., 61 (2000), pp. 633–658.
[30] M. Nikolova, *Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers*, SIAM J. Numer. Anal., 40 (2002), pp. 965–994.
[31] E. L. O'Neill, *Introduction to Statistical Optics*, Addison-Wesley, Reading, MA, 1963.
[32] W. K. Pratt, *Digital Image Processing*, 2nd ed., Wiley, New York, 1991.
[33] L. Rudin and S. Osher, *Total variation based image restoration with free local constraints*, in Proceedings of the 1994 IEEE International Conference on Image Processing, Austin, TX, Vol. 1, 1994, IEEE Press, Los Alamitos, CA, pp. 31–35.
[34] A. Sandage, *The Hubble Atlas of Galaxies*, Publication 618, Carnegie Institution of Washington, Washington, DC, 1961.
[35] A. M. Stokolos and W. Trebels, *On the rate of almost everywhere convergence of Abel-Cartwright means on $L^p(R^n)$*, Result. Math., 34 (1998), pp. 373–380.
[36] M. H. Taibleson, *Lipschitz classes of functions and distributions in E_n*, Bull. Amer. Math. Soc., 69 (1963), pp. 487–493.
[37] M. H. Taibleson, *On the theory of Lipschitz spaces of distributions on Euclidean n-space.* I. *Principal properties*, J. Math. Mechanics, 13 (1964), pp. 407–478.
[38] A. E. Vladár, M. T. Postek, and M. P. Davidson, *Image sharpness measurement in scanning electron microscopy*, Scanning, 20 (1998), pp. 24–34.
[39] J. Weickert, *Anisotropic Diffusion in Image Processing*, Teubner, Stuttgart, Germany, 1998.
[40] K. Yosida, *Fractional powers of infinitesimal generators and the analyticity of the semigroups generated by them*, Proc. Japan Acad., 36 (1960), pp. 86–89.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1775–1809

SELF-SIMILAR BLOW-UP IN HIGHER-ORDER SEMILINEAR PARABOLIC EQUATIONS*

C. J. BUDD†, V. A. GALAKTIONOV‡, AND J. F. WILLIAMS§

Abstract. We study the Cauchy problem in $\mathbb{R}\times\mathbb{R}_+$ for one-dimensional $2m$th-order, $m>1$, semilinear parabolic PDEs of the form ($D_x=\partial/\partial x$)

$$u_t=(-1)^{m+1}D_x^{2m}u+|u|^{p-1}u, \text{ where } p>1, \quad \text{and} \quad u_t=(-1)^{m+1}D_x^{2m}u+e^u$$

with bounded initial data $u_0(x)$. Specifically, we are interested in those solutions that blow up at the origin in a finite time T. We show that, in contrast to the solutions of the classical second-order parabolic equations $u_t=u_{xx}+u^p$ and $u_t=u_{xx}+e^u$ from combustion theory, the blow-up in their higher-order counterparts is asymptotically *self-similar*. In particular, there exist exact nontrivial self-similar blow-up solutions, $u_*(x,t)=(T-t)^{-1/(p-1)}f(y)$ in the case of the polynomial nonlinearity and $u(x,t)=-\ln(T-t)+f(y)$ for the exponential nonlinearity, where $y=x/(T-t)^{1/2m}$ is the backward higher-order heat kernel variable. The profiles $f(y)$ satisfy related semilinear ODEs that share the same non–self-adjoint higher-order linear differential operators. We show that there are *at least* $2\lfloor\frac{m}{2}\rfloor$ nontrivial self-similar solutions to the full PDEs. Numerical solution of the ODEs for $m=2$ and 3 supports this, and the time dependent solutions of the PDEs for $m=2$ are then studied by using a scale invariant adaptive numerical method. It is shown that those functions $f(y)$, which have the simplest spatial shape (e.g., a single maximum), correspond to *stable* self-similar solutions. A further countable subset of nonsimilarity blow-up patterns can be constructed by linearization and matching with similarity solutions of a first-order Hamilton–Jacobi equation.

Key words. semilinear parabolic equation, blow-up, similarity solutions, asymptotic behavior

AMS subject classifications. 35K55, 35K65

DOI. 10.1137/S003613990241552X

1. Introduction.

Scaling and self-similarity have been known since the 1930s to give a fundamental insight into many systems that develop singularities in finite time. A general treatment of blow-up processes naturally occurred in the 1930s–1950s in the context of N. N. Semenov's chain reaction theory, adiabatic explosion, and combustion theory (the first blow-up result was by O.M. Todes [43]); see [26, section 15] and [48]. On the other hand, in the same period there was a strong influence from the study of blow-up singularities in gas dynamics; in particular, the intense explosion (focusing) problem, admitting similarity solutions of the second kind, was considered by Bechert, Guderley, and Sedov in the 1940s; see [4, p. 127] and [49]. Another classical area of blow-up processes in the 1960s was nonlinear optics, where the main model is the nonlinear (cubic) Schrödinger equation defined in $\mathbb{R}^2$ or $\mathbb{R}^3$ that admits blow-up self-focusing solutions; see the references in [42].

1.1. On second-order semilinear and quasi-linear heat equations from combustion theory: Singularity formation.

Because of their importance to many applications, canonical equations from combustion theory such as the nonstationary

*Received by the editors October 1, 2002; accepted for publication (in revised form) January 7, 2004; published electronically July 23, 2004. This research was supported by the TMR network ERB FMRX CT98-0201 and RTN network HPRN-CT-2002-00274.

http://www.siam.org/journals/siap/64-5/41552.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (cjb@maths.bath.ac.uk).

‡Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK, and Keldysh Institute of Applied Mathematics, Miusskaya Sq. 4, 125047 Moscow, Russia (vag@maths.bath.ac.uk).

§Centrum voor Wiskunde en Informatica, Postbus 94079, 1090 GB, Amsterdam, The Netherlands (williams@cwi.nl).

semilinear one-dimensional *Frank-Kamenetskii equation* [19] (the solid fuel model [48]),

$$u_t = u_{xx} + e^u, \quad x \in \mathbb{R}, \ t > 0, \tag{1.1}$$

and its counterpart with a power nonlinearity,

$$u_t = u_{xx} + u^p, \quad x \in \mathbb{R}, \ t > 0, \quad \text{with exponent } p > 1 \quad (u(x,t) \ge 0), \tag{1.2}$$

have been well studied for the past thirty years. It is known that these both exhibit singularities in finite time. While exact self-similar solutions are known to exist for the related second-order reaction-diffusion *quasi-linear* problems (see references to Chapter 4 in [40] and [8])

$$u_t = (|u_x|^\sigma u_x)_x + e^u \quad \text{or} \quad u_t = (u^\sigma u_x)_x + u^p \quad \text{with} \ \ \sigma > 0, \tag{1.3}$$

it is somewhat paradoxical that none exist for the above semilinear problems. Instead, the generic stable asymptotic blow-up behavior is described by approximate similarity solutions satisfying first-order Hamilton–Jacobi equations; see the references in the books [5, 40] and the surveys [35, 23]. For example, in the quasi-linear problem (1.3) with the power nonlinearity u^p, for any $p > 1$ and $\sigma > 0$, there exists an *exact* nontrivial self-similar solution of the form

$$u_S(x,t) = (T-t)^{-1/(p-1)} f(y), \quad y = \frac{x}{(T-t)^{(p-1-\sigma)/2(p-1)}}, \tag{1.4}$$

where T is the finite blow-up time and f is not identically constant and solves a related ODE; see [40, Chapter 4]. Other types of blow-up in quasi-linear heat equations via Hamilton–Jacobi asymptotics are described in [24, Chapters 9, 10].

In comparison, for the semilinear equation (1.2), looking for the same similarity solution

$$u_S(x,t) = (T-t)^{-1/(p-1)} f(y), \quad y = \frac{x}{(T-t)^{1/2}}$$

yields that, for the corresponding ODE, the only nonzero similarity profile is the trivial constant one $f \equiv \beta^\beta$, where $\beta = \frac{1}{p-1}$. Such nonexistence results are known from the 1970s; see [31] ($p = 3$), [1] ($p > 1$), and [27] for the corresponding equation in $\mathbb{R}^N$ with $1 < p \le \frac{N+2}{(N-2)_+}$. This means that, for a wide "dense" subset of general solutions $u(x,t)$ blowing up at $t = T$ at the origin $x = 0$, the similarity rescaling satisfies (see [22, 28])

$$\theta(y,t) \equiv (T-t)^{1/(p-1)} u(x,t) \to \beta^\beta \quad \text{as } t \to T^-$$

uniformly on compact subset in y. The spatial variation of the blow-up solutions can be observed on larger subsets, and the generic asymptotic behavior is as follows:

$$u(x,t) = [(p-1)(T-t)(1 + C_* \eta^2)]^{-1/(p-1)} (1 + o(1)) \tag{1.5}$$

uniformly on compact subsets in $\eta = x/[(T-t)|\ln(T-t)|]^{1/2}$, where the constant $C_* = \frac{p-1}{4p}$ does not depend on initial data (nor, in fact, on the space dimension). The non–scaling-invariant "hot-spot variable" η with an extra logarithmic factor was first formally derived in 1972 (see [31]) and was rigorously established twenty years later

(see [7, 18, 30, 37, 45, 46] and the survey [23]). The stable behavior (1.5) is essentially equivalent to the fact that the ODE for the self-similar solutions, which is obtained by a symmetry reduction of the original PDE, has no solution (other than the constant one) with an appropriate decay rate at infinity. Comparing (1.4) and (1.5) shows that nonexistence of nontrivial ODE similarity profiles implies a fundamental change of the basic spatial scale of singularity formation phenomena. The observation that the blow-up behavior of these second-order problems is only approximately self-similar with a new logarithmically perturbed backward heat kernel variable is an essential feature of many related reaction diffusion problems and the corresponding parabolic equations under consideration.

1.2. Main higher-order semilinear models, results, and plan of the paper. Higher-order semilinear parabolic equations arise in many physical applications such as thin film theory, convection-explosion theory, lubrication theory, flame and wave propagation (the Kuramoto–Sivashinsky equation and the extended Fisher–Kolmogorov equation), phase transition at critical Lifschitz points, bistable systems, and applications to structural mechanics. The effect of fourth-order terms on self-focusing problems in nonlinear optics has also recently been considered in [17, 6]. Indeed, fourth- (and higher-) order terms are increasingly recognized as being significant in many physical models, which has led to the burgeoning literature including the recent book [39], which lists a number of models and references. Therefore, it is important to know whether higher-order semilinear equations exhibit singularity behavior analogous to their classical second-order counterparts where the exact self-similar behavior is unavailable.

In the present paper we show that the higher-order generalizations of the second-order model (1.1), the *extended Frank-Kamenetskii equation*,

$$u_t = (-1)^{m+1} D_x^{2m} u + e^u, \quad x \in \mathbb{R},\ t > 0 \quad (D_x = \partial/\partial x), \tag{1.6}$$

and of (1.2),

$$u_t = (-1)^{m+1} D_x^{2m} u + |u|^{p-1} u, \quad x \in \mathbb{R},\ t > 0, \tag{1.7}$$

have self-similar blow-up solutions, and hence their evolution is somewhat simpler than in the case $m = 1$, though, of course, for $m > 1$ the problem of rigorous justification of the results becomes much more delicate. Fundamentally, we would like to understand the importance of the semilinear structure in (1.6) and (1.7) and its role in self-similarity. This study is an attempt to further mathematical understanding of higher-order parabolic equations and, in particular, the corresponding singularity formation phenomena, an area of increasing physical and mathematical importance. In particular, a model, admitting blow-up, from convection-explosion theory has been described in [33] and takes the form

$$u_t = -u_{xxxx} - [(2 - (u_x)^2) u_x]_x - \alpha u + q e^{su}. \tag{1.8}$$

Here the formation of such finite time singularities was shown to be self-similar [25] with a number of analogous properties to the generic equations (1.7) and (1.6).

In section 2 we introduce the relevant mathematical definitions, formulation of similarity variables, and rescaled equations. In section 3 we present the properties of the underlying linearized operator, which governs the "dynamics" of both equations (1.6) and (1.7) near certain blow-up solutions.

In section 4 we consider an extension of the linearized problem that makes clear the structure of the subset of nonlinear evolution patterns. In particular, we analyze bifurcation points associated with the linearized operator and present an argument for the existence of self-similar solutions. This analytic argument is strengthened with numerical and asymptotic evidence. Section 5 is devoted to the asymptotic behavior of the solutions close to bifurcation points.

Lastly, in sections 6 and 7 we construct the blow-up profiles asymptotically and compare them with numerical solutions of both the ODE for the self-similar profile and rescaled profiles from simulations of the full PDEs. A number of our results and techniques can be applied to the blow-up of radially symmetric solutions of the N-dimensional semilinear equations

$$u_t = -(-\Delta)^m u + |u|^{p-1}u \quad \text{and} \quad u_t = -(-\Delta)^m u + e^u.$$

Spectral properties of the corresponding linearized operators in $L^2_\rho(\mathbf{R}^N)$ can be found in [15, 20].

This paper is mainly devoted to the study of self-similar blow-up for higher-order semilinear parabolic equations, though we discuss some related center manifold structures. Countable spectra of other blow-up patterns that are approximately self-similar and are constructed by matching of different asymptotic regions are studied in [20]; see also [25] for (1.8).

2. Finite time blow-up solutions and similarity variables.

2.1. Blow-up solutions. Central to singularity formation phenomena for $2m$th-order reaction-diffusion equations is the concept of finite time blow-up, where the solution of the Cauchy problem with uniformly bounded initial data $u_0(x)$ becomes unbounded at some time $T \in \mathbb{R}_+$ in the sense that $u(x,t)$ exists and is classical on any time-interval $[0,T']$ with $T' \in (0,T)$ and

$$\sup_{x\in\mathbb{R}} |u(x,t)| \to \infty \quad \text{as } t \to T^-. \tag{2.1}$$

Finite time blow-up for higher-order semilinear and quasi-linear parabolic equations is well known from the 1970s. There are several techniques for proving blow-up, including the concavity methods [35], test functions methods [38] (see also [14] and references therein), and an extension of Kaplan's idea based on derivation of an ordinary differential inequality for the first Fourier coefficient of the solutions [21, 10].

2.2. Similarity variables and rescaled PDEs. Finite time blow-up singularities involve a delicate balance between the spatial and temporal derivatives and the reaction terms driving the blow-up. This balance is made naturally apparent by considering the scaling invariance of the underlying PDE. This scaling structure is also important for the numerical methods employed in integrating the full PDE; see section 5.

Because of their semilinear structure, the PDEs (1.6) and (1.7) have similar scaling symmetries, so that (1.7) is invariant with respect to the scaling transformations

$$t \mapsto \lambda t, \quad x \mapsto \lambda^{1/2m} x, \quad u \mapsto \lambda^{-1/(p-1)} u \quad \text{for all } \lambda > 0,$$

while (1.6) is invariant under the group of transformations

$$t \mapsto \lambda t, \quad x \mapsto \lambda^{1/2m} x, \quad u \mapsto u - \ln \lambda.$$

Without loss of generality we may assume that the solution $u(x,t)$ blows up at finite time $t=T$ in the sense of (2.1), and the blow-up set $B[u_0]$ defined by

$$B[u_0]=\big\{x\in I:\ \exists\ \{x_k\}\to x,\ \{t_k\}\to T^-\ \text{such that}\ u(x_k,t_k)\to\infty\big\} \tag{2.2}$$

contains the origin, $0\in B[u_0]$. Motivated by this assumption and looking for invariants of the above groups of transformations, we introduce the following self-similar spatial variable:

$$y=\frac{x}{(T-t)^{1/2m}}:\mathbb{R}\to\mathbb{R},\quad t\in[0,T),$$

and the new time variable

$$\tau=-\ln(T-t):(0,T)\to(\tau_0,\infty)\quad\text{with}\ \ \tau_0=-\ln T.$$

Then for the polynomial nonlinearity we define a new dependent variable (the rescaled solution) $\theta(y,\tau)$ by

$$u(x,t)=(T-t)^{-1/(p-1)}\theta(y,\tau) \tag{2.3}$$

and for the exponential nonlinearity by

$$u(x,t)=-\ln(T-t)+\theta(y,\tau). \tag{2.4}$$

Rescaling (1.7) in terms of the new variables by substituting (2.3), we obtain the following PDE for the rescaled solution θ:

$$\theta_\tau=\boldsymbol{\mathcal{L}}\theta+G_p(\theta),\quad y\in\mathbb{R},\ \tau>\tau_0,\quad\text{where}\ \ G_p(\theta)=|\theta|^{p-1}\theta-\frac{1}{p-1}\,\theta, \tag{2.5}$$

and the linear differential operator $\boldsymbol{\mathcal{L}}$ is given by

$$\boldsymbol{\mathcal{L}}\equiv(-1)^{m+1}D_y^{2m}-\frac{y}{2m}\,D_y. \tag{2.6}$$

Similarly, rescaling (1.6) leads to the PDE

$$\theta_\tau=\boldsymbol{\mathcal{L}}\theta+G_e(\theta),\quad y\in\mathbb{R},\ \tau>\tau_0,\quad\text{where}\ \ G_e(\theta)=e^\theta-1. \tag{2.7}$$

It is important that, unlike the well understood case $m=1$, for any $m>1$ the operators on the right-hand sides *are not potential*, and (2.5) and (2.7) do not possess Lyapunov functions.

2.3. Preliminaries: Local and asymptotic properties of self-similar solutions. Exact (not just asymptotic) self-similar solutions are those that are invariant under the group of transformations, i.e., correspond to suitable stationary solutions $\theta(y)$ that are independent of the rescaled time τ. Any exact self-similar solution to (1.7) takes the form

$$u_S(x,t)=(T-t)^{-1/(p-1)}f(y), \tag{2.8}$$

where $f(y)$ satisfies the ODE

$$\boldsymbol{\mathcal{L}}f+G_p(f)=0\quad\text{in}\ \ \mathbb{R}. \tag{2.9}$$

It is natural to impose the symmetry conditions at the origin

$$f'(0) = f'''(0) = \cdots = f^{(2m-1)}(0) = 0. \tag{2.10}$$

Then a stable (generic) self-similar solution with a suitable similarity profile f in (2.8) means that, for a sufficiently wide and dense subset of global symmetric nonstationary solutions to (2.5), there holds

$$\theta(y,\tau) \to f(y) \quad \text{as} \quad \tau \to \infty$$

in a suitable metric. For such a stable similarity solution (2.8) to have nonvanishing trace in the limit $t \to T^-$ and to rule out constant solutions, we need to impose a special decay condition on $f(y)$ as $y \to \infty$. In particular, we will demand that there exist a finite limit $u(x,t) \to u(x,T^-)$ as $t \to T^-$ for arbitrarily small fixed $|x| > 0$.

2.3.1. Asymptotic behavior at infinity. First we need to describe possible asymptotics of small solutions to (2.5) satisfying $f(y) \to 0$ as $y \to +\infty$. Consider the linearization of (2.9) about $f = 0$,

$$\boldsymbol{\mathcal{L}} f - \frac{1}{p-1} f = 0, \quad y > 0. \tag{2.11}$$

Setting $z = y^\nu$ with $\nu = \frac{2m}{2m-1}$ reduces it to

$$f^{(2m)} - a_1 f' - a_2 z^{-1} f + \mathbf{B}(z) f = 0, \tag{2.12}$$

where $a_1 = (-1)^{m+1} \frac{1}{2m} \nu^{1-2m}$, $a_2 = (-1)^{m+1} \frac{1}{p-1} \nu^{-2m}$, and

$$\mathbf{B}(z) f = \sum_{j=1}^{2m-1} \gamma_j z^{j-2m} f^{(j)}$$

is a linear operator with bounded coefficients as $z \to \infty$, where the first coefficient of derivative f' is of order $O(z^{1-2m})$. By the perturbation theory of higher-order linear ODEs (see Chapters III–V in [12]), we have that the leading terms of exponentially decaying solutions are described by the operator in (2.12) with constant coefficients,

$$f^{(2m)} - a_1 f' = 0. \tag{2.13}$$

Setting $f = e^{pz}$, $p \neq 0$, gives the characteristic equation $p^{2m} - a_1 p = 0$, whence

$$p^{2m-1} = a_1 = \frac{(-1)^{m+1}}{2m\nu^{2m-1}} \equiv \rho_0^{2m-1} (-1)^{m+1}, \quad \text{where } \rho_0 > 0. \tag{2.14}$$

For any $m \geq 1$, there exist $2m-1$ roots $\{p_0, p_1, \dots, p_{2m-2}\}$ given by

$$p_k = \rho_0 e^{i(2k+1)\pi/(2m-1)}, \ m = 2l; \quad p_k = \rho_0 e^{i2\pi k/(2m-1)}, \ m = 2l+1, \tag{2.15}$$

where $m-1$ roots have negative real parts ($\operatorname{Re} p_k < 0$). These correspond to $l \leq k \leq 3l-2$ for even $m = 2l$ and $l+1 \leq k \leq 3l$ for odd $m = 2l+1$. The linearized equation (2.11) has a κ_m-dimensional subspace of exponentially decaying solutions as $y \to \infty$, where $\kappa_m = 2m-3$ for m even and $\kappa_m = 2(m-1)$ for m odd. For the second-order case $m = 1$, it is empty.

On the other hand, (2.12) admits a solution with algebraic decay (rather than exponential) as $z \to \infty$ described by the first-order operator

$$-a_1 f' - a_2 z^{-1} f = 0 \quad \Longrightarrow \quad f(z) = c\, z^{-(2m-1)/(p-1)}.$$

Existence of solutions with such a decay for the perturbed equation (2.12) is established by a standard expansion analysis by calculating solutions via Kummer-type series converging uniformly for $z \gg 1$. For the linearized equation (2.11), the leading order behavior is algebraic,

$$f(y) = C|y|^{2m/(p-1)}(1 + o(1)) \quad \text{as } y \to \infty, \text{ with } C \neq 0. \tag{2.16}$$

In summary, these results yield that (2.11) admits a

$$(\kappa_m{+}1)\text{-dimensional subset of admissible solutions as } y \to \infty. \tag{2.17}$$

Actually, for the nonlinear equation (2.9) we are going to look for profiles $f(y)$ having the algebraic decay (2.16). Then for such similarity solutions (2.8), the limit-time profile is bounded for any $x \neq 0$ and is given by

$$u_S(x, T^-) = C|x|^{-2m/(p-1)}.$$

Asymptotic and numerical computations suggest that the solutions of (2.9) which satisfy (2.16) are *isolated* and that the constant C plays a role of a *nonlinear* eigenvalue. In section 5 we give an asymptotic formula for one value of C valid in a certain limit.

Likewise for (1.6), the self-similar solution is given by

$$u_S(x, t) = -\ln(T - t) + f(y), \tag{2.18}$$

where the function $f(y)$ satisfies the ODE

$$\boldsymbol{\mathcal{L}} f + G_e(f) = 0 \tag{2.19}$$

with the symmetry conditions (2.10). We look for similarity profiles $f(y) \to -\infty$ "slowly" as $y \to \infty$. The limit $\lim_{f \to -\infty} G_e(f) = -1$, so we first consider the "linearized" equation

$$\boldsymbol{\mathcal{L}} f = 1. \tag{2.20}$$

Setting $f(y) = -2m \ln y + g(y)$ for $y > 0$, we obtain

$$\boldsymbol{\mathcal{L}} g = 1 + 2m\, \boldsymbol{\mathcal{L}} \ln y \,=\, 2m(-1)^{m+1} D_y^{2m} \ln y = O(y^{-2m}) \quad \text{as } y \to +\infty. \tag{2.21}$$

As above, the homogeneous equation $\boldsymbol{\mathcal{L}} g = 0$ has a κ_m-dimensional subspace of exponentially decaying solutions. The nonhomogeneous equation (2.21) has solutions $g(y) = C + o(1)$ as $y \to +\infty$, so that (2.17) holds for (2.20), admitting a $\kappa_m + 1$-dimensional subset of solutions satisfying

$$f(y) = -2m \ln |y| + C + o(1) \quad \text{as} \quad y \to \infty. \tag{2.22}$$

In this case the limit-time profile is given by

$$u_S(x, T^-) = -2m \ln |x| + C,$$

where again the constant $C \in \mathbb{R}$ is a certain isolated nonlinear eigenvalue which can be approximated asymptotically.

Obviously, ODEs (2.9) and (2.19) admit constant solutions $f^*_{p,e}$ satisfying

$$G_p(f^*_p) = 0,\ f^*_p = \beta^\beta, \quad \text{and} \quad G_e(0) = 0,\ f^*_e = 0,$$

respectively. The trivial solution $f = 0$ also solves (2.9). The linearizations of the operator $\boldsymbol{\mathcal{L}} + G_p$ about β^β and $\boldsymbol{\mathcal{L}} + G_e$ about 0 coincide and are equal to $\boldsymbol{\mathcal{L}} + I$, where I is the identity operator. The spectral properties of this nonsymmetric operator in a weighted L^2-space play an important part in our analysis and help to describe the perturbation of the solutions from the constant state. They are essential to describe the long time dynamics of both of the PDEs (2.5) and (2.7). We will describe the properties of the linearized operator in the next section.

3. Spectral properties of $\boldsymbol{\mathcal{L}}$ and its adjoint. In this section we study the spectral properties of the linear differential operator $\boldsymbol{\mathcal{L}}$ and its adjoint $\boldsymbol{\mathcal{L}}^*$ given by

$$\boldsymbol{\mathcal{L}}^* = (-1)^{m+1} D_y^{2m} + \frac{1}{2m}\, y \frac{d}{dy} + \frac{1}{2m}\, I. \tag{3.1}$$

Both operators are nonsymmetric and do not admit a self-adjoint extension. To determine the nature of the stability of the constant solution and also to apply the Fredholm alternative to computing asymptotic solutions of the ODEs, it is necessary to determine the spectrum and corresponding eigenfunctions of both $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{L}}^*$. We present some results from [15] and [20] which describe these.

3.1. The fundamental solution. We start by determining the spectrum and the eigenfunctions of the adjoint operator $\boldsymbol{\mathcal{L}}^*$. In order to find the null eigenfunction, we begin with the fundamental solution of the corresponding linear $2m$th-order parabolic operator. Consider the linear equation

$$u_t = (-1)^{m+1} D_x^{2m} u \quad \text{in } \mathbb{R} \times \mathbb{R}_+. \tag{3.2}$$

The fundamental solution of (3.2) has the standard self-similar form

$$b(x,t) = t^{-1/2m} F(y), \quad y = \frac{x}{t^{1/2m}}. \tag{3.3}$$

Substituting $b(x,t)$ into (3.2) yields that the radially symmetric profile $F(y)$ is the unique even square integrable solution of the linear ODE

$$\boldsymbol{\mathcal{L}}^* F = 0 \quad \text{in } \mathbb{R} \tag{3.4}$$

and is a null eigenfunction of $\boldsymbol{\mathcal{L}}^*$. Taking a Fourier transform leads to

$$F(y) = \alpha \int_0^\infty e^{-s^{2m}} \cos(sy) ds. \tag{3.5}$$

The coefficient α is chosen to normalize $\int F = 1$, so that

$$\alpha = \left(\int_0^\infty \int_0^\infty e^{-s^{2m}} \cos(sy)\, ds\, d\eta \right)^{-1}.$$

The rescaled kernel $F(\eta)$ then satisfies a standard pointwise estimate (see [16])

$$|F(y)| \le d_1 e^{-d_2 |y|^\nu} \quad \text{in } \mathbb{R},$$

where d_1 and d_2 are positive constants. Applying the Fourier transform to (3.2) and performing the rescaling, we have

$$\mathcal{F}(b(\cdot,t))(\xi) = e^{-\xi^{2m}t} \quad \text{and} \quad \hat{F}(\omega) = \mathcal{F}(F(\cdot))(\omega) = e^{-\omega^{2m}}. \tag{3.6}$$

3.2. The discrete real spectrum and eigenfunctions of the adjoint operator $\boldsymbol{\mathcal{L}^*}$. We describe the spectrum $\sigma(\boldsymbol{\mathcal{L}}^*)$ of the adjoint operator in the weighted space $L^2_{\rho^*}(\mathbf{R})$ with the exponential weight

$$\rho^*(y) = e^{a|y|^\nu} > 0 \quad \text{in } \mathbb{R}, \quad \nu = \frac{2m}{2m-1}, \tag{3.7}$$

where $a < 2d_2$ is a sufficiently small positive constant. Denoting by $\langle\cdot,\cdot\rangle_*$ and $\|\cdot\|_*$ the corresponding inner product and induced norm, respectively, we introduce a Hilbert space of functions $H^{2m}_{\rho^*}(\mathbb{R})$ with the inner product and norm

$$\langle v, w\rangle_* = \int_{\mathbb{R}} \rho^*(y) \sum_{k=0}^{2m} D^k v(y)\, \overline{D^k w(y)}\, dy, \qquad \|v\|_*^2 = \int_{\mathbb{R}} \rho^*(y) \sum_{k=0}^{2m} |D^k v(y)|^2\, dy.$$

Then $H^{2m}_{\rho^*}(\mathbb{R}) \subset L^2_{\rho^*}(\mathbb{R}) \subset L^2(\mathbb{R})$, and $\boldsymbol{\mathcal{L}}^*$ is a bounded linear operator from $H^{2m}_{\rho^*}(\mathbb{R})$ to $L^2_{\rho^*}(\mathbb{R})$. With these definitions, the spectral properties of the operator $\boldsymbol{\mathcal{L}}$ are given by the following lemma.

LEMMA 3.1. (i) *The spectrum of $\boldsymbol{\mathcal{L}}^*$ (and hence of $\boldsymbol{\mathcal{L}}$) comprises real simple eigenvalues only,*

$$\sigma(\boldsymbol{\mathcal{L}}^*) = \left\{ \lambda_k = -\frac{k}{2m},\ k = 0, 1, 2, \dots \right\}. \tag{3.8}$$

(ii) *The eigenfunctions $\psi^*_k(y)$ are given by*

$$\psi^*_k(y) = \frac{(-1)^k}{\sqrt{k!}} D^k F(y) \tag{3.9}$$

and form a complete subset in $L^2(\mathbf{R})$ and in $L^2_{\rho^*}(\mathbf{R})$. (Here F is as defined in (3.5).)

(iii) *The resolvent $(\boldsymbol{\mathcal{L}}^* - \lambda I)^{-1} : L^2_{\rho^*}(\mathbb{R}) \to L^2_{\rho^*}(\mathbb{R})$ is a compact integral operator.*

Most importantly, the operators $\boldsymbol{\mathcal{L}}^*$ and $\boldsymbol{\mathcal{L}}$ have zero Morse index (no eigenvalues have positive real part).

3.3. The polynomial eigenfunctions of the operator $\boldsymbol{\mathcal{L}}$. We now consider the operator (2.6) in the weighted space $L^2_\rho(\mathbb{R})$ ($\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ are the inner product and the norm), with the exponentially decaying weight function

$$\rho(y) \equiv \frac{1}{\rho^*(y)} = e^{-a|y|^\nu} > 0, \tag{3.10}$$

and ascribe to $\boldsymbol{\mathcal{L}}$ the domain $H^{2m}_\rho(\mathbb{R})$ that is dense in $L^2_\rho(\mathbb{R})$. Then $\boldsymbol{\mathcal{L}} :\ H^{2m}_\rho(\mathbb{R}) \to L^2_\rho(\mathbb{R})$ is a bounded linear operator, $\boldsymbol{\mathcal{L}}^*$ is adjoint to $\boldsymbol{\mathcal{L}}$, and, denoting by $\langle\cdot,\cdot\rangle$ the inner product on $L^2(\mathbb{R})$, we have

$$\langle \boldsymbol{\mathcal{L}} v, w\rangle = \langle v, \boldsymbol{\mathcal{L}}^* w\rangle \quad \text{for any } v \in H^{2m}_\rho(\mathbb{R}),\ w \in H^{2m}_{\rho^*}(\mathbb{R}). \tag{3.11}$$

The eigenfunctions of $\boldsymbol{\mathcal{L}}$ take a particularly simple polynomial form and are as follows.

LEMMA 3.2. (i) *The eigenfunctions $\psi_k(y)$ of $\boldsymbol{\mathcal{L}}$ are polynomials in y of order k given by*

$$\psi_k(y) = \frac{1}{\sqrt{k!}} \sum_{j=0}^{\lfloor -\lambda_k \rfloor} \frac{(-1)^{mj}}{j!} D^{2mj} y^k, \quad k = 0, 1, 2, \dots , \tag{3.12}$$

and form a complete subset in $L^2_\rho(\mathbb{R})$. (Here $\lfloor \cdot \rfloor$ denotes the integer part.)

(ii) *$\boldsymbol{\mathcal{L}}$ has compact resolvent $(\boldsymbol{\mathcal{L}} - \lambda I)^{-1}$ in $L^2_\rho(\mathbb{R})$.*

COROLLARY 3.3. *With the definition* (3.9) *of the adjoint basis, integrating by parts, we have that the orthonormality condition holds*

$$\langle \psi_k, \psi_l^* \rangle = \delta_{k,l} \quad \textit{for any } k, l \geq 0, \tag{3.13}$$

where $\delta_{\kappa,l}$ is the Kronecker delta.

COROLLARY 3.4. *If $m = 2$, then there are coefficients α_j (depending on k) such that, for $k = 4r + 2$ and $k = 4r$,*

$$\psi_{4r+2} = y^2 \sum_{j=0}^{r} \alpha_j y^{4j} \quad \textit{and} \quad \psi_{4r} = \sum_{j=0}^{r} \alpha_j y^{4j}, \qquad \alpha_0 \neq 0.$$

For example, if $m = 2$ (a case we consider in detail first), then the first four even eigenfunctions are

$$\psi_0(y) = 1, \quad \psi_2(y) = \frac{y^2}{\sqrt{2}}, \quad \psi_4(y) = \frac{(y^4 + 24)}{\sqrt{24}}, \quad \psi_6(y) = \frac{y^2(720 + y^4)}{\sqrt{6!}}, \tag{3.14}$$

with corresponding eigenvalues $0, -\frac{1}{2}, -1, -\frac{3}{2}$.

4. Local asymptotic analysis: Invariant subspaces and bifurcation points. In this section we use the spectral properties of the linearized operators to determine the local stability of the constant solutions of the rescaled PDEs (2.5) and (2.7). We begin with the linearized stability analysis and describe invariant subspaces.

4.1. Invariant eigenspaces. Since the nonlinearities under consideration satisfy $G_p'(\beta^\beta) = G_e'(0) = 1$, let us consider solutions of (2.5) and (2.7) as perturbations of the constant solution of the form

$$\theta(y, \tau) = f^* + g(y, \tau) \quad \text{with} \quad \|g\| \ll 1.$$

In both cases g satisfies a perturbed PDE

$$g_\tau = (\boldsymbol{\mathcal{L}} + I)g + \bar{G}(g), \quad \text{where } \bar{G}(g) = G(f^* + g) - g, \tag{4.1}$$

with a quadratic nonlinear perturbation $\bar{G}$,

$$\bar{G}(g) = c_2 g^2 + c_3 g^3 + \cdots \quad \text{as} \quad g \to 0, \tag{4.2}$$

and the coefficients depending on the nonlinearity, $c_2 = \frac{1}{2}$, $c_3 = \frac{1}{6}, \dots$ for G_e and $c_2 = \frac{1}{2}p(p-1)^{1/(p-1)}$, $c_3 = \frac{1}{6}p(p-1)^{2/(p-1)}(p-2), \dots$ for G_p.

In what follows, we restrict our attention to symmetric in x solutions $u = u(|x|, t)$ and hence to symmetric in y rescaled solutions $\theta = \theta(|y|, \tau)$ and $g = g(|y|, \tau)$. In

the space $L^2_{0,\rho}(\mathbb{R})$ of symmetric functions, it follows from (3.8) that $\boldsymbol{\mathcal{L}} + I$ has the spectrum

$$\sigma(\boldsymbol{\mathcal{L}} + I) = \left\{ \tilde{\lambda}_k = 1 - \frac{k}{2m},\ k = 0, 2, 4, \dots \right\}. \tag{4.3}$$

Let $\tilde{L}^2_{0,\rho} \subseteq L^2_{0,\rho}$ be the subspace of eigenfunction expansions, where $\{\psi_k\}$ is closed, obtained as the closure of the subset of finite sums $\{v = \sum c_k \psi_k\}$ in the norm $\|\cdot\|$ [15]. Then

$$\tilde{L}^2_{0,\rho}(\mathbb{R}) = E^u(0) \oplus E^c(0) \oplus E^s(0),$$

where $E^u(0)$, $E^c(0)$, and $E^s(0)$ are the unstable, center, and stable subspaces of $\boldsymbol{\mathcal{L}} + I$ given by

$$\begin{aligned} E^u(0) &= \mathrm{Span}\{\psi_0, \psi_2, \dots, \psi_{2m-2}\}, \\ E^c(0) &= \mathrm{Span}\{\psi_{2m}\}, \\ E^s(0) &= \mathrm{Span}\{\psi_{2m+2},\ \psi_{2m+4}, \dots\}. \end{aligned}$$

In particular, the dimension of the unstable subspace is precisely m.

Consider the two one-dimensional unstable subspaces corresponding to positive eigenvalues of the operator $\boldsymbol{\mathcal{L}} + I$, namely

$$\tilde{\lambda}_0 = 1,\ \ \psi_0(y) = 1 \quad \text{and} \quad \tilde{\lambda}_2 = 1 - \frac{1}{m},\ \ \psi_2(y) = \frac{y^2}{\sqrt{2}}. \tag{4.4}$$

As is usual in blow-up problems, the first unstable mode with $k = 0$ corresponds to the instability of blow-up behavior with respect to perturbations of the blow-up time T.

In contrast, the second mode with $k = 2$ describes an actual instability of the constant solution which is in the direction of $\psi_2(y)$ and is in the space of rescaled solutions having the same fixed blow-up time T. From our asymptotic calculations and numerical experiments we expect that the orbits that arise from the instability of the constant solution in the PDEs (2.7) and (2.5) when $m > 1$ are uniformly bounded and stabilize to one of the self-similar solutions. Namely, the first such unstable mode with $\tilde{\lambda}_2 = 1 - \frac{1}{m} > 0$ gives a heteroclinic connection of f^* with a nonconstant stable (generic) similarity profile $f_1(y)$.

It is significant that when $m = 1$, there is no such unstable mode. In contrast, the dimension of the unstable subspace is one corresponding only to the change in the blow-up time. The eigenfunction ψ_2 then has eigenvalue zero, and the behavior of the perturbations of the constant solution must be studied on the center manifold. It is this which leads to the approximate self-similar behavior (1.5) described in the introduction.

Before performing some formal invariant manifold analysis for higher-order PDEs, note that the main properties of connecting equilibria and transversality of intersections of the corresponding stable and unstable manifolds are known for the one-dimensional second-order parabolic equations

$$u_t = u_{xx} + f(x, u) \quad \text{in } (0,1) \times \mathbb{R}_+, \quad u = 0 \text{ at } x = 0, 1 \text{ for } t > 0,$$

and were obtained in [29, 3] and [11] using Sturm's theorem on the nonincrease of the number of zeros (intersections) of solutions to linear second-order parabolic equations. This Sturmian property is not true for the fourth- and higher-order parabolic

equations (owing to the lack of a maximum principle in these cases), for which there are some particular results (see references in [39]), and in general the structure of connecting orbits remains an important open problem.

4.2. The center subspace. Consider the center subspace $E^c(0)$ in the case of general m. From Lemma 3.2, it follows that the null eigenfunction of the operator $\boldsymbol{\mathcal{L}}$ is given by ψ_{2m} so that

$$\tilde{\lambda}_{2m} = 0 \quad \text{and} \quad \psi_{2m}(y) = \frac{[y^{2m} + (-1)^m (2m)!]}{\sqrt{(2m)!}}\,. \tag{4.5}$$

We now present a simple calculation showing that the behavior on the center manifold is semistable.

PROPOSITION 4.1. *Let $g(\cdot,\tau) \in H^{2m}_{0,\rho}(\mathbb{R})$ exhibit the center subspace dominance, i.e.,*

$$g(\cdot,\tau) = a_{2m}(\tau)\psi_{2m}(\cdot) + w(\cdot,\tau) \quad \textit{for } \tau \gg 1, \tag{4.6}$$

where $w(\cdot,\tau) \in \mathcal{L}^{\perp}\{\psi_{2m}\}$ and $w(\cdot,\tau) = o(\|g(\cdot,\tau)\|) = o(|a_{2m}(\tau)|)$ as $\tau \to \infty$. Then

$$a_{2m}(\tau) = -\frac{1}{\gamma_0 \tau}(1 + o(1)) \quad \textit{as } \tau \to \infty, \textit{ where } \gamma_0 = c_2 \langle (\psi_{2m})^2, \psi^*_{2m} \rangle \neq 0. \tag{4.7}$$

It follows from (4.7) that $a_{2m}(\tau)$ cannot change sign in any neighborhood of $\tau = \infty$, meaning a one-sided instability of the center manifold behavior.

Proof. We look for a solution of (4.1) via a uniformly convergent eigenfunction expansion

$$g(\cdot,\tau) = \sum a_k(\tau)\psi_k(\cdot). \tag{4.8}$$

Substituting this expression into (4.1) and multiplying by ψ^*_k in $L^2(\mathbb{R})$, we arrive at a dynamical system for the expansion coefficients

$$\dot{a}_k = \tilde{\lambda}_k a_k + \langle \bar{G}(g), \psi^*_k \rangle, \quad k = 0, 2, \dots\,. \tag{4.9}$$

Consider an equation for the coefficient a_{2m} with $\tilde{\lambda}_{2m} = 0$. In view of assumption (4.6) and (4.2), assuming that $|a_{2m}(\tau)| \ll 1$, it follows that

$$\dot{a}_{2m} = (\gamma_0 + o(1)) a^2_{2m} \quad \text{for } \tau \gg 1. \tag{4.10}$$

Calculating γ_0 by using the adjoint eigenfunction $\psi^*_{2m} = D^{2m}_y F/\sqrt{2m!}$ and (4.5), we obtain that

$$\gamma_0 = c_2 (-1)^{m+1} \sqrt{(2m!)} \left(\frac{(4m)!}{[(2m)!]^2} - 2 \right). \tag{4.11}$$

Integrating (4.10) as a standard ODE, we deduce that any small solution for $\tau \gg 1$ has the asymptotic behavior (4.7). ∎

It follows from the quadratic "ODE" (4.10) that the center manifold behavior exhibits a typical semistable ("saddle-node") structure. Because the constant profile β^β is only semistable, small perturbations in the unstable direction may evolve to self-similar solutions. We present some evidence for this conjecture and the role of the parity of m in sections 5 and 6.

In view of known spectral and sectorial properties of operators $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{L}}^*$ [15, 20], we expect that the center (and stable; see section 7) manifold behavior can be justified by the invariant manifold theory in interpolation spaces; see [36, Chapter 9].

4.3. Bifurcation points. In this subsection we extend the ODEs (2.9) and (2.19) for similarity profiles and consider the family of ODEs with a parameter $\mu \geq 0$:

$$(-1)^{m+1} D_y^{2m} f - \mu f' y + G(f) = 0 \quad \text{for } y > 0 \text{ with conditions (2.10).} \tag{4.12}$$

Recall that, for single-point blow-up, we need to impose an extra condition (of the type (2.16) or (2.22) with $\frac{1}{2m} \mapsto \mu$) on the decay of $f(y)$ at infinity.

If we take $\mu = \frac{1}{2m}$ and the appropriate nonlinearity, $G = G_p$ or G_e, then we obtain the ODEs (2.9) and (2.19) for the rescaled self-similar profiles. More generally, suitable solutions of (4.12) depend smoothly upon $\mu \approx \frac{1}{2m}$ and coincide with the self-similar solutions when $\mu = \frac{1}{2m}$. In either case we define a corresponding linearized operator $\boldsymbol{\mathcal{L}}_\mu$ by

$$\boldsymbol{\mathcal{L}}_\mu = (-1)^{m+1} D_y^{2m} - \mu y D_y + I \equiv \boldsymbol{\mathcal{L}} + \left(1 - \mu + \frac{1}{2m}\right) I. \tag{4.13}$$

Changing the independent variable to

$$y = \frac{z}{(2m\mu)^{1/2m}}, \tag{4.14}$$

we have

$$\frac{1}{2m\mu} \boldsymbol{\mathcal{L}}_\mu = (-1)^{m+1} D_z^{2m} - \frac{1}{2m} z \frac{d}{dz} + \frac{1}{2m\mu} I \equiv \boldsymbol{\mathcal{L}} + \frac{1}{2m\mu} I. \tag{4.15}$$

Hence $\boldsymbol{\mathcal{L}}_\mu : H^{2m}_{0,\rho}(\mathbb{R}) \to L^2_{0,\rho}(\mathbb{R})$ is a bounded linear operator (with a change in the coefficient a in the weight function (3.10) if necessary). By Lemma 3.1, the spectrum $\boldsymbol{\mathcal{L}}_\mu$ in the space $L^2_{0,\rho}(\mathbb{R})$ of radial functions is given by

$$\sigma(\boldsymbol{\mathcal{L}}_\mu) \equiv 2m\mu\, \sigma\left(\boldsymbol{\mathcal{L}} + \frac{1}{2m\mu} I\right) = \{1 - 2\mu l,\ l = 0, 1, 2, \dots\}, \tag{4.16}$$

with eigenfunctions ψ_{2l} as before, rescaled according to the transformation (4.14).

We next compute bifurcation points from the constant solution f^*. Since the weight function (3.10) is exponentially decaying as $y \to \infty$, in general, the inclusion $f \in H^{2m}_\rho$ does not imply the boundedness of f, unlike the adjoint case of the increasing weight (3.7), where $H^{2m}_\rho \subset C$. Nonlinearity $G(f)$ is not uniformly Lipschitz continuous on bounded subsets from H^{2m}_ρ. Therefore, we truncate the nonlinearity in (4.12) by replacing G by G_n, which satisfies

$$G_n(f) \equiv G(f) \quad \text{for } |f| \leq n, \ \ n = 1, 2, \dots,$$

and $G_n(f)$ is sufficiently smooth and uniformly Lipschitz continuous in $\mathbb{R}$. For $G = G_e$, we need only perform the truncation for $f > n$. We have

$$G_n(f) \to G(f) \quad \text{as } n \to \infty \text{ uniformly on compact subsets.}$$

Replacing the full problem by the truncated one

$$(-1)^{m+1} D_y^{2m} f - \mu y f' + G_n(f) = 0 \tag{4.17}$$

is permissible because we are interested in bounded solutions f, for which the nonlinearities $G_p(f)$ and $G_e(f)$ have finite range.

PROPOSITION 4.2. *For any $m \geq 1$, the values of μ for which the spectrum of $\boldsymbol{\mathcal{L}}_\mu$ contains zero,*

$$1 - 2\mu l = 0 \implies \mu_l = \frac{1}{2l}, \quad l = 1, 2, \dots , \tag{4.18}$$

are bifurcation points for problem (4.17).

Proof. Using rescaling (4.14) and setting $f = f^* + g$, equation (4.17) takes the form

$$(\boldsymbol{\mathcal{L}} - I)g = \tilde{\mu} g + (1 + \tilde{\mu}) G_n(g), \quad \text{where } \tilde{\mu} = -1 - \frac{1}{2m\mu}. \tag{4.19}$$

Consider the Hammerstein operator $(\boldsymbol{\mathcal{L}} - I)^{-1} G_n$. By Lemma 3.2, $(\boldsymbol{\mathcal{L}} - I)^{-1}$ is a compact operator in $L^2_{0,\rho}$ with simple eigenvalues $\{-1/(1 + \frac{l}{m}) \leq -1,\ l = 0, 1, 2, \dots\}$. By construction, G_n is uniformly Lipschitz continuous, $|G_n(g)| \leq C_1 + C_2|g|$ in $\mathbb{R}$, and hence $G_n : L^2_{0,\rho} \to L^2_{0,\rho}$. Therefore, the product $(\boldsymbol{\mathcal{L}} - I)^{-1} G_n$ is a compact operator in $L^2_{0,\rho}$; see [34, Chapter V]. Hence, in the nonlinear integral equation written as a fixed point problem

$$g = \mathbf{A}(g, \tilde{\mu}) \equiv \tilde{\mu} (\boldsymbol{\mathcal{L}} - I)^{-1} g + (1 + \tilde{\mu}) (\boldsymbol{\mathcal{L}} - I)^{-1} G_n(g), \tag{4.20}$$

bifurcation from the origin occurs iff $\tilde{\mu}$ coincides with characteristic values of $(\boldsymbol{\mathcal{L}} - I)^{-1}$ (simple eigenvalues of $\boldsymbol{\mathcal{L}} - I$), i.e., at $\tilde{\mu}_l = -1 - \frac{l}{m}$ (see [34]). This yields (4.18). ∎

Passing to the limit $n \to \infty$, some of the bifurcation sub-branches (which are not of physical interest) may disappear, so that we always need to check which sub-branches are available for $n = \infty$. On the other hand, it is interesting to know for which values of μ, less or greater than μ_l, there exist nonconstant solutions and how many. Since the spectrum of the Frechet derivative $\mathbf{A}'(0, \tilde{\mu}_l)$,

$$\sigma(\mathbf{A}'(0, \tilde{\mu}_l)) = \left\{ \frac{(1 + \frac{l}{m})}{(1 + \frac{k}{m})},\ k = 0, 1, 2, \dots \right\}, \tag{4.21}$$

always contains 1 (for $k = l$), the local asymptotic behavior of bifurcation branches for $\mu \approx \mu_l$ is a delicate problem, and often there exist at least two solutions even in the cases of analytic nonlinearities; see a general theory in [44]. Therefore we will need an extra matching analysis to specify "correct" branches, which have the required behavior at infinity and hence correspond to single-point blow-up similarity profiles.

It is important to mention the main reason for extending the operator (2.6) in (2.9) and (2.19) to the operator in (4.12) parameterized by μ. Setting $\mu = 0$, in the case of the polynomial nonlinearity with $G = G_p$, we recover a well-studied Hamiltonian system (see [2] and the book [39]), and the solutions considered in this case can, in principle, be followed as μ increases to the physically important value of $\frac{1}{2}m$. Alternatively, by setting μ close to the bifurcation points (4.18), we can construct asymptotic descriptions of solutions that are local perturbations of the constant solution. This calculation is presented in the next section. Once we have constructed such solutions, we may again extend varying μ to determine branches of solutions that persist until the value $\mu_m = \frac{1}{2}m$.

In other words, problem (4.12) for $\mu \in [0, \frac{1}{2m}]$ describes the *transition* phenomenon between Hamiltonian systems for $\mu = 0$, with a potential and leading self-adjoint differential operators, and the singularity formation problem for $\mu = \frac{1}{2m}$, with no potential structure or symmetry properties of operators involved.

4.4. Conjecture on existence of self-similar solutions. For any $m > 1$, the question of the solvability of problem (4.12) for $\mu = \frac{1}{2}m$ (with the appropriate decay of $f(y)$ at infinity) and of the number of solutions seem to be very hard. Proving solvability is a multidimensional problem of matching of the $(\kappa_m{+}1)$-dimensional bundle of orbits as $y \to \infty$ (see (2.17)) with the m-dimensional bundle at $y \approx 0$ depending on the parameters $\{f(0), f''(0), \dots, f^{(2m-2)}(0)\}$ (a multidimensional shooting problem whose complexity increases dramatically as m increases). For $m = 1$, such a problem for quasi-linear equations (1.3) is well understood in one dimension (see [40] and [8]), though a complete proof of the number, finite or infinite, of solutions for equations in $\mathbb{R}$ and in $\mathbb{R}^N$ is still missing.

We now use the above local bifurcation analysis to estimate the number of solutions from below. In view of Proposition 4.2, there exist branches of solutions $f(y;\mu)$ emanating at $\mu = \frac{1}{2l}$ from constant solutions $f = f^*$ for each value of $l = 1, \dots, m-1$ (though we still do not know which bifurcation branches correspond to single point blow-up profiles with required decay at infinity). In particular, if we fix m, then a self-similar solution occurs at $\mu_m = \frac{1}{2m}$. However, there are $m-1$ bifurcation points at $\mu_l = \frac{1}{2l} > \mu_m = \frac{1}{2m}$ for $l = 1, \dots, m-1$. The numerical calculations of section 6 strongly imply that each such bifurcation leads to a branch of solutions $f(y)$ with far-field behavior of the type (2.16) or (2.22) persisting until μ_m, giving rise to a self-similar solution. Furthermore, due to the semistability properties of the center manifold patterns (see further comments in section 5), we expect from the observations of the previous section that there is an additional solution of the ODE when m is even. This detail is also supported by both the asymptotic calculations presented in section 5 and the numerical calculations of section 6. Combining these observations, let us state the following conjecture suggested by our understanding of the dynamics of the linearized operator, asymptotic constructions, and a number of numerical experiments.

CONJECTURE 4.3. *For all $m > 1$, the problems* (2.19) *and* (2.9) *have at least* $2\lfloor \frac{m}{2} \rfloor$ *(self-similar) solutions.*

Hence, we conjecture that the nonexistence of exact self-similar blow-up solutions is a feature only of the second-order semilinear equations, not of all the semilinear equations of the forms (1.6) and (1.7). This conjecture is indeed a lower bound and is based only on properties of the linear operator presented in this paper. In fact, we expect that there are $m(m-1)$ solutions. This estimate is topological and characterizes a typical matching of two multidimensional bundles at $y = \infty$ and $y = 0$, respectively, in the presence of sufficiently strong oscillatory character of the ODE; see further results below.

Further, we note that bifurcations in the limit problem (4.12) hold for arbitrary $L^2_{0,\rho}$-solutions of (4.19), not necessarily satisfying the appropriate decay conditions at infinity. There may also exist nonconstant solutions that correspond to stabilization as $y \to \infty$ to another equilibrium,

$$f(y) \to \beta^\beta \ \text{ for } \ G = G_p \quad \text{and} \quad f(y) \to 0 \ \text{ for } \ G = G_e. \tag{4.22}$$

One can see from (2.8) and (2.18) that these self-similar solutions create *global blow-up*, where

$$u(x,t) \to \infty \quad \text{as } \ t \to T^- \ \text{ uniformly in } \ \mathbb{R}. \tag{4.23}$$

Such behavior is unavailable for $m = 1$ as the dimension of the stable manifold about f^* is $2(m-1)$ for m odd. For $m > 1$, no such solutions have yet been detected, numerically or otherwise.

5. The asymptotic behavior of the solutions close to the bifurcation points. In this section we again consider μ to be a continuous parameter in (4.12) and construct an asymptotic description of solutions $f(y;\mu)$ (with the appropriate decay at infinity; see a precise statement below) for μ close to the bifurcation points at $\mu_l = \frac{1}{2l}$. We set

$$\mu = \mu_l + \sigma_l \varepsilon \quad \text{with } 0 < \varepsilon \ll 1 \text{ and } \sigma_l^2 = 1, \tag{5.1}$$

and look for solutions to the ODEs in $\mathbb{R}_+$ for $l = 1, 2, \dots,$

$$(-1)^{m+1} f^{(2m)} - (\mu_l + \sigma_l \varepsilon)\, y f' + G_p(f) = 0, \tag{5.2}$$

$$(-1)^{m+1} f^{(2m)} - (\mu_l + \sigma_l \varepsilon)\, y f' + G_e(f) = 0. \tag{5.3}$$

We seek solutions with symmetry conditions (2.10) satisfying the decay condition

$$f(y) = C y^{-1/(p-1)\mu}(1 + o(1)) \quad \text{or} \quad f(y) = -\frac{1}{\mu} \ln y + C + o(1) \quad \text{as } y \to +\infty. \tag{5.4}$$

Here $\sigma_l = \pm 1$ indicates the direction that the branch departs from the constant solution, which we shall show depends upon l and m. Because of the polynomial structure of the eigenfunctions of the linear operator $\boldsymbol{\mathcal{L}}$ (and hence of $\boldsymbol{\mathcal{L}}_\mu$), the asymptotic calculations are similar in spirit for each bifurcation point, $\mu = \frac{1}{2l}$, although for each order $2m$ of the differential operator there are m slightly different types of expansion. As such, we will illustrate the calculations by first considering the case $m = 2$ close to arbitrary bifurcation points, then close to the particular bifurcation points of interest to fourth-order PDEs, namely, $\mu_1 = \frac{1}{2}$ and $\mu_2 = \frac{1}{4}$. Lastly, we construct solutions close to the specific bifurcation points $\mu_m = \frac{1}{2m}$ for the case of general m to complement the calculations of the center manifold behavior described in the previous section and our conjecture regarding the existence of self-similar solutions of the ODE when $\mu = \frac{1}{2m}$.

5.1. The case of fourth-order ODEs: $m = 2$. We shall first consider the two ordinary differential problems, namely finding the slowly growing/bounded solutions of the fourth-order equations with $l = 1, 2, \dots,$

$$-f'''' - (\mu_l + \sigma_l \varepsilon)\, y f' + |f|^{p-1} f - \frac{1}{p-1} f = 0, \tag{5.5}$$

$$-f'''' - (\mu_l + \sigma_l \varepsilon)\, y f' + e^f - 1 = 0. \tag{5.6}$$

The calculation proceeds by identifying three key regions in which asymptotic solutions of three different scalings of the above equations are derived. The three different asymptotic descriptions of the solutions are then matched together. The first region is given by considering solutions for which $\varepsilon^\gamma y$ is small and where

$$\gamma = \begin{cases} \frac{1}{4l} \text{ for } l \text{ odd}, \\ \frac{1}{2l} \text{ for } l \text{ even}. \end{cases} \tag{5.7}$$

Here the solution is near constant, and we can express the solution in terms of the eigenfunctions of the linear operator $\boldsymbol{\mathcal{L}}_\mu$ in (4.13). Next is a midrange region, for which $\varepsilon^{-\gamma} < y < e^{1/\varepsilon}$, where the appropriately rescaled differential equations reduce to an integrable first-order equation. Lastly, there is the region $\{y > e^{1/\varepsilon}\}$, where the solution satisfies the far-field behavior (5.4).

5.1.1. The behavior of $f(y)$ for $\varepsilon^\gamma y \ll 1$. We begin by seeking solutions of (5.5) and (5.6), which are valid for small $\varepsilon^\gamma|y|$ and which are close to the constant solutions of the respective nonlinearities. Consider the corresponding equation (4.20) for fixed points. Since, by (4.21), 1 is an eigenvalue of $\mathbf{A}'(0,\tilde{\mu}_l)$ with the one-dimensional eigenspace E_l, according to the general branching theory [44, Chapter 5], in this special case we seek solutions of the form of the rational series

$$f(y) = f_0 + \varepsilon^q f_1(y) + \varepsilon^{2q} f_2(y) + \cdots, \tag{5.8}$$

where we define $f_0 = f^*$. For convenience, we perform this equivalent expansion analysis directly for the ODEs and avoid using the integral equation (4.20) with compact operators. The exponent $q = 1/n$ with an unknown integer $n \geq 1$ is to be determined from the solvability of the corresponding nonlinear systems on the expansion coefficients (the branching equation). Since $\dim E_l = 1$, the branching equation is always one-dimensional. Note that, for analytic nonlinearities, i.e., (5.6) with any odd p and (5.5), in the case of one- (or two-) dimensional eigenspace E_l, finite solvability of such systems (existence of a finite number of solutions) implies convergence of the series (5.8) for sufficiently small ε, although we can expect there to be at least two different bifurcating branches of solutions; see [44, pp. 209–211]. We then determine the correct branch by matching to solutions with the appropriate decay properties at infinity.

The rational power q of the order parameter depends on the coefficients of the branching equation, which are different depending on whether l is even or odd. Substituting the expansion (5.8) into the ODEs leads, at lowest order, to an ODE for $f_1(y)$ of the form

$$\boldsymbol{\mathcal{L}}_{1/2l} f_1 \equiv -f_1'''' - \frac{1}{2l}\, y f_1' + f_1 = 0.$$

Accordingly, the leading order approximation to $f - f_0$ is given by a linear multiple of the eigenfunction $\psi_{2l}((2/l)^{1/4}y)$; see (4.14). From the description of the spectrum of the operator $\boldsymbol{\mathcal{L}}$ given in Lemma 3.1, using Corollary 3.4, we know that (as $m = 2$) the transformed operator $\boldsymbol{\mathcal{L}}_{1/2l}$ has null eigenfunctions ψ_{2l} which are polynomials and which take the form

$$\psi_{2l}(y) = y^2 \sum_{j=0}^{(l-1)/2} \alpha_j y^{4j} \quad \text{for } l \text{ odd} \qquad \text{and} \qquad \psi_{2l}(y) = \sum_{j=0}^{l/2} \alpha_j y^{4j} \quad \text{for } l \text{ even},$$

as defined by (3.12) after the change of variable $y \mapsto y(2/l)^{1/4}$.

The difference between the cases of l even and l odd is as follows. In the asymptotic expansion, the higher powers of $f_1(y)$ become forcing terms to equations involving the operator $\boldsymbol{\mathcal{L}}_{1/2l} f_j$. In the case of odd l, these terms will always be polynomials in y^4. These may have no contribution, which resonates with the null eigenfunction ψ_{2l} of $\boldsymbol{\mathcal{L}}$. In contrast, the powers of $f_1(y)$ for even l will always have contributions, which resonate with $\psi_{2l}(y)$. As a consequence, the cases l even and l odd lead to distinctly different forms of asymptotic expansion, in particular $q = \frac{1}{2}$ for odd l and $q = 1$ for even l. In other words, for l even and odd the branching equation changes its type. Generically, there will be m distinct expansions in powers of $\varepsilon^{i/m}$, $i = 1, 2, \ldots, m$; see a general classification in [44, section 12].

A. *The case of $m = 2$ and l odd.* We take $l = 2r+1$ so that the bifurcation point is at $\mu = 1/(4r+2)$, $r = 0, 1, \ldots$. We express $f(y)$ as an asymptotic expansion ($q = \frac{1}{2}$)

$$f = f_0 + \varepsilon^{1/2} f_1 + \varepsilon f_2 + \varepsilon^{3/2} f_3 + \cdots. \tag{5.9}$$

This expansion corresponds to the case of the branching equation described in Theorem 12.2 in [44], where there exist two solutions either for $\mu < \mu_l$ or for $\mu > \mu_l$. Substituting the expansion (5.9) into either (5.5) or (5.6) gives a sequence of ODE problems of the form

$$O(\varepsilon^{1/2}): \quad \boldsymbol{\mathcal{L}}_{1/2l} f_1 \equiv -f_1'''' - \frac{1}{4r+2}\, y f_1' + f_1 = 0, \tag{5.10}$$

$$O(\varepsilon): \quad \boldsymbol{\mathcal{L}}_{1/2l} f_2 = -c_2 f_1^2, \tag{5.11}$$

$$O(\varepsilon^{3/2}): \quad \boldsymbol{\mathcal{L}}_{1/2l} f_3 = \sigma_l y f_1' - 2c_2 f_1 f_2 - c_3 f_1^3, \dots, \tag{5.12}$$

where $c_2, c_3, \dots$ are as given in (4.2). In each case we seek solutions from $H^{0,2m}_{\rho}(\mathbb{R})$. In view of asymptotic properties for linearized operators in section 2, the solutions are assumed to grow slowly (at worst polynomially) as y increases and will ultimately be matched to solutions of the ODEs (5.2) and (5.3) that have the correct behavior at infinity, (5.4).

As observed above, it follows from (4.16) that the lowest-order equation (5.10) can be solved in terms of a rescaling of the null eigenfunction ψ_{2l} of $\boldsymbol{\mathcal{L}}_{2l}$. Applying in (3.12) the scaling $y \mapsto (2/(2r+1))^{1/4} y$, it follows that there is a constant α such that

$$f_1(y) = \alpha \tilde{f}_1(y), \quad \text{where } \tilde{f}_1(y) = \sum_{j=0}^{r} \left(\frac{2r+1}{2}\right)^{j-r-1/2} \frac{1}{j!} D^{4j} y^{4r+2}. \tag{5.13}$$

For example, $f_1(y) = \alpha y^2$ when $r = 0$ and $\mu = \frac{1}{2}$. Here the constant α is unspecified at this level of expansion and will be determined by a solvability condition for the higher-order terms.

The Fredholm alternative gives the orthogonality condition for the second equation (5.11) at order ε to have a solution in $H^{2m}_{0,\rho}(\mathbb{R})$,

$$\langle f_1^2, \psi^*_{2l} \rangle = 0, \tag{5.14}$$

where $\psi^*_{2l} = \psi^*_{2l}((2/l)^{1/4} y)$ defined in (3.9) is the eigenfunction of the adjoint operator $\boldsymbol{\mathcal{L}}^*_{1/2l}$. If $r = 0$ and $l = 1$, then the first three even eigenfunctions of $\boldsymbol{\mathcal{L}}_{1/2}$ are given in (3.14). Since ψ_2^* is the null eigenfunction of $\boldsymbol{\mathcal{L}}^*_{1/2l}$, it follows that $\langle \psi_2^*, \psi_0 \rangle = 0$ and $\langle \psi_2^*, \psi_4 \rangle = 0$. Hence $\langle \psi_2^*, y^4 \rangle = \langle \psi_1, f_1^2 \rangle = 0$, so that the orthogonality (5.14) holds and there exist solutions of (5.11) at this order. This is the lack of resonance condition, which we described earlier.

For arbitrary r, by (5.13),

$$f_1^2(y) = \alpha^2 \sum_{j=0}^{2r} a_j y^{4j+4},$$

and we find a particular polynomial solution of (5.11) in the form

$$\alpha^2 \tilde{f}_2(y) = -c_2 \alpha^2 \sum_{j=-1}^{2r} b_j y^{4j+4}. \tag{5.15}$$

Substituting it into the equation and equating the coefficients gives

$$b_{2r} = -a_{2r}, \quad b_{-1} = 4!\, b_0 \quad \text{and} \tag{5.16}$$

$$b_j = \frac{2r+1}{2(r-j)-1} \left[a_j + b_{j+1} \frac{(8+4j)!}{(4+4j)!} \right] \quad \text{for } j = 2r-1, \dots, 0. \tag{5.17}$$

Hence, the orthogonality condition (5.14) holds. The general solution of (5.11) is given by

$$f_2(y) = \alpha^2 \tilde{f}_2(y) + \alpha_1 \tilde{f}_1(y), \tag{5.18}$$

where α_1 is an extra real unknown.

The unknowns α and α_1 are determined by applying the Fredholm alternative at the next orders of expansion. In (5.12), similar to (5.14), the solvability condition is given by

$$\left\langle \sigma_l y f_1' - 2c_2 f_1 f_2 - c_3 f_1^3, \psi_{2l}^* \right\rangle = 0. \tag{5.19}$$

Substituting (5.13) and (5.18) yields the algebraic equation

$$\alpha A - \alpha^3 B + \alpha\alpha_1 C = 0, \tag{5.20}$$

where $A = \langle \sigma_l y f_1', \psi_{2l}^* \rangle$, $B = \langle c_3 \tilde{f}_1^3 + 2c_2 \tilde{f}_1 \tilde{f}_2, \psi_{2l}^* \rangle$, and the third coefficient C vanishes by the first solvability criterion (5.14),

$$C = -2C_2 \langle \tilde{f}_1^2, \psi_{2l}^* \rangle = 0. \tag{5.21}$$

Equation (5.20) is a cubic equation for the first unknown α only, $\alpha(\alpha^2 - \sigma_l \gamma) = 0$, where γ can be computed explicitly. The $\alpha = 0$ case simply corresponds to the constant solution (the trivial expansion (5.9)) and can be discarded. Hence, we have two solutions

$$\alpha = \pm\sqrt{\sigma_l \gamma}\,. \tag{5.22}$$

The sign of σ_l is thus the same as that of γ, while the sign of α follows from matching to the far field solution (see section 5.2). In general, the second unknown α_1 (together with an extra one α_3 obtained from the homogeneous equation (5.12), etc.) is to be determined from the solvability conditions of equations for the coefficients f_4, $f_5, \ldots$ of higher-order perturbations. Although not presented, higher approximations follow in a similar manner to those here.

Example 1. To illustrate this calculation, we now look at the two cases of $l = 1$ and $l = 3$ for the quadratic nonlinearity with $p = 2$, where $G_p(f) = |f|f - f$. These are chosen so that the corresponding bifurcation points at $\mu = \frac{1}{2}$ and $\mu = \frac{1}{6}$ are on either side of the "self-similar" value of $\mu_2 = \frac{1}{4}$, as indicated in Figure 1.

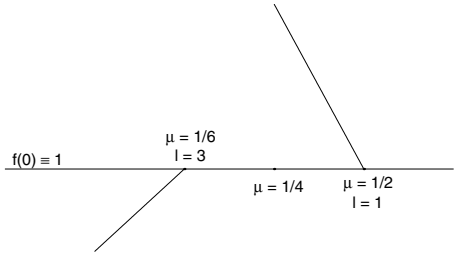


FIG. 1. *Sketch of the bifurcation points under consideration.*

The first bifurcation point: $\mu_1 = \frac{1}{2}$ ($l = 1$, $r = 0$). As observed above, when $l = 1$, we have $f_0 = 1$ and $f_1 = \alpha y^2$. A simple calculation then gives $f_2 = \alpha^2(y^4 + 24)$, and expansion (5.9) takes the form

$$f(y) = 1 + \alpha\varepsilon^{1/2} y^2 + \varepsilon\left[\alpha^2(y^4 + 24) + \alpha_1 y^2\right] + \varepsilon^{3/2} f_3(y) + \cdots. \tag{5.23}$$

Observe that since $c_3 = 0$, the solvability condition (5.19) for f_3 is then given by

$$\langle 2\sigma_1\alpha y^2 - 2\alpha^3 y^2(y^4+24), \psi_2^*\rangle = 0, \quad \psi_2^* = \psi_2^*(2^{1/4}y). \tag{5.24}$$

To calculate α, we exploit the fact that $\hat{\psi}_2^*(\omega) = -\omega^2 e^{-\omega^4}/\sqrt{2}$ by (3.6) and (3.9). Recall also that if a function $f(y)$ has Fourier transform $\hat{f}(\omega)$, then

$$\langle f, y^{2n}\rangle = (-1)^n \hat{f}^{(2n)}(0). \tag{5.25}$$

Taking $\psi_2^* = \psi_2^*(2^{1/4}y)$ yields $\langle \psi_2^*(2^{1/4}y), y^2\rangle = 1/2^{1/4}$ and $\langle \psi_2^*(2^{1/4}y), y^6\rangle = -180/2^{1/4}$, the solvability condition (5.24) reduces to the cubic equation $\sigma_1\alpha + 156\alpha^3 = 0$, and hence

$$\sigma_1 = -1 \quad \text{and} \quad \alpha = \pm\frac{1}{\sqrt{156}} = \pm\frac{\sqrt{39}}{78}, \tag{5.26}$$

so that (5.23) yields

$$f(y) = 1 \pm \varepsilon^{1/2}\frac{y^2}{\sqrt{156}} + \varepsilon\left(\frac{1}{156}(y^4+24) + \alpha_1 y^2\right) + \cdots. \tag{5.27}$$

The sign of α will be determined by matching to the solution in the midrange. We show presently that $\alpha < 0$ so that

$$f(y) = 1 - \varepsilon^{1/2}\frac{y^2}{\sqrt{156}} + \varepsilon\left(\frac{1}{156}(y^4+24) + \alpha_1 y^2\right) + \cdots,$$

and, in particular, since $\varepsilon > 0$,

$$f(0) = 1 + \frac{2}{13}\,\varepsilon + \cdots > 1. \tag{5.28}$$

The resulting branch thus bifurcates to the left and exists locally only for $\mu < \frac{1}{2}$; there is no possible matching to a decaying solution for $\mu > \frac{1}{2}$. The numerical calculations reported in the next section indicate that the branch persists globally, so that solutions exist at the self-similar value $\mu_2 = \frac{1}{4}$.

The third bifurcation point: $\mu_l = \frac{1}{6}$ ($l = 3$, $r = 1$). We again have $f_0 = 1$, and now $f_1(y) = \alpha(y^6 + 540y^2)$ and $f_2 = \alpha^2(y^{12} - 32400y^8 - 164170800y^4 - 3940099200)$, so that the expansion is

$$\begin{aligned} f = {} & 1 + \varepsilon^{1/2}\alpha\left(y^6 + 540y^2\right) \\ & + \varepsilon[\alpha^2\left(y^{12} - 32400y^8 - 164170800y^4 - 3940099200\right) + \alpha_1\tilde{f}_1] + \cdots. \end{aligned}$$

A similar (but much longer) analysis of the orthogonality condition (5.19) with eigenfunction $\psi_6((2/3)^{1/4}y) = (y^6+540y^2)/12\sqrt{5}$ then indicates that the branch again bifurcates to the left and exists locally for $\mu < \frac{1}{6}$.

B. *The case of $m = 2$ and l even.* In the case $l = 2r$ the bifurcation occurs at the point $\mu_{2r} = \frac{1}{4r}$. Because of the presence of a constant term in the eigenfunction ψ_{2l}, the effect of the "forcing terms" yf' comes in at lower order than in the previous case. This leads to a standard asymptotic expansion for $f(y)$ of the form (cf. Theorem 12.1 in [44])

$$f = f_0 + \varepsilon f_1 + \varepsilon^2 f_2 + \cdots. \tag{5.29}$$

Substituting this expression for f into (5.5) or (5.6) gives

$$O(\varepsilon): \qquad \boldsymbol{\mathcal{L}}_{1/2l} f_1 \equiv -f_1'''' - \frac{1}{4r}\, y f_1' + f_1 = 0, \tag{5.30}$$

$$O(\varepsilon^2): \qquad \boldsymbol{\mathcal{L}}_{1/2l} f_2 = \sigma_l y f_1' - c_2 f_1^2. \tag{5.31}$$

As before, we express f_1 as a multiple of the (scaled) eigenfunction $\psi_{2l}(r^{-1/4}y)$,

$$f_1(y) = \alpha \tilde{f}_1(y) \equiv \alpha \sum_{j=0}^{r} \frac{r^{j-r}}{j!} D^{4j} y^{4r}. \tag{5.32}$$

The value of α is determined by considering the solvability condition for (5.31) at $O(\varepsilon^2)$. From the analysis above, it follows that, for f_2 to exist, we must have

$$\langle \sigma_l y f_1' - c_2 f_1^2, \psi_{2l}^* \rangle = 0 \quad \text{with } \psi_{2l}^* = \psi_{2l}^* \left(\left(\frac{2}{l} \right)^{1/4} y \right). \tag{5.33}$$

This leads to a quadratic equation in α of the form $\alpha(\alpha - \gamma) = 0$, where γ may again be determined explicitly. This is the case of a unique nontrivial solution existing for both $\mu > \mu_l$ and $\mu < \mu_l$, and again we will need an extra matching argument to determine the correct sub-branch.

To illustrate this calculation, we again take $p = 2$, $G_p(f) = |f|f - f$ and now consider the case of $l = 2$. This is an especially important value as it corresponds to $\mu_2 = \frac{1}{4}$, at which the self-similar solution exists. In this case we have $f_0 = 1$ and $f_1 = \alpha(y^4 + 24)$. The solvability condition for α is now

$$\langle \sigma_2 y f_1' - f_1^2, \psi_4^*(y) \rangle = \langle 4\sigma_2 \alpha y^4 - \alpha^2 (y^4 + 24)^2, \psi_4^*(y) \rangle = 0.$$

We have that $\hat{\psi}_4^*(\omega) = \omega^4 e^{-\omega^4}/2\sqrt{6}$, and it follows that the quadratic equation satisfied by α is given by

$$96\sigma_2 \alpha + 39168\alpha^2 = 0 \quad \Longrightarrow \quad \alpha = -\frac{1}{408}\, \sigma_2. \tag{5.34}$$

We show presently that, to match with the midrange, we have to have $\alpha < 0$ so that $\sigma_2 = 1$. Hence

$$f(y) = 1 - \frac{\varepsilon}{408} \left(y^4 + 24 \right) + \varepsilon^2 \left[\tilde{f}_2(y) + \alpha_1 \tilde{f}_1(y) \right] + \cdots, \tag{5.35}$$

where the third term (actually we do not need to compute it) explains the spatial nonmonotonicity of such a solution. If $\varepsilon > 0$, then

$$f(0) = 1 - \frac{1}{17}\, \varepsilon + O(\varepsilon^2) < 1. \tag{5.36}$$

5.1.2. The midrange $\boldsymbol{\varepsilon^\gamma < y < e^{1/\varepsilon}}$. The midrange behavior is governed by the solutions of a first-order equation, which is different for each nonlinearity. However, the calculation now takes the same form for both l even and odd and uses a regular asymptotic expansion. To study the midrange, we rescale the underlying ODEs in space according to the transformation

$$s = \varepsilon^\gamma y \geq 0 \qquad (\gamma \text{ as in (5.7)}). \tag{5.37}$$

The outer limit of the inner region can be matched to the midrange region by taking s to be small and y to be large.

A. *The case* $G_p(f) = |f|^{p-1}f - \frac{1}{p-1}\,f$. Under the spatial rescaling (5.37), equation (5.5) becomes

$$-\varepsilon^{4\gamma} f'''' - (\mu_l + \sigma_l \varepsilon)\, s f' + |f|^{p-1} f - \frac{1}{p-1}\, f = 0, \quad l = 1, 2, \dots , \tag{5.38}$$

where $' = d/ds$. To solve this, we pose a regular asymptotic expansion

$$f = f_0 + \varepsilon^{4\gamma} f_1 + \varepsilon^{8\gamma} f_2 + \cdots . \tag{5.39}$$

To leading order we have simply the first-order ODE $-\frac{1}{2l} y f_0' + |f_0|^{p-1} f_0 - \frac{1}{p-1} f_0 = 0$, which has a family of bounded positive exact solutions

$$f_0(s) = [(p-1) + \kappa s^{2l}]^{-1/(p-1)}, \tag{5.40}$$

where $\kappa > 0$ is a positive constant.

Note that, for *small* s, we have

$$f_0(s) = \beta^\beta \left(1 - \frac{\kappa}{(p-1)^2} s^{2l} + \frac{p\kappa^2}{2(p-1)^4} s^{4l} + O(s^{8l})\right), \tag{5.41}$$

while for *large* s,

$$f_0(s) = \kappa^{-1/(p-1)} s^{-2l/(p-1)} + \cdots . \tag{5.42}$$

We now consider the next term in the asymptotic expansion, looking at the two cases of small s and large s separately. The function f_1 satisfies the equation

$$-\frac{1}{2l}\, s f_1' + \left[\frac{p}{(p-1) + \kappa s^{2l}} - \frac{1}{(p-1)}\right] f_1 = \sigma_l s f_0' + f_0'''' .$$

We consider for simplicity the case of $p = 2$, and look at the three cases of $l = 1, 2$, and 3.

If $l = 1$, then $4\gamma = 1$, and for small s we have $f_0(s) = 1 - \kappa s^2 + \frac{1}{2}\kappa^2 s^4 + \cdots$; thus the leading order contribution to $\sigma_l s f_0' + f_0''''$ is simply $12\kappa^2$, and hence we have, to leading order as $s \to 0$,

$$f_1(s) = 12\kappa^2 + \cdots .$$

If $l = 2$, then $4\gamma = 1$, and for small s, $f_0(s) = 1 - \kappa s^4 + \frac{1}{2}\kappa^2 s^8 + \cdots$ so that, to leading order,

$$f_1(s) = -24\kappa + \cdots .$$

If $l = 3$, then $4\gamma = 1/3$, and for small s, $f_0(s) = 1 - \kappa s^6 + \frac{1}{2}\kappa^2 s^{12} + \cdots$ so that, to leading order, $f_0'''' = -360\kappa s^2$ and

$$f_1(s) = -540\kappa s^2 + \cdots .$$

We conclude that the small s limit of the midrange solution is

$$f = 1 - \kappa s^2 + \frac{1}{2}\,\kappa^2 s^4 + \cdots + \varepsilon(12\kappa^2 + \cdots) \quad \text{if} \quad l = 1, \tag{5.43}$$

$$f = 1 - \kappa s^4 + \frac{1}{2}\,\kappa^2 s^8 + \cdots - \varepsilon(24\kappa + \cdots) \quad \text{if} \quad l = 2, \tag{5.44}$$

$$f = 1 - \kappa s^6 + \frac{1}{2}\,\kappa^2 s^{12} + \cdots - \varepsilon(540\kappa^2 s^2 + \cdots) \quad \text{if} \quad l = 3. \tag{5.45}$$

In terms of the original variable y we have

$$f = 1 - \varepsilon^{1/2}\kappa y^2 + \varepsilon\kappa^2\left(\frac{1}{2}\,y^4 + 12\right) + \cdots \quad \text{if} \quad l = 1, \tag{5.46}$$

$$f = 1 - \varepsilon\kappa(y^4 + 24) + \frac{1}{2}\,\varepsilon^2\kappa^2 y^8 + \cdots \quad \text{if} \quad l = 2, \tag{5.47}$$

$$f = 1 - \varepsilon^{1/2}\kappa(y^6 + 540y^2) + \frac{1}{2}\,\varepsilon\kappa^2 y^{12} + \cdots \quad \text{if} \quad l = 3. \tag{5.48}$$

We can now consider matching the above expressions to the expansions given in the last sections.

If $l = 1$, then comparing with (5.27), we have a perfect match, provided that $\kappa = -\alpha$. As $\kappa > 0$, it follows that $\alpha = -1/\sqrt{156}$. Thus in the midrange when $l = 1$ we have

$$f_0(y) = \left(1 + \frac{\varepsilon^{1/2}y^2}{\sqrt{156}}\right)^{-1}.$$

As remarked earlier, this bifurcation branch exists only if $\mu < \frac{1}{2}$.

If $l = 2$, then comparing with (5.35), we again have a perfect match if $\kappa = -\alpha > 0$. In the midrange when $l = 2$ there holds

$$f_0(y) = \left(1 + \frac{1}{408}\,\varepsilon^{1/2}y^4\right)^{-1}.$$

Note that this expression is only meaningful if $\epsilon > 0$. As in this case $\sigma_2 = 1$, it follows that locally the branch of solutions that bifurcates from $\mu = \frac{1}{4}$ exists only if $\mu > \frac{1}{4}$. Numerically we observe that this curve continues globally for values of $\mu < \frac{1}{4}$, and hence there is a fold bifurcation at some point $\mu = \mu_* > \frac{1}{4}$, with a *nonzero* solution on the branch existing at $\mu = \frac{1}{4}$. This corresponds to a self-similar solution distinct from that lying on the branch bifurcating from the point $\mu = \frac{1}{2}$. The existence of such a solution is consistent with the semistability of the center manifold determined in section 4.

If $l = 3$, then comparing with the inner expansion, we again have a match if $\kappa = -\alpha > 0$, and in the midrange

$$f_0(y) = (1 - \alpha\varepsilon^{1/2}y^6)^{-1} \quad (\alpha < 0).$$

Now consider the behavior for $s \gg 1$ when $p = 2$. For these values of s, to leading order, the function f_1 satisfies the ODE $-\mu_l s f_1' - f_1 = -2l\sigma_l/\kappa s^{2l} + \cdots$; hence,

$$f_1(s) = \frac{4l^2\sigma_l \ln s}{\kappa s^{2l}} + \cdots \quad \text{as } s \to \infty\,.$$

Or, returning to the original variable y,

$$f(y) = \frac{1}{\kappa\varepsilon^{l/2}y^{2l}}\left(1 + 4l^2\sigma_l \ln y + \cdots\right) \quad \text{as} \quad y \to \infty. \tag{5.49}$$

B. *The case of* $G = e^f - 1$. Under the same spatial rescaling as before, (5.6) becomes

$$-\varepsilon f'''' - (\mu_l + \sigma_l\varepsilon)\, s f' + e^f - 1 = 0, \quad l = 1, 2, \ldots\,.$$

Posing expansion (5.39), substituting into the ODE, and solving the leading order equation gives

$$f_0(s) = -\ln(1+\kappa s^{2l}). \tag{5.50}$$

The analysis now proceeds as above, and again matching in the limit $s \to 0$ fixes $\kappa > 0$.

5.1.3. Far field behavior. The correct far field behavior is determined by assuming slow growth in both (5.2) and (5.3), $f''''(y) \to 0$ as $y \to \infty$, and hence $|f|f \ll f$ for small $f > 0$ there in (5.5), while $e^f \ll 1$ for $f \ll -1$ in (5.6). In the case of (5.5), this gives

$$f = Cy^{-1/\mu}(1+o(1)) \equiv Cy^{-2l/(1+2l\sigma_l\varepsilon)}(1+o(1)) \quad \text{as } y \to \infty.$$

Expanding this for $\varepsilon \ll 1$, we have

$$f = Cy^{-2l}(1+4l^2\sigma_l\varepsilon \ln y) + \cdots \quad (\varepsilon|\ln y| \ll 1).$$

This matches with (5.49) if $C = 1/\kappa\varepsilon^{l/2}$ (note that $\kappa = |\alpha|/\varepsilon$ for $l = 3$).

5.2. Bifurcations from $\mu_m = \frac{1}{2m}$ for general m. As remarked, for $m = 2$ we can also postulate existence of the new profile f_2 from the shape of the branch associated with $\mu_2 = \frac{1}{4}$, as the branch leaves the bifurcation point to the right and then is expected to fold back. In fact, this behavior can be understood for general m.

For all m, the bifurcation point $\mu_m = \frac{1}{2m}$ is associated with a zero eigenvalue of the linearized operator $\boldsymbol{\mathcal{L}} + I$ in the PDE (4.1). Further evidence for the existence of a nonlinear pattern associated with this point comes from the local structure of the bifurcation diagram. Looking for small solutions near this point, we solve

$$(-1)^{m+1}D_y^{2m}f - \mu_m y D_y f + f + \sigma_m \varepsilon y D_y f + \bar{G}(f - f_0) = 0,$$

where $\bar{G}$ is the quadratic perturbation (4.2). At $\mu_m = \frac{1}{2m}$ we have the regular expansion (5.29), and expanding as before gives

$$\mathcal{L}_{1/2m} f_1 = 0 \quad \Longrightarrow \quad f_1 = \alpha\left[y^{2m} + (-1)^m(2m)!\right] \quad \text{with unknown } \alpha \in \mathbb{R}.$$

At the next order we have

$$\mathcal{L}_{1/2m} f_2 = \sigma_m y f_1' - c_2 f_1^2 = 2m\sigma_m \alpha y^{2m} - c_2\alpha^2\left([(2m)!]^2 + 2(-1)^m(2m)!y^{2m} + y^{4m}\right). \tag{5.51}$$

By the Fredholm alternative this can be solved only if

$$\left\langle 2m\sigma_m\alpha y^{2m} - c_2\alpha^2\left([(2m)!]^2 + 2(-1)^m(2m)!y^{2m} + y^{4m}\right), \psi_{2m}^*(y)\right\rangle = 0.$$

By (3.9), $\hat{\psi}_{2m}^*(\omega) = \omega^{2m}e^{-2m}/\sqrt{(2m)!}$ so that, after a little manipulation noting that

$$\langle 1, \psi_{2m}^*\rangle = 0, \quad \langle y^{2m}, \psi_{2m}^*\rangle = (2m)!, \quad \langle y^{4m}, \psi_{2m}^*\rangle = (-1)^{m+1}(2m)!,$$

the solvability condition becomes

$$2m\sigma_m\alpha(2m)! - c_2\alpha^2(-1)^{m+1}\left((4m)! - 2[(2m)!]^2\right) = 0.$$

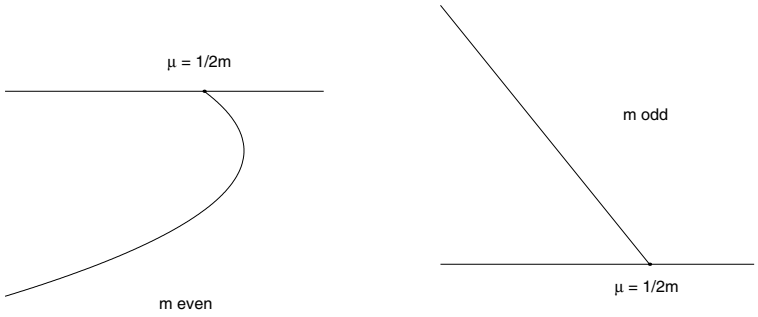


FIG. 2. *Schematic of the distinction between even and odd m.*

Thus the solvability condition implies that

$$\alpha = (-1)^{m+1}\frac{\sigma_m}{c_2}\frac{2m(2m)!}{(4m)! - 2[(2m)!]^2} \tag{5.52}$$

and hence

$$f(y) = f_0 + (-1)^{m+1}\varepsilon\frac{\sigma_m}{c_2}\frac{2m(2m)!}{(4m)! - 2[(2m)!]^2}\left(y^{2m} + (-1)^m(2m)!\right) + \cdots .$$

However, to match with the midrange, we require that $f_1 \to -\infty$ as $y \to \infty$, i.e., $\alpha < 0$ in (5.52), which sets

$$\sigma_m = (-1)^m \quad \text{for} \quad \varepsilon > 0.$$

Hence, by (5.1) for even m, the branches initially increase in μ and thus, if they have folded back, contribute an extra similarity profile $f_m(y)$ at $\mu = \frac{1}{2m}$, whereas there need be no such contribution for odd m; see Figure 2.

The existence of a *second* self-similar solution to the ODE in the case $m = 2$ is suggested by the center manifold analysis in Proposition 4.1. More precisely, consider the *unstable* center manifold behavior (4.6) for any even m,

$$g(y,\tau) = -\frac{1}{\gamma_0\tau}\psi_{2m}(y) + \cdots \to 0 \quad \text{as } \tau \to -\infty, \tag{5.53}$$

where $\psi_{2m} > 0$ is given by (4.5). We suppose that $g(\cdot,\tau)$ becomes sufficiently large as $\tau \approx -0$. Hence, $g(y,\tau) < 0$ for $\tau \ll -1$ on any compact subset in y, i.e., the corresponding solution of the PDE (2.5) (or (2.7)) satisfies $\theta(y,\tau) = \beta^\beta + g(y,\tau) < \beta^\beta$. Such a solution can be extended as above to satisfy $\theta(y,\tau) \to 0$ as $y \to \infty$; see also [20]. This shows that such an orbit cannot be a heteroclinic connection $\beta^\beta \to 0$, since for $\tau \approx -0$ this would mean that $|\theta(y,\tau)|$ gets essentially smaller than the constant blow-up profile β^β. Hence $\theta(y,\tau)$ cannot correspond to a solution $u(x,t)$ of the PDE that blows up at the fixed $t = T$; see L^∞-estimates of the blow-up rate in [10] and [21]. Therefore, this $\theta(\cdot,\tau)$ can be assumed to describe an orbital connection $\beta^\beta \to f_m(y)$ to a new nontrivial similarity profile f_m existing at $\mu = \mu_m$. Note that, by construction, it is expected that a certain approximated order occurs, meaning that $f_m(y) \lesssim \beta^\beta$ in $\mathbb{R}$ in a natural sense.

On the other hand, for odd m's, $\psi_{2m}(y)$ in (4.5) changes sign and we do not have such a contradiction. (One can see that an orbital connection $\beta^\beta \to 0$ is possible; see such a center manifold pattern in [25].)

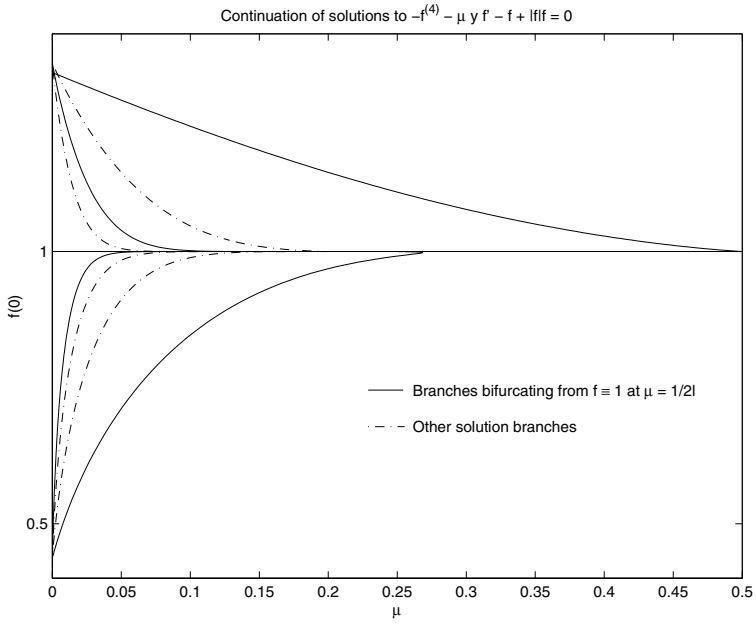


FIG. 3. *Bifurcation diagram for* $m = 2$.

6. Numerical calculations of the self-similar profiles. We next present a numerical calculation of the solutions of the problem (4.12) parameterized by μ and taking $G_p(f) = |f|f - f$ for $p = 2$. (As indicated from the analysis of the previous sections, the case $G_e(f) = e^f - 1$ is fundamentally the same and is omitted for the sake of brevity.) This calculation allows us to extend the asymptotic analysis of the previous section, and, in particular, to study the global behavior of the branches that bifurcate from the first two bifurcation points at $\mu_1 = \frac{1}{2}$ and $\mu_2 = \frac{1}{4}$. The solutions were obtained using a collocation code that guarantees a small residual tolerance [41]. The initial points on each curve were obtained by setting $\mu = 0$. The continuation of each solution was then done by using the pseudo arc-length routine in AUTO [13]. Symmetry conditions were imposed at the origin, and minimal growth was enforced at the far field by solving the problem on the finite interval $(0, 1000)$ and setting the highest derivatives to zero at the right-hand boundary.

6.1. The fourth-order case $m = 2$. In Figure 3 we present the results of the numerical calculations for different values of the parameter μ, looking at the fourth-order differential equations given by taking $m = 2$. In this figure we use $f(0)$ as a measure of the size of the solution. The existence of branches bifurcating from each of the points $\mu_l = \frac{1}{2l}$ (displayed as solid lines) is clear. Also plotted in dashed lines are other solutions obtained from continuing solutions from $\mu = 0$ that do not bifurcate from the constant solution $f \equiv 1$. In this format it is difficult to distinguish the solutions that bifurcate from the linear spectrum from the additional "nonlinear" solutions. To make this distinction clear we plot the same solutions in Figure 4 using the L^2_ρ-norm as the solution measure.

We observe first that the curve bifurcating from $\mu_1 = \frac{1}{2}$ appears to exist for all values of $\mu \in [0, \frac{1}{2}]$ and, in particular, there is a nonconstant solution $f_s(y)$ (the subscript s denotes stable; see section 7) of (4.12) for the value of $\mu_2 = \frac{1}{4}$. This

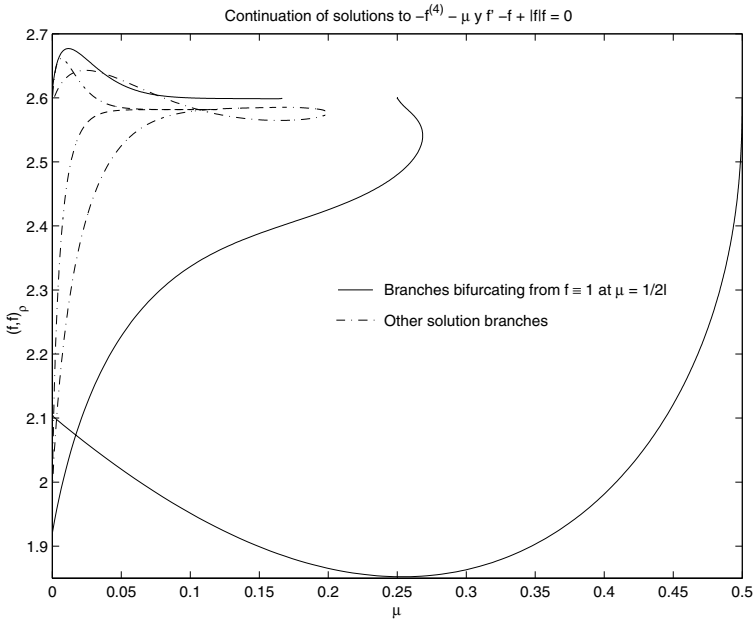


FIG. 4. *Bifurcation diagram for* $m = 2$ *in* L^2_ρ.

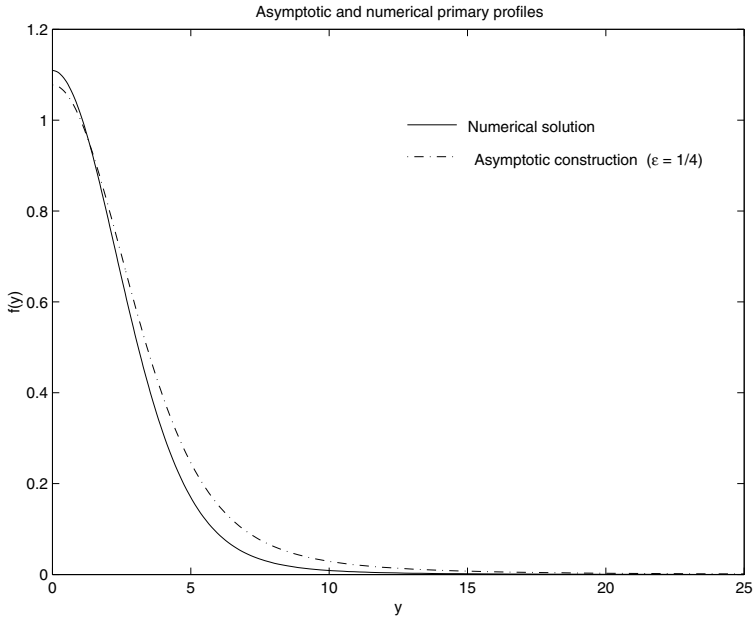


FIG. 5. *Comparison of asymptotic and numeric solutions.*

solution gives a self-similar solution of the underlying PDE (1.7). In Figure 5 we compare the numerical solution to the boundary-value problem (2.9), (2.10) with the asymptotic construction (5.27). In Figure 6 we present an enlargement of Figure 3

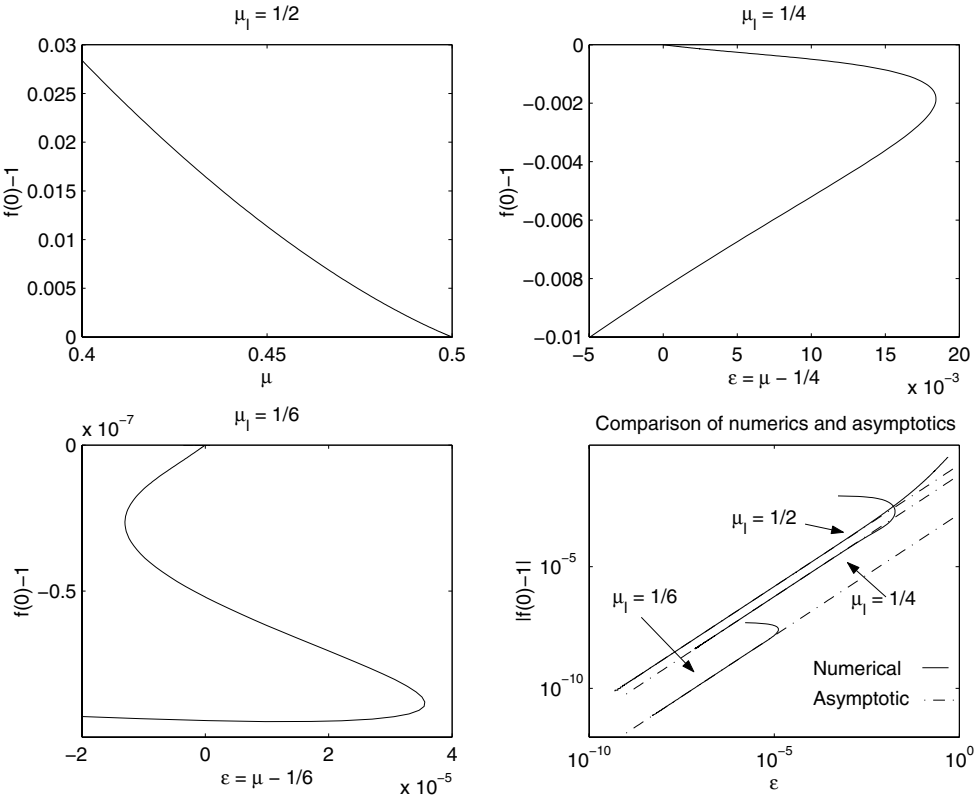


FIG. 6. *Detail of branches at* $\mu = \frac{1}{2}, \frac{1}{4}, \frac{1}{6}$ *for* $m = 2$ *and* $p = 2$.

close to the point $\mu = \frac{1}{2}$, allowing a direct comparison with the asymptotic calculation of $f(0)$ given by (5.28).

In contrast, the curve bifurcating from $\mu = \frac{1}{4}$ appears to exist for all $\mu \in [0, \frac{1}{4} + \delta]$, where δ is a small positive constant. This behavior can be seen more clearly in the enlargement of Figure 3 close to $\mu = \frac{1}{4}$, which is presented in Figure 6. Again, we can compare this figure to the asymptotic calculation of $f(0)$ given by (5.36), and the associated discussion on the unstable center manifold behavior in section 5, which predicts the existence of the bifurcating curve for a range of values of $\varepsilon > \frac{1}{4}$. This asymptotic calculation is clearly valid only for a small range of values of $\mu > \frac{1}{4}$, and the curve of solutions folds back at $\mu \simeq 0.26841\ldots$.

In particular, we observe a second nonzero solution $f_u(y)$ (the subscript u denotes unstable; see section 7) of (4.12) at $\mu = \frac{1}{4}$. The existence of this solution implies the existence of a further self-similar solution of the PDE. As remarked earlier, this result is consistent with the semistability of the center manifold when $m = 2$. The profiles of the two distinct self-similar solutions $f_s(y)$ and $f_u(y)$ are given in Figure 7.

Observe that the form of $f_s(y)$ is qualitatively similar to the profile of the solution computed close to $\mu = \frac{1}{2}$ and described asymptotically in the previous section. In particular, it appears to be a monotone decreasing function of y. In contrast, the self-similar solution $f_u(y)$ is *increasing* for small values of y and decreasing for larger values. This possible small nonmonotonicity in the expansion (5.35) is described by the terms $O(\varepsilon^2)$.

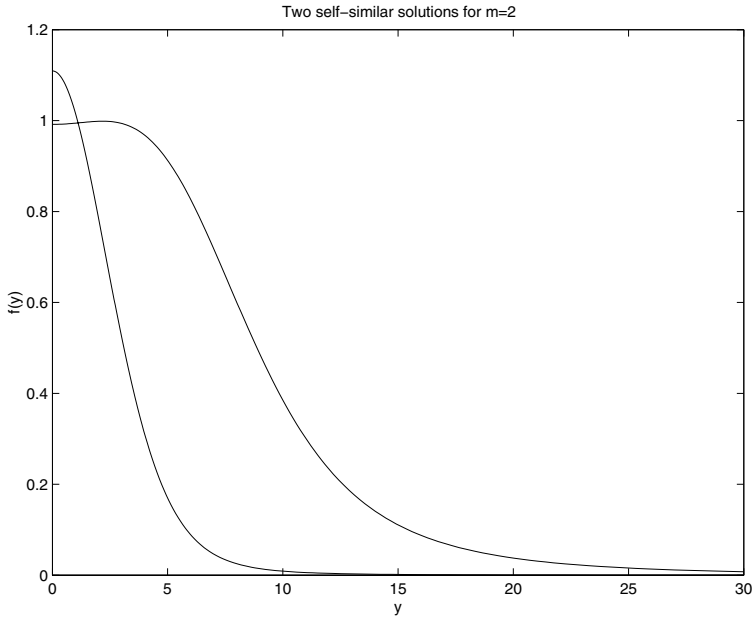


FIG. 7. *The two self-similar profiles f_s, f_u for $m = 2$.*

We also present in Figure 6 a detail of the neighborhood of $\mu = \frac{1}{6}$. Although this branch does not lead to a self-similar solution, its local form is interesting. As predicted by the asymptotic analysis, it bifurcates to the left but then folds back twice locally before continuing backwards to $\mu = 0$. Following these calculations, we make the following conjecture.

CONJECTURE 6.1. *If $m = 2$, then each of the curves bifurcating from the point $\mu_l = \frac{1}{2l}$ continues globally to include the point $\mu = 0$ and has $l - 1$ fold bifurcations in the vicinity of μ_l.*

Such fold bifurcations can occur [34] if

$$0 \in \sigma(\boldsymbol{\mathcal{L}}_\mu + G'(f)I). \tag{6.1}$$

This equation determines a difficult eigenvalue problem for higher-order operators with nonconstant coefficients. The eigenvalues of this problem correspond to the turning points of the solution branches indicated in Figure 6.

Lastly, in Figure 6 we compare our asymptotic construction of the bifurcation diagram with the numerical computations. Away from all folds, the agreement is excellent even with only a linear approximation.

6.2. The sixth-order case $m = 3$. A bifurcation diagram similar to Figure 3, and now for the case of the sixth order differential equations when $m = 3$, is presented in Figure 8. The far-field boundary condition is (5.4).

This picture is qualitatively similar to Figure 3, with the solutions at $\mu_3 = \frac{1}{6}$ being of interest. As before, the monotone decreasing (in a neighborhood of the origin) solution bifurcating from $\mu = \frac{1}{2}$ extends backwards to $\mu_3 = \frac{1}{6}$, as does the solution bifurcating from $\mu = \frac{1}{4}$. This leads to two self-similar solutions f_s and f_u. A detail of Figure 8 in the neighborhood of $\mu = \frac{1}{6}$ is given in Figure 9. As predicted by

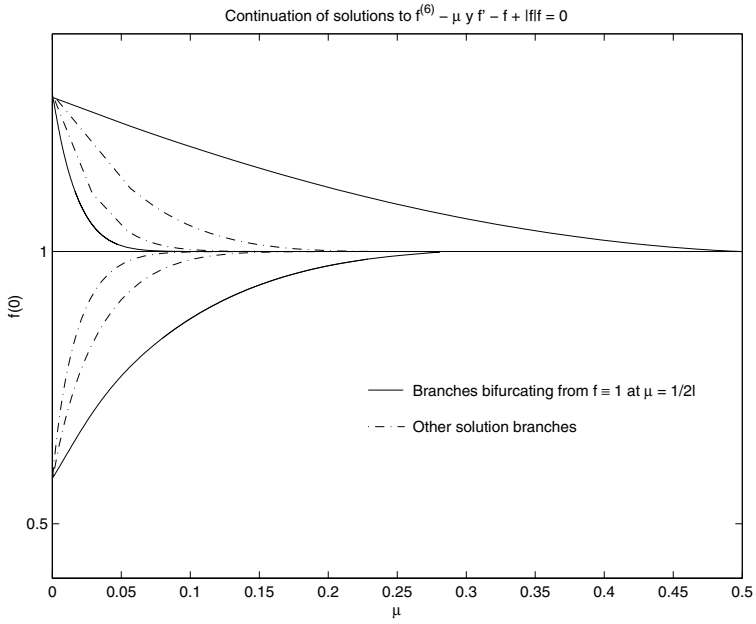


FIG. 8. *Bifurcation diagram for* $m = 3$.

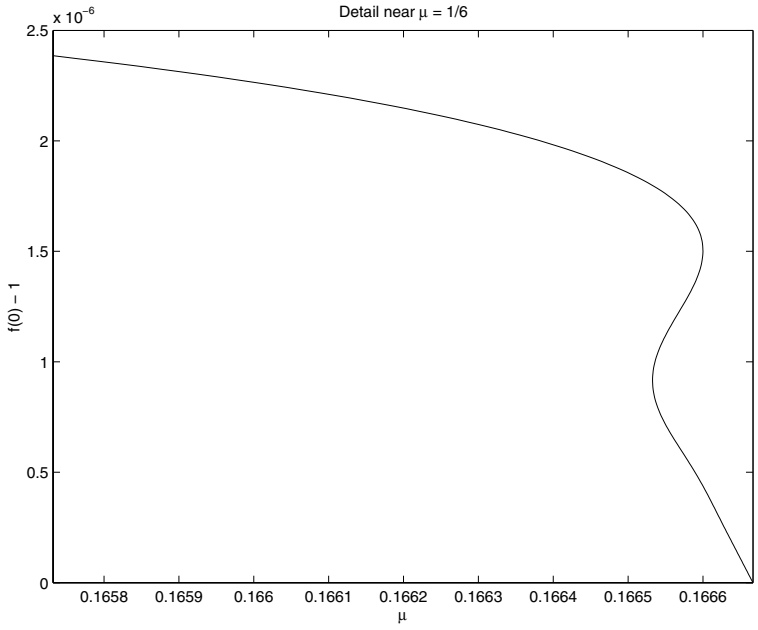


FIG. 9. *Detail of bifurcation diagram for* $m = 3$ *near* $\mu = \frac{1}{6}$.

the asymptotic analysis of section 5, this curve bifurcates *to the left*, and there are no nonzero (and hence no self-similar) solutions on this branch when $\mu = \frac{1}{4}$. Consistent with the previous analysis, we observe two self-similar solutions associated with the unstable subspace and none associated with the center subspace.

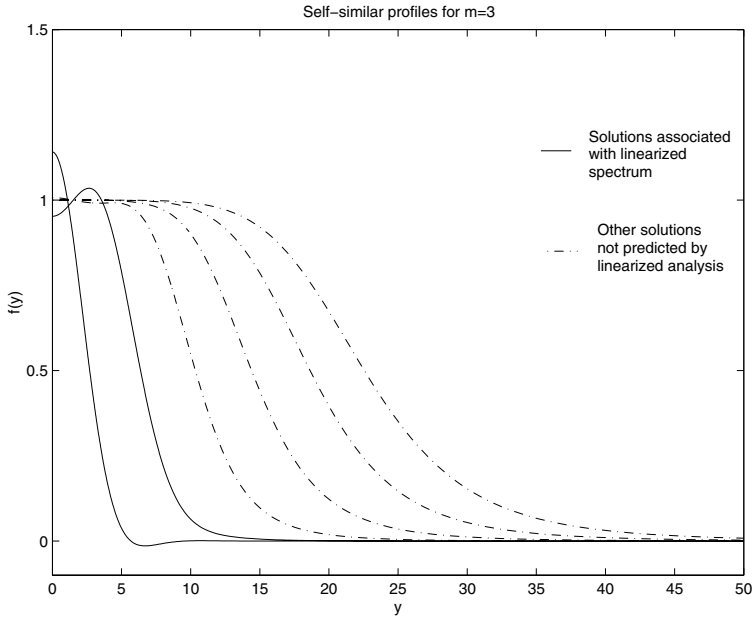


FIG. 10. *Six self-similar profiles for $m = 3$, of which two arise from the analytic bifurcation analysis and four do not.*

A plot of f_s and f_u is given in Figure 10. It is of special interest that, for this value of m, we see *four* other self-similar solutions that arise from paths that start at $\mu = 0$. These are also plotted in Figure 9. Observe that $2 + 4 = 6 = m(m-1)$ for $m = 3$; cf. the last comment in section 4.

7. Numerical simulations of the solutions of the PDE. While the self-similar solutions of (1.7) and (1.6) are important, they give only a partial picture of the overall dynamical behavior of the solutions of these systems. For example, we have not even established whether the self-similar solutions are stable. As we have mentioned, for $m > 1$, the operators in (2.5) and (2.7) are not potentials and do not generate gradient flows as in the second-order case. For $m = 1$, a Lyapunov function exists and this essentially simplifies the asymptotic analysis; see the first results in [22] for $N = 1$ and in [27, 28] for $N \geq 1$. Moreover, compactness of the rescaled orbits $\{\theta(\tau), \tau > \tau_0\}$ remains an open problem (the only known L^∞-estimate for the blow-up rate is a lower one [10, 21]). This makes the asymptotic stability for higher-order equations extremely difficult.

In this section we investigate the dynamics of (1.7) in the case of $m = 2$ by using a *scale-invariant* adaptive numerical method. A general description of the philosophy and implementation of these methods is given in [32, 9, 47] and the references therein. Scale-invariant methods are extremely well suited to computing the solution of systems of PDEs, which have solutions blowing up in finite time and which are also invariant under the action of scaling symmetries. In particular, the underlying PDE is semi-discretized in *space* by using a collocation method on a moving grid. This leads to a system of (stiff) ODEs, which are then solved by using an implicit method. The spatial grid is chosen to *equidistribute* a monitor function $M(u)$ chosen to be

$$M(u) = |u|^{p-1}. \tag{7.1}$$

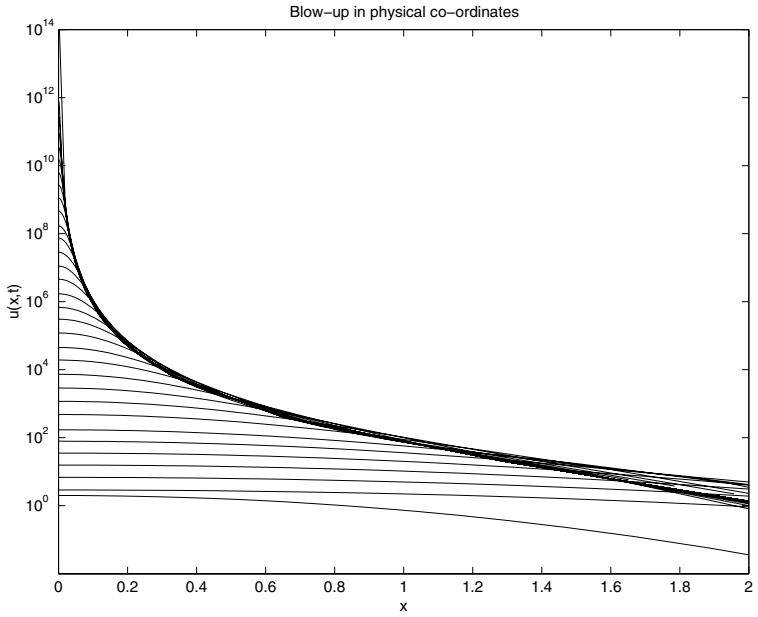


FIG. 11. *The solution of* (1.7) *in the physical variables.*

By doing this, mesh points are clustered where $M(u)$, and hence u is large. The particular choice of $M(u)$ given above leads to a discrete system of equations that is invariant to changes in the scale of the solution and gives relative truncation errors that are *independent of scale*. This is the key to the accuracy of the numerical calculations of this section.

Example 2. For the first calculation we consider the polynomial nonlinearity in (1.7) with as initial data the function

$$u_0(x) = 2e^{-x^2}.$$

First, we present the evolution of this data in the original variables in Figure 11; here the formation of the singularity can be seen clearly. In Figure 12 we present *the same* data, this time in the scaled variables θ and y. Here, the blow-up time T is estimated by a least squares fit of $u(0,t) = f_0/(T-t)^{1/(p-1)}$, with both f_0 and T unknown. The most significant aspect of this figure is that the solutions rapidly converge (exponentially in τ) to the first monotone function $f_s(y)$. The solution of the ODE (2.9) is plotted on Figure 12 for comparison and is indistinguishable from the large τ solutions to the full PDE.

Example 3. For our final calculation, we take as initial data the second solution to (2.9), the solution that extends from the bifurcation point $\mu_2 = \frac{1}{4}$,

$$u(x,0) = f_u(x).$$

This is seen to be unstable. While remaining close to the initial data as the PDE solution increases over several orders of magnitude, eventually the rescaled solution converges to the primary profile as in Example 2; see Figure 13.

Calculations have also been done for the case of the exponential nonlinearity and are fundamentally the same as those presented here; see also [25].

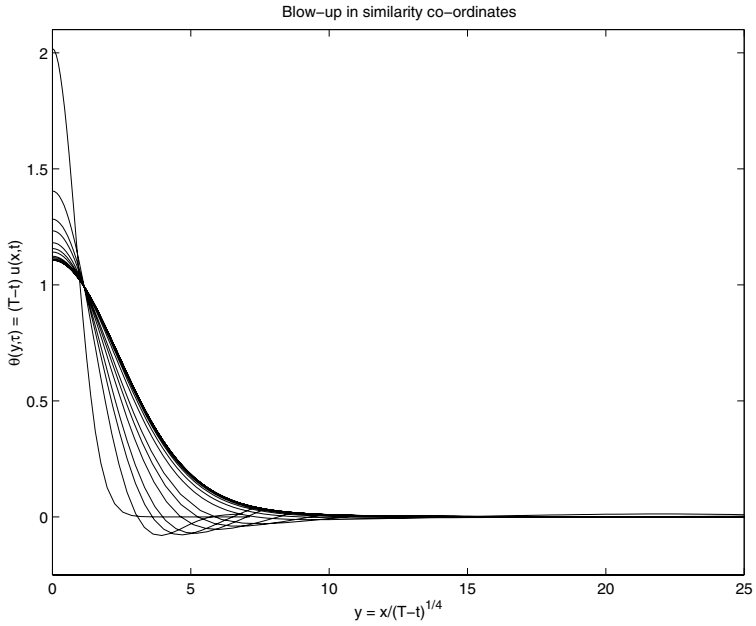


FIG. 12. *The solution of* (1.7) *in the rescaled variables.*

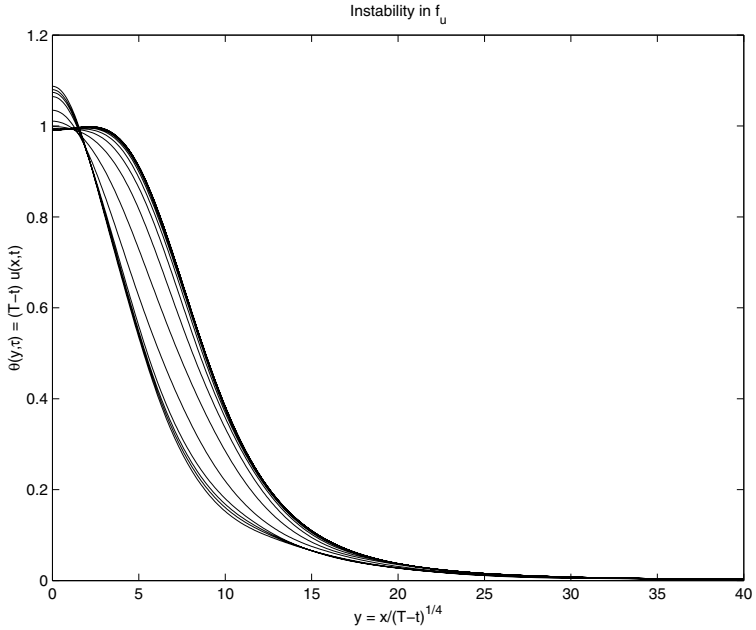


FIG. 13. *The solution of* (1.7) *in the rescaled variables.*

8. Conclusions. It is clear from this study that the (self-similar) behavior of the blow-up solutions of a relatively straightforward higher-order PDE is quite different, and in a sense simpler, than that of related second-order equations. It is very likely that similar behavior will be observed in a much wider class of higher-order equations.

The numerical and asymptotic calculations presented in this paper have suggested a number of open questions in analysis, which deserve further investigation, in particular a fully rigorous proof of the existence of the self-similar solutions and the uniqueness and stability of the "most" monotone stable profiles. We leave this as a subject for future study.

Acknowledgments. The authors wish to thank J.F. Toland for several useful discussions.

REFERENCES

[1] M. M. Ad'jutov and L. A. Lepin, *Absence of blowing up similarity structures in a medium with a source for constant thermal conductivity*, Differential Equations, 20 (1984), pp. 1279–1281.

[2] C. J. Amick and J. F. Toland, *Homoclinic orbits in the dynamic phase space analogy of an elastic strut*, European J. Appl. Math., 3 (1992), pp. 97–114.

[3] S. B. Angenent, *The Morse–Smale property for a semi-linear parabolic equation*, J. Differential Equations, 62 (1986), pp. 427–442.

[4] G. I. Barenblatt, *Scaling, Self-similarity, and Intermediate Asymptotics*, Cambridge University Press, Cambridge, UK, 1996.

[5] J. Bebernes and D. Eberly, *Mathematical Problems from Combustion Theory*, Springer-Verlag, New York, 1989.

[6] M. Ben-Artzi, H. Koch, and J. C. Saut, *Dispersion estimates for fourth order Schrödinger equations*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 87–92.

[7] A. Bressan, *Stable blow-up patterns*, J. Differential Equations, 98 (1992), pp. 57–75.

[8] C. J. Budd and V. A. Galaktionov, *Stability and spectra of blow-up in problems with quasilinear gradient diffusivity*, Proc. Roy. Soc. London A, 454 (1998), pp. 2371–2407.

[9] C. J. Budd, W. Huang, and R. D. Russell, *Moving mesh methods for problems with blow-up*, SIAM J. Sci. Comput., 17 (1996), pp. 305–327.

[10] M. Chaves and V. A. Galaktionov, *Regional blow-up for a higher-order semilinear parabolic equation*, European J. Appl. Math., 12 (2001), pp. 601–623.

[11] M. Chen, X. Y. Chen, and J. K. Hale, *Structural stability for time-periodic one-dimensional parabolic equations*, J. Differential Equations, 96 (1992), pp. 355–418.

[12] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, London, 1955.

[13] E. J. Doedel, *AUTO, A program for the automatic bifurcation analysis of autonomous systems*, Congr. Numer., 30 (1981), pp. 265–384.

[14] Y. V. Egorov, V. A. Galaktionov, V. A. Kondratiev, and S. Pohozaev, *On the necessary conditions of existence to a quasilinear inequality in the half-space*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 93–98.

[15] Y. V. Egorov, V. A. Galaktionov, V. A. Kondratiev, and S. I. Pohozaev, *Asymptotic behavior of global solutions to higher-order semilinear parabolic equations in the supercritical range*, C. R. Acad. Sci. Paris Sér. I Math., 335 (2002), pp. 805–810.

[16] S. D. Eidelman, *Parabolic Systems*, North–Holland, Amsterdam, London, 1969.

[17] G. Fibich, B. Ilan, and G. Papanicolau, *Self-focusing with fourth-order dispersion*, SIAM J. Appl. Math., 62 (2002), pp. 1437–1462.

[18] S. Filippas and R. V. Kohn, *Refined asymptotics for the blow-up of* $u_t - \Delta u = u^p$, Comm. Pure Appl. Math., 45 (1992), pp. 821–869.

[19] D. A. Frank-Kamenetskii, *Towards temperature distributions in a reaction vessel and the stationary theory of thermal explosion*, Dokl. Acad. Nauk SSSR, 18 (1938), pp. 411–412.

[20] V. A. Galaktionov, *On a spectrum of blow-up patterns for a higher-order semilinear parabolic equations*, Proc. Roy. Soc. London A, 457 (2001), pp. 1623–1643.

[21] V. A. Galaktionov and S. I. Pohozaev, *Existence and blow-up for higher-order semilinear parabolic equations: Majorizing order-preserving operators*, Indiana Univ. Math. J., 51 (2002), pp. 1321–1338.

[22] V. A. Galaktionov and S. A. Poshashkov, *Application of new comparison theorems to the investigation of unbounded solutions of nonlinear parabolic equations*, Differential Equations, 22 (1986), pp. 809–815.

[23] V. A. Galaktionov and J. L. Vazquez, *The problem of blow-up in nonlinear parabolic equations*, Discrete Contin. Dynam. Systems, 8 (2002), pp. 399–433.

[24] V. A. Galaktionov and J. L. Vazquez, *A Stability Technique for Evolution Partial Differential Equations. A Dynamical Systems Approach*, Birkhäuser, Boston, Berlin, 2004.
[25] V. A. Galaktionov and J. F. Williams, *Blow-up in a fourth-order semilinear parabolic equation from convection explosion theory*, European J. Appl. Math., 14 (2002), pp. 1–20.
[26] I. M. Gel'fand, *Some problems in the theory of quasilinear equations*, Amer. Math. Soc. Transl. Ser. 2, 29 (1963), pp. 295–381.
[27] Y. Giga and R. Kohn, *Asymptotically self-similar blow-up of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297–319.
[28] Y. Giga and R. Kohn, *Characterizing blowup using similarity variables*, Indiana Univ. Math. J., 36 (1987), pp. 1–40.
[29] D. Henry, *Some infinite-dimensional Morse–Smale systems defined by parabolic partial differential equations*, J. Differential Equations, 59 (1985), pp. 165–205.
[30] M. A. Herrero and J. J. L. Velázquez, *Blow-up behavior of one-dimensional semilinear parabolic equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 131–189.
[31] L. Hocking, K. Stewartson, and J. Stuart, *A nonlinear instability burst in plane parallel flow*, J. Fluid Mech., 51 (1972), pp. 702–735.
[32] W. Huang and R. Russell, *A moving collocation method for solving time dependent partial differential equations*, Appl. Numer. Math., 20 (1996), pp. 101–116.
[33] G. Joulin, A. Mikishev, and G. I. Sivashinsky, *A Semenov–Rayleigh–Benard Problem*, preprint.
[34] M. A. Krasnosel'skii, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, Oxford, Paris, 1964.
[35] H. A. Levine, *The role of critical exponents in blowup theorems*, SIAM Rev., 32 (1990), pp. 262–288.
[36] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel, Berlin, 1995.
[37] F. Merle and H. Zaag, *Stability of the blow-up profile for equations of the type* $u_t = \Delta u + |u|^{p-1}u$, Duke Math. J., 86 (1997), pp. 143–195.
[38] E. Mitidieri and S. I. Pohozaev, *A priori estimates and blow-up of solutions to nonlinear partial differential equations and inequalities*, Proc. Steklov Inst. Math., 234 (2001), pp. 1–362.
[39] L. Peletier and W. Troy, *Spatial Patterns: Higher-order Models in Physics and Mechanics*, Birkhäuser, Boston, Berlin, 2001.
[40] A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailov, *Blow-up in Quasilinear Parabolic Equations*, Walter de Gruyter, Berlin, New York, 1995.
[41] L. F. Shampine and J. Kierzenka, *A BVP solver based on residual control and the MATLAB PSE*, ACM Trans. Math. Software, 27 (2001), pp. 299–316.
[42] C. Sulem and P. L. Sulem, *The Nonlinear Schrödinger Equation*, Appl. Math. Sci. 139, Springer-Verlag, Berlin, 1999.
[43] O. M. Todes, Zh. Fiz. Khim, 4 (1933), p. 71 (in Russian).
[44] M. A. Vainberg and V. A. Trenogin, *Theory of Branching of Solutions of Non-Linear Equations*, Noordhoff, Leiden, The Netherlands, 1974.
[45] J. J. L. Velazquez, *Estimates on* $(N-1)$*-dimensional Hausdorff measure of the blow-up set for a semilinear heat equation*, Indiana Univ. Math. J., 42 (1993), pp. 445–476.
[46] J. J. L. Velázquez, V. A. Galaktionov, and M. A. Herrero, *The space structure near a blow-up point for semilinear heat equations: A formal approach*, Comput. Math. Math. Phys., 31 (1991), pp. 46–55.
[47] J. F. Williams, X. Xu, and R. D. Russell, *MovCol*4: *A Moving Collocation Method for Higher-Order Parabolic Equations*, in preparation.
[48] Y. B. Zel'dovich, G. I. Barenblatt, V. B. Librovich, and G. M. Makhviladze, *The Mathematical Theory of Combustion and Explosions*, Consultants Bureau (Plenum), New York, London, 1985.
[49] Y. B. Zel'dovich and Y. P. Raizer, *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena*, Vols. I and II, Academic Press, New York, 1966.

SIAM J. APPL. MATH.
Vol. 64, No. 5, pp. 1810–1838

TIME REVERSAL FOR DISPERSIVE WAVES IN RANDOM MEDIA*

JEAN-PIERRE FOUQUE†, JOSSELIN GARNIER‡, AND ANDRÉ NACHBIN§

Abstract. Refocusing for time reversed waves propagating in disordered media has recently been observed experimentally and studied mathematically. This surprising effect has many potential applications in domains such as medical imaging, underwater acoustics, and wireless communications. Time refocusing for one-dimensional acoustic waves is now mathematically well understood. In this paper the important case of one-dimensional dispersive waves is addressed. Time reversal is studied in reflection and in transmission. In both cases we derive the self-averaging properties of time reversed refocused pulses. An asymptotic analysis allows us to derive a precise description of the combined effects of randomness and dispersion. In particular, we study an important regime in transmission, where the coherent front wave is destroyed while time reversal of the incoherent transmitted wave still enables refocusing.

Key words. dispersive waves, inhomogeneous media, asymptotic theory, time reversal

AMS subject classifications. 76B15, 35Q99, 60F05

DOI. 10.1137/S0036139903422371

1. Introduction.

Time reversal for ultrasound has been extensively studied by Fink and his collaborators at the "Laboratoire Ondes et Acoustique" in Paris; for a description of these experiments we refer, for instance, to the papers [11, 12]. A time reversal mirror is, roughly speaking, a device which is capable of receiving an acoustic signal in time, keeping it in memory, and sending it back into the medium in the reversed direction of time. Time reversal refocusing properties are well understood mathematically for one-dimensional acoustic waves propagating in random media [9] and for three-dimensional waves in layered media [16] or in the paraxial regime [3, 6, 23, 5, 4].

In this paper we consider a case of dispersive waves, namely the Boussinesq model derived in [20]. We first revisit time reversal for reflected signals generated by a pulse sent in a random half-space. The main property of time reversal is the refocusing of the pulse with a shape that depends only on the statistical properties of the medium, and not on the particular realization. This has been mathematically studied in the high-frequency regime for acoustic waves in [9]. We extend this result to the case of dispersive waves. In Theorem 6.1 we derive the deterministic shape of the refocused pulse, which depends on the statistical properties of the medium and the strength of the dispersion. This result is obtained in the regime of weak fluctuations of the medium, a correlation length of the order of magnitude of the pulse carrier wavelength, and long distances of propagation. The underlying asymptotic analysis is based on the techniques of separation of scales presented, for instance, in [2]. In particular, we generalize the system of transport equations that characterize the multiple scattering of the wave.

*Received by the editors February 5, 2003; accepted for publication (in revised form) December 3, 2003; published electronically July 23, 2004. This work was supported by ONR grant N00014-02-1-0089 while the second and third authors were visiting NC State University.

http://www.siam.org/journals/siap/64-5/42237.html

†Department of Mathematics, North Carolina State University, Raleigh NC 27695-8205 (fouque@math.ncsu.edu).

‡Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex 4, France (garnier@cict.fr). This author was supported by program ACI-NIM-2003-94.

§Instituto de Matemática Pura e Aplicada, Est. D. Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (nachbin@impa.br).

Time reversal refocusing can also be obtained from transmitted waves generated by a pulse propagating through a slab of random medium. For dispersive waves, in contrast with acoustic waves, there is an interesting regime where the coherent front wave is destroyed. We show in this paper that by time reversing the incoherent part of the transmitted wave it is still possible to refocus at the source. We provide a precise analysis of the interplay between randomness and dispersion. In particular, Theorem 7.1 gives the precise description of the refocused pulse.

One potential application discussed in this paper is the characterization by waveform inversion for water waves of the initial sea surface displacement due to tsunamigenic earthquakes [24]. In [24] an adjoint method is proposed. In the synthetic numerical experiments presented there, a shallow water system in two space dimensions is used for the forward propagation, while a linear adjoint method is adopted for the backward identification of the tsunami source. The authors claim that, in principle, the adjoint method can be applied to nonlinear hydrodynamic models. Their method is also applied to real tide gauge series for the small Gorringe Tsunami of 1969, indicating improvements over previous methods. Here we consider a one-dimensional dispersive system, which is valid for longer propagation distances than the hyperbolic shallow water system. Recently we have produced the first analysis for the time reversal of a nonlinear, one-dimensional hyperbolic shallow water system [13]. In particular, we have shown how randomness dramatically improves time reversal experiments. In [13] we have shown that in the presence of randomness one can perform time reversal beyond the shock propagation distance. Randomness acts as an apparent viscosity and regularizes the shock. Extension to linear hyperbolic systems in higher dimensions has been accomplished, for example, in [16]. Hence time reversal for more realistic models in higher dimensions is a promising technique.

Another important fact, regarding applications, is that we have accomplished a mathematical theory for both the time reversal of dispersive waves (the present paper) and also for weakly nonlinear hyperbolic waves [13]. These two papers are important steps in obtaining a mathematical theory for the time reversal of weakly dispersive weakly nonlinear waves, namely solitary waves. This might have a great impact on other models supporting solitons. As a consequence of these two papers, numerical experiments were performed for the time reversal of solitary waves [14].

The paper is organized as follows. In section 2 we introduce the Boussinesq equation including randomness and dispersion, and we describe the different scales arising in the problem. In section 3 we show how the wave can be decomposed into left- and right-propagating modes in the dispersive nonrandom case. This decomposition is crucial in the following sections where the analysis of the random case is performed. In section 4 we establish the system satisfied by the right- and left-going waves in the random case. We also give the integral representation of the transmitted and reflected waves in terms of the mode transmission and reflection coefficients. In section 5 we introduce the time reversal procedures in reflection (TRR) and in transmission (TRT) and derive the corresponding integral representations for the time reversed waves. The two subsequent sections are devoted to the asymptotic analysis of the refocused pulses and comparisons with numerical simulations.

2. The terrain-following Boussinesq model. We consider the Boussinesq equation that describes the evolution of surface waves in shallow channels [20]:

$$M(z)\frac{\partial \eta}{\partial t} + \frac{\partial u}{\partial z} = 0, \tag{2.1}$$

$$\frac{\partial u}{\partial t} + \frac{\partial \eta}{\partial z} - \beta \frac{\partial^3 u}{\partial z^2 \partial t} = 0, \tag{2.2}$$

where η is the wave elevation, u is the depth-averaged velocity, and z and t are the space and time coordinates, respectively. The spatial variations of the coefficient M are imposed by the bottom profile

$$M(z) = 1 + \varepsilon m(z),$$

where 1 stands for the constant mean depth and the dimensionless small parameter ε characterizes the size of the relative fluctuations of the bottom modeled by the zero-mean stationary random process $m(z)$. The process m is assumed to be bounded by a deterministic constant, differentiable, and to have strong mixing properties, such as a rapidly decaying function [22]. We may think, for instance, that $m(z) = f(\nu(z))$, where f is a smooth bounded function and ν is a stationary Gaussian process with Gaussian autocorrelation function and we assume that $\mathbb{E}[f(\nu(0))] = 0$. Note that in that case the realizations of the process ν are of class $\mathcal{C}^\infty$ almost surely. This hypothesis is consistent with the terrain-following coordinate system adopted in deriving (2.1)–(2.2) [20].

We consider the problem on the finite slab $-L \le z \le 0$, where boundary conditions will be imposed at $-L$ and 0 corresponding to a pulse entering the slab from the right at $z = 0$. The quantities of interest, the transmitted and reflected waves, will be observed in time at the extremities $z = -L$ and $z = 0$, respectively.

The coefficient β measures the dispersion strength. In this paper we consider the case where the dispersion parameter β is either of order 1 or small. We consider a pulse whose support is comparable to the correlation length of the random medium, that is, of order 1. In order to see the effect of the small random fluctuations, we consider a long distance of propagation. As we shall see, the interesting regime arises when the propagation distance is of order $1/\varepsilon^2$.

3. The propagating modes of the homogeneous Boussinesq equation. Consider the homogeneous Boussinesq equation (with $m \equiv 0$):

$$\frac{\partial \eta}{\partial t} + \frac{\partial u}{\partial z} = 0, \tag{3.1}$$

$$\frac{\partial u}{\partial t} + \frac{\partial \eta}{\partial z} - \beta \frac{\partial^3 u}{\partial z^2 \partial t} = 0, \tag{3.2}$$

with a smooth initial condition

$$u(t = 0, z) = u_0(z), \qquad \eta(t = 0, z) = \eta_0(z).$$

Taking the space Fourier transform

$$\check{u}(t, k) = \frac{1}{2\pi} \int u(t, z) \exp(ikz) dz, \qquad \check{\eta}(t, k) = \frac{1}{2\pi} \int \eta(t, z) \exp(ikz) dz,$$

the Boussinesq equation (3.1)–(3.2) reduces to a set of ordinary differential equations:

$$\frac{\partial \check{\eta}}{\partial t} = ik\check{u}, \tag{3.3}$$

$$\left(1 + \beta k^2\right) \frac{\partial \check{u}}{\partial t} = ik\check{\eta}. \tag{3.4}$$

Introducing the pulsation corresponding to the wavenumber k through the *dispersion relation*

$$\omega(k) = \frac{k}{\sqrt{1+\beta k^2}}, \tag{3.5}$$

we get closed form expressions for the solutions:

$$\begin{aligned}
\check{u}(t,k) &= \frac{1}{2}\left(\check{u}_0(k) + \frac{\omega}{k}\check{\eta}_0(k)\right)\exp(i\omega t) + \frac{1}{2}\left(\check{u}_0(k) - \frac{\omega}{k}\check{\eta}_0(k)\right)\exp(-i\omega t),\\
\check{\eta}(t,k) &= \frac{1}{2}\left(\frac{k}{\omega}\check{u}_0(k) + \check{\eta}_0(k)\right)\exp(i\omega t) - \frac{1}{2}\left(\frac{k}{\omega}\check{u}_0(k) - \check{\eta}_0(k)\right)\exp(-i\omega t).
\end{aligned}$$

From these expressions we can conclude that any solution can be decomposed as the superposition of left-propagating modes $(u^{(l)}, \eta^{(l)})$ and right-propagating modes $(u^{(r)}, \eta^{(r)})$:

$$\begin{aligned}
u(t,z) &= u^{(r)}(t,z) + u^{(l)}(t,z),\\
\eta(t,z) &= \eta^{(r)}(t,z) + \eta^{(l)}(t,z),
\end{aligned}$$

where

$$\begin{aligned}
u^{(r)}(t,z) &= \int \frac{1}{2}\left(\check{u}_0(k) + \frac{\omega}{k}\check{\eta}_0(k)\right)\exp\left(i\omega(k)t - ikz\right)dk,\\
\eta^{(r)}(t,z) &= \int \frac{k}{2\omega}\left(\check{u}_0(k) + \frac{\omega}{k}\check{\eta}_0(k)\right)\exp\left(i\omega(k)t - ikz\right)dk,\\
u^{(l)}(t,z) &= \int \frac{1}{2}\left(\check{u}_0(k) - \frac{\omega}{k}\check{\eta}_0(k)\right)\exp\left(-i\omega(k)t - ikz\right)dk,\\
\eta^{(l)}(t,z) &= -\int \frac{k}{2\omega}\left(\check{u}_0(k) - \frac{\omega}{k}\check{\eta}_0(k)\right)\exp\left(-i\omega(k)t - ikz\right)dk.
\end{aligned}$$

This decomposition will be used in the nonhomogeneous case in the next section. In [18] a hyperbolic mode decomposition was used as an approximation for the right- and left-propagating modes. Here the mode decomposition is exact for dispersive waves.

4. Propagator formulation. In this section we first express the scattering problem as a two-point boundary value problem in the frequency domain, and then rewrite it as an initial value problem in terms of the propagator. This is the standard approach for acoustic equations [2] that we generalize to the dispersive case using the decomposition introduced in the previous section.

4.1. Mode propagation in the frequency domain. We consider the random Boussinesq equation (2.1)–(2.2) and take the time Fourier transform

$$\hat{u}(\omega,z) = \frac{1}{2\pi}\int u(t,z)\exp(-i\omega t)dt, \qquad \hat{\eta}(\omega,z) = \frac{1}{2\pi}\int \eta(t,z)\exp(-i\omega t)dt,$$

so that the system reduces to a set of ordinary differential equations:

$$\left(1 - \beta\omega^2(1+\varepsilon m(z))\right)\frac{\partial\hat{\eta}}{\partial z} + i\omega\hat{u} - \varepsilon\beta\omega^2 m'(z)\hat{\eta} = 0, \tag{4.1}$$

$$\frac{\partial\hat{u}}{\partial z} + i\omega\left(1+\varepsilon m(z)\right)\hat{\eta} = 0, \tag{4.2}$$

where m' stands for the spatial derivative of m. We introduce the wavenumber k corresponding to the pulsation ω,

$$k(\omega) = \frac{\omega}{\sqrt{1-\beta\omega^2}}, \tag{4.3}$$

so that we can decompose the wave into *right-going modes* A^ε and *left-going modes* B^ε over distances of propagation of order $1/\varepsilon^2$. We show explicitly the dependence on the small parameter ε:

$$A^\varepsilon(\omega,z) = \frac{1}{2}\left(\hat\eta\left(\omega,\frac{z}{\varepsilon^2}\right) + \frac{k}{\omega}\hat u\left(\omega,\frac{z}{\varepsilon^2}\right)\right), \tag{4.4}$$

$$B^\varepsilon(\omega,z) = \frac{1}{2}\left(\hat\eta\left(\omega,\frac{z}{\varepsilon^2}\right) - \frac{k}{\omega}\hat u\left(\omega,\frac{z}{\varepsilon^2}\right)\right). \tag{4.5}$$

The modes $(A^\varepsilon, B^\varepsilon)$ satisfy

$$\begin{aligned}\frac{\partial A^\varepsilon}{\partial z} &= -\frac{ik}{\varepsilon^2}A^\varepsilon - \frac{ik}{2\varepsilon}m\left(\frac{z}{\varepsilon^2}\right)(A^\varepsilon+B^\varepsilon) + \frac{\beta k^2}{2\varepsilon}m'\left(\frac{z}{\varepsilon^2}\right)(A^\varepsilon+B^\varepsilon)\\ &\quad - \frac{i\omega^2}{2k\varepsilon^2}\left(\frac{1}{1-\beta\omega^2(1+\varepsilon m(z/\varepsilon^2))} - \frac{1}{1-\beta\omega^2}\right)(A^\varepsilon - B^\varepsilon)\\ &\quad + \frac{\beta\omega^2}{2\varepsilon}m'\left(\frac{z}{\varepsilon^2}\right)\left(\frac{1}{1-\beta\omega^2(1+\varepsilon m(z/\varepsilon^2))} - \frac{1}{1-\beta\omega^2}\right)(A^\varepsilon+B^\varepsilon),\end{aligned} \tag{4.6}$$

$$\begin{aligned}\frac{\partial B^\varepsilon}{\partial z} &= \frac{ik}{\varepsilon^2}B^\varepsilon + \frac{ik}{2\varepsilon}m\left(\frac{z}{\varepsilon^2}\right)(A^\varepsilon+B^\varepsilon) + \frac{\beta k^2}{2\varepsilon}m'\left(\frac{z}{\varepsilon^2}\right)(A^\varepsilon+B^\varepsilon)\\ &\quad - \frac{i\omega^2}{2k\varepsilon^2}\left(\frac{1}{1-\beta\omega^2(1+\varepsilon m(z/\varepsilon^2))} - \frac{1}{1-\beta\omega^2}\right)(A^\varepsilon - B^\varepsilon)\\ &\quad + \frac{\beta\omega^2}{2\varepsilon}m'\left(\frac{z}{\varepsilon^2}\right)\left(\frac{1}{1-\beta\omega^2(1+\varepsilon m(z/\varepsilon^2))} - \frac{1}{1-\beta\omega^2}\right)(A^\varepsilon+B^\varepsilon).\end{aligned} \tag{4.7}$$

We expand the last terms of the right-hand sides up to $O(\varepsilon^3)$ terms

$$\frac{\omega^2}{1-\beta\omega^2(1+\varepsilon m(z/\varepsilon^2))} - \frac{\omega^2}{1-\beta\omega^2} = \varepsilon\beta k^4 m\left(\frac{z}{\varepsilon^2}\right) + \varepsilon^2\beta^2k^6m^2\left(\frac{z}{\varepsilon^2}\right) + O(\varepsilon^3), \tag{4.8}$$

where the $O(\varepsilon^3)$ is a term that can be bounded by $\varepsilon^3\beta^3k^8\|m\|_\infty^3/(1-\varepsilon\beta k^2\|m\|_\infty)$. We now look at the waves along the *frequency-dependent modified characteristics* defined by

$$a^\varepsilon(\omega,z) = A^\varepsilon(\omega,z)\exp\left(\frac{ikz}{\varepsilon^2}\right)\exp\left(-\frac{\varepsilon\beta k^2}{2}m\left(\frac{z}{\varepsilon^2}\right) - \frac{\varepsilon^2\beta^2k^4}{4}m\left(\frac{z}{\varepsilon^2}\right)^2\right), \tag{4.9}$$

$$b^\varepsilon(\omega,z) = B^\varepsilon(\omega,z)\exp\left(-\frac{ikz}{\varepsilon^2}\right)\exp\left(-\frac{\varepsilon\beta k^2}{2}m\left(\frac{z}{\varepsilon^2}\right) - \frac{\varepsilon^2\beta^2k^4}{4}m\left(\frac{z}{\varepsilon^2}\right)^2\right), \tag{4.10}$$

which satisfy the linear equation

$$\frac{\partial}{\partial z}\begin{pmatrix} a^\varepsilon \\ b^\varepsilon\end{pmatrix}(\omega,z) = Q^\varepsilon(\omega,z)\begin{pmatrix} a^\varepsilon \\ b^\varepsilon\end{pmatrix}(\omega,z). \tag{4.11}$$

The complex 2×2 matrix Q^ε is given by

$$Q^\varepsilon(\omega,z) = \begin{pmatrix} Q_1^\varepsilon(\omega,z) & Q_2^\varepsilon(\omega,z)e^{\frac{2ikz}{\varepsilon^2}} \\ \overline{Q_2^\varepsilon}(\omega,z)e^{-\frac{2ikz}{\varepsilon^2}} & \overline{Q_1^\varepsilon}(\omega,z)\end{pmatrix} \tag{4.12}$$

FIG. 4.1. *Scattering problem.*

with

$$Q_1^\varepsilon(\omega, z) = -\frac{ik}{2\varepsilon}\left(1+\beta k^2\right) m\left(\frac{z}{\varepsilon^2}\right) - \frac{i\beta^2 k^5}{2} m^2\left(\frac{z}{\varepsilon^2}\right) + O(\varepsilon), \tag{4.13}$$

$$Q_2^\varepsilon(\omega, z) = -\frac{ik}{2\varepsilon}\left(1-\beta k^2\right) m\left(\frac{z}{\varepsilon^2}\right) + \frac{\beta k^2}{2\varepsilon} m'\left(\frac{z}{\varepsilon^2}\right) + \frac{i\beta^2 k^5}{2} m^2\left(\frac{z}{\varepsilon^2}\right)$$

$$+ \frac{\beta^2 k^4}{2} m\left(\frac{z}{\varepsilon^2}\right) m'\left(\frac{z}{\varepsilon^2}\right) + O(\varepsilon). \tag{4.14}$$

The small terms of order ε come from the $O(\varepsilon^3)$ term in the expansion (4.8).

4.2. Boundary values. We assume that a left-going pulse is incoming from the right and is scattered into a reflected wave at $z = 0$ and a transmitted wave at $z = -L/\varepsilon^2$ (see Figure 4.1).

The incoming pulse shape is given by the elevation function $f(t)$, where f is assumed to be a L^1 function compactly supported in the Fourier domain:

$$u_{inc}(t, z=0) = -\int \frac{\omega}{k(\omega)} \hat{f}(\omega) \exp(i\omega t) d\omega, \tag{4.15}$$

$$\eta_{inc}(t, z=0) = \int \hat{f}(\omega) \exp(i\omega t) d\omega, \tag{4.16}$$

with $\mathrm{supp}(\hat{f}) \subset (-1/\sqrt{\beta}, 1/\sqrt{\beta})$. We also impose a radiation condition at $-L/\varepsilon^2$ corresponding to the absence of right-going waves at the left-hand side of the slab $[-L/\varepsilon^2, 0]$. The two-point boundary value problem consisting of the system (4.11) for $z \in [0, L]$, together with the conditions

$$b^\varepsilon(\omega, z=0) = \hat{f}(\omega), \qquad a^\varepsilon(\omega, z=-L) = 0,$$

is then well posed.

4.3. Propagator. It is convenient to transform the two-point boundary value problem into an initial value problem by introducing the propagator $Y^\varepsilon(\omega, -L, z)$, which is a complex 2×2 matrix solution of

$$\frac{\partial Y^\varepsilon}{\partial z}(\omega, -L, z) = Q^\varepsilon(\omega, z) Y^\varepsilon(\omega, -L, z), \qquad Y^\varepsilon(\omega, -L, z=-L) = Id_{\mathbb{C}^2}$$

such that

$$Y^\varepsilon(\omega, -L, z) \begin{pmatrix} a^\varepsilon(\omega, -L) \\ b^\varepsilon(\omega, -L) \end{pmatrix} = \begin{pmatrix} a^\varepsilon(\omega, z) \\ b^\varepsilon(\omega, z) \end{pmatrix}.$$

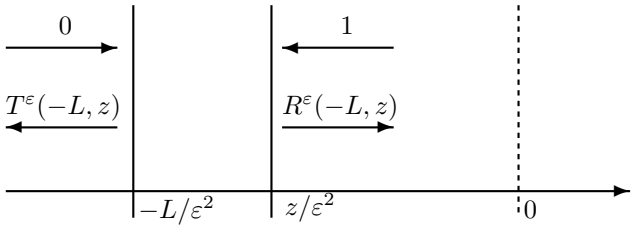


FIG. 4.2. *Reflection and transmission coefficients.*

By the form (4.12) of the matrix Q^ε, if the column vector $(a_1^\varepsilon, b_1^\varepsilon)^T$ is solution of (4.11) with the initial conditions

$$a_1^\varepsilon(\omega,-L)=1, \quad b_1^\varepsilon(\omega,-L)=0, \tag{4.17}$$

then the column vector $(\overline{b_1^\varepsilon}, \overline{a_1^\varepsilon})^T$ is another solution linearly independent of the first solution, so that the propagator matrix Y^ε can be written as

$$Y^\varepsilon(\omega,-L,z)=\begin{pmatrix} a_1^\varepsilon & \overline{b_1^\varepsilon} \\ b_1^\varepsilon & \overline{a_1^\varepsilon} \end{pmatrix}(\omega,z).$$

Note also that the matrix Q^ε has zero trace because $\overline{Q_1^\varepsilon} = -Q_1^\varepsilon$. As a consequence, the determinant of Y^ε is conserved, and $(a_1^\varepsilon, b_1^\varepsilon)$ satisfies the relation

$$\det Y^\varepsilon = |a_1^\varepsilon|^2 - |b_1^\varepsilon|^2 = 1. \tag{4.18}$$

We can now define the *transmission and reflection coefficients* $T^\varepsilon(\omega,-L,z)$ and $R^\varepsilon(\omega,-L,z)$, respectively, for a slab $[-L,z]$ by (see also Figure 4.2)

$$Y^\varepsilon(\omega,-L,z)\begin{pmatrix} 0 \\ T^\varepsilon(\omega,-L,z) \end{pmatrix} = \begin{pmatrix} R^\varepsilon(\omega,-L,z) \\ 1 \end{pmatrix}.$$

In terms of the propagator entries, they are given by

$$R^\varepsilon(\omega,-L,z)=\frac{\overline{b_1^\varepsilon}}{\overline{a_1^\varepsilon}}(\omega,z), \quad T^\varepsilon(\omega,-L,z)=\frac{1}{\overline{a_1^\varepsilon}}(\omega,z),$$

and they satisfy the closed form nonlinear differential system

$$\frac{\partial R^\varepsilon}{\partial z} = 2Q_1^\varepsilon(\omega,z)R^\varepsilon - e^{-\frac{2ikz}{\varepsilon^2}}\overline{Q_2^\varepsilon}(\omega,z)(R^\varepsilon)^2 + e^{\frac{2ikz}{\varepsilon^2}}Q_2^\varepsilon(\omega,z), \tag{4.19}$$

$$\frac{\partial T^\varepsilon}{\partial z} = -T^\varepsilon\left(e^{-\frac{2ikz}{\varepsilon^2}}\overline{Q_2^\varepsilon}(\omega,z)R^\varepsilon + \overline{Q_1^\varepsilon}(\omega,z)\right), \tag{4.20}$$

with the initial conditions at $z=-L$

$$R^\varepsilon(\omega,-L,z=-L)=0, \quad T^\varepsilon(\omega,-L,z=-L)=1.$$

Note that (4.18) implies the conservation of energy relation

$$|R^\varepsilon|^2 + |T^\varepsilon|^2 = 1 \tag{4.21}$$

and in turn the uniform boundedness of the transmission and reflection coefficients. Note also that R^ε and T^ε are the reflection and transmission coefficients for the modified characteristics (4.9)–(4.10). In terms of the real characteristics, the reflection and transmission coefficients are R^ε and $T^\varepsilon \exp(-ikL/\varepsilon^2)$, respectively.

4.4. Quantities of interest. The transmitted wave at time t, denoted by $(u^\varepsilon_{tr}, \eta^\varepsilon_{tr})$, is the left-going wave, which admits the following integral representation in terms of the transmission coefficients:

$$u^\varepsilon_{tr}\left(t, z=-\frac{L}{\varepsilon^2}\right) = -\int \frac{\omega}{k(\omega)}\hat{f}(\omega)T^\varepsilon(\omega,-L,0)\exp\left(i\omega t - ik(\omega)\frac{L}{\varepsilon^2}\right)d\omega, \tag{4.22}$$

$$\eta^\varepsilon_{tr}\left(t, z=-\frac{L}{\varepsilon^2}\right) = \int \hat{f}(\omega)T^\varepsilon(\omega,-L,0)\exp\left(i\omega t - ik(\omega)\frac{L}{\varepsilon^2}\right)d\omega. \tag{4.23}$$

Similarly, the reflected wave $(u^\varepsilon_{ref}, \eta^\varepsilon_{ref})$ can be expressed in terms of the reflection coefficients as

$$u^\varepsilon_{ref}(t, z=0) = \int \frac{\omega}{k(\omega)}\hat{f}(\omega)R^\varepsilon(\omega,-L,0)\exp\left(i\omega t\right)d\omega, \tag{4.24}$$

$$\eta^\varepsilon_{ref}(t, z=0) = \int \hat{f}(\omega)R^\varepsilon(\omega,-L,0)\exp\left(i\omega t\right)d\omega. \tag{4.25}$$

These are the quantities that we will use as new initial conditions for the time reversal experiments.

5. Time reversal setups.

5.1. Time reversal in reflection (TRR). The first step of the time reversal procedure consists of recording the reflected signal at $z = 0$ up to a certain time. It turns out that as $\varepsilon \to 0$ the interesting asymptotic regime arises when we record the signal up to a large time of order $1/\varepsilon^2$, which we denote by t_1/ε^2 with $t_1 > 0$. In the context of shallow water waves, one records only the elevation η_{ref}. If the recording were sufficiently long, one could deduce the depth-averaged velocity u_{ref} by using (4.24), (4.25), but this is not usually the case. If the recording is done over an approximately flat region, then, through (4.15), (4.16) and the proper zero-padding for Fourier transforming the elevation data $\eta_{ref} \equiv f$, the consistent incoming velocity field for the time reversal experiment can be well approximated. The zero-padding is due to the cut-off function of the recorded signal, as explained below.

In the second step of the time reversal procedure a piece of the recorded signal is cut using a cut-off function $s \mapsto G_{t_1}(\varepsilon^2 s)$, where the support of G_{t_1} is included in $[0, t_1]$:

$$\eta^\varepsilon_{ref,cut}\left(\frac{t}{\varepsilon^2}\right) = \eta^\varepsilon_{ref}\left(\frac{t}{\varepsilon^2}\right)G_{t_1}(t).$$

One then time reverses that piece of signal and re-emits the corresponding elevation field with a two-fold amplification. No velocity field is generated. This gives rise to a new wave that can be decomposed as the sum of a right-going wave and a left-going wave. The right-going wave propagates freely in the homogeneous right half-space, and it can be forgotten. The left-going wave is the new incoming signal. Accordingly, the elevation of the time reversed wave sent back into the medium is given by

$$\begin{aligned}\eta^\varepsilon_{inc(TRR)}\left(\frac{t}{\varepsilon^2}, z=0\right) &= \eta^\varepsilon_{ref}\left(\frac{t_1-t}{\varepsilon^2}\right)G_{t_1}(t_1-t)\\ &= \frac{1}{\varepsilon^2}\int\int \exp\left(\frac{i\omega(t_1-t)}{\varepsilon^2}\right)\hat{\eta}^\varepsilon_{ref}(\omega')\hat{G}_{t_1}\left(\frac{\omega-\omega'}{\varepsilon^2}\right)d\omega' d\omega\\ &= \frac{1}{\varepsilon^2}\int\int \exp\left(\frac{i\omega(t-t_1)}{\varepsilon^2}\right)\overline{\hat{\eta}^\varepsilon_{ref}}(\omega')\overline{\hat{G}}_{t_1}\left(\frac{\omega-\omega'}{\varepsilon^2}\right)d\omega' d\omega,\end{aligned} \tag{5.1}$$

where *TRR* stands for "time reversal in reflection." Here we have used the fact that η^ε_{ref} is a real-valued signal, and also that $k(-\omega) = -k(\omega)$, by (4.3), which is actually a direct consequence of the *time reversibility* of the Boussinesq equation. The new incoming (left-going) velocity is given by

$$u^\varepsilon_{inc(TRR)}\left(\frac{t}{\varepsilon^2}, z=0\right) = -\frac{1}{\varepsilon^2}\int\int \frac{\omega}{k(\omega)} e^{\frac{i\omega(t-t_1)}{\varepsilon^2}} \overline{\hat\eta^\varepsilon_{ref}}(\omega')\overline{\hat G}_{t_1}\left(\frac{\omega-\omega'}{\varepsilon^2}\right) d\omega' d\omega. \tag{5.2}$$

A right-going velocity wave is also generated, but it propagates freely with the right-going elevation wave mentioned above, and it can be forgotten as well. Note that the reason why we have amplified the generated elevation field by a factor two is that it gives rise to two counter-propagating waves which both contain half of the generated energy.

The new incoming signal (5.1)–(5.2) repropagates into the same medium and generates a new reflected signal which we observe at the time $t_2/\varepsilon^2 + t$, that is, around the time t_2/ε^2 in the scale of the initial pulse $f(t)$. In terms of the reflection coefficients the observed reflected elevation signal is given by

$$\eta^\varepsilon_{ref(TRR)}\left(\frac{t_2}{\varepsilon^2}+t, z=0\right) = \int \hat\eta^\varepsilon_{inc(TR)}(\omega) R^\varepsilon(\omega, -L, 0) e^{\frac{i\omega t_2}{\varepsilon^2}+i\omega t} d\omega.$$

Substituting the expression of $\hat\eta^\varepsilon_{inc(TRR)}$ into this equation yields the following representation of the reflected signal:

$$\eta^\varepsilon_{ref(TRR)}\left(\frac{t_2}{\varepsilon^2}+t, z=0\right) = \frac{1}{\varepsilon^2}\int\int e^{i\omega t} e^{\frac{i\omega(t_2-t_1)}{\varepsilon^2}} \overline{\hat f}(\omega')\overline{\hat G}_{t_1}\left(\frac{\omega-\omega'}{\varepsilon^2}\right) \times R^\varepsilon(\omega, -L, 0)\overline{R^\varepsilon}(\omega', -L, 0) d\omega' d\omega.$$

After the change of variable $\omega' = \omega - \varepsilon^2 h$, the representation becomes

$$\eta^\varepsilon_{ref(TRR)}\left(\frac{t_2}{\varepsilon^2}+t, z=0\right) = \int\int e^{i\omega t} e^{\frac{i\omega(t_2-t_1)}{\varepsilon^2}} \overline{\hat f}(\omega-\varepsilon^2 h)\overline{\hat G}_{t_1}(h) \times R^\varepsilon(\omega, -L, 0)\overline{R^\varepsilon}(\omega-\varepsilon^2 h, -L, 0) dh\, d\omega. \tag{5.3}$$

Note that, by (4.21), the reflection coefficients are bounded, and we shall show in section 6 that the rapid phase $\exp(i\omega(t_2-t_1)/\varepsilon^2)$ averages out the integral except when $t_2 = t_1$. This means that refocusing can be observed only at the time $t_2/\varepsilon^2 = t_1/\varepsilon^2$. The precise description of the refocused pulse, taking into account the interaction between randomness and dispersion, will be carried out in section 6.

5.2. Transmitted front wave. Before going into time reversal in transmission, we give an integral representation for the *coherent transmitted wavefront* observed at $z = -L/\varepsilon^2$ around the effective arrival time L/ε^2. By (4.23), the transmitted elevation front is given by

$$\eta^\varepsilon_{tr}\left(\frac{L}{\varepsilon^2}+t, z=-\frac{L}{\varepsilon^2}\right) = \int e^{i\omega t} e^{i(\omega-k(\omega))\frac{L}{\varepsilon^2}} \hat f(\omega) T^\varepsilon(\omega, -L, 0) d\omega. \tag{5.4}$$

Note that expressions like $t + L$ arise because constants have been set to one, so that the mean velocity is one. Due to dispersion, $k(\omega)$ is different from ω (see (4.3)). As a consequence, if $\beta = O(1)$, then the rapid phase $\exp(i(\omega - k(\omega))L/\varepsilon^2)$ makes the integral vanish as $\varepsilon \to 0$. This is in dramatic contrast with the hyperbolic case ($\beta = 0$),

where the coherent transmitted wave persists in this regime as a manifestation of the well known O'Doherty–Anstey theory studied in [8, 17, 25] in various situations.

In the dispersive case, the front will be present if the dispersion parameter β is small enough. This has been characterized and observed numerically in [18]. In particular, in the regime where $\beta = \varepsilon^2\beta_0$, we can derive the precise shape of the front resulting from the interplay of randomness and dispersion. In that regime, by expanding the dispersion relation $\omega \mapsto k(\omega)$, we get that the front is given by

$$\eta^\varepsilon_{tr}\left(\frac{L}{\varepsilon^2}+t, z=-\frac{L}{\varepsilon^2}\right) = \int e^{i\omega t}e^{-i\beta_0\omega^3 L}\hat{f}(\omega)T^\varepsilon(\omega,-L,0)d\omega + O(\varepsilon^2).$$

The transmission coefficients are given by $T^\varepsilon(\omega,-L,0) = 1/\overline{a^\varepsilon_1}(\omega,0)$, where a^ε_1 satisfies (4.11) with the initial conditions (4.17). In the case $\beta = \varepsilon^2\beta_0$, the entries of the matrix Q^ε can be expanded as

$$Q^\varepsilon_1(\omega,z)|_{\beta=\beta_0\varepsilon^2} = -\frac{ik}{2\varepsilon}m\left(\frac{z}{\varepsilon^2}\right) + O(\varepsilon),$$
$$Q^\varepsilon_2(\omega,z)|_{\beta=\beta_0\varepsilon^2} = -\frac{ik}{2\varepsilon}m\left(\frac{z}{\varepsilon^2}\right) + O(\varepsilon),$$

so that we get the same system as in the hyperbolic case up to terms of order ε. The limit of η^ε_{tr} has been derived for the hyperbolic case with small fluctuations [2, 25]. In our case the derivation of the limit follows the same lines except for the deterministic phase $\exp(-i\beta_0\omega^3 L)$ due to the small dispersion. The process $(\eta^\varepsilon_{tr}(\frac{L}{\varepsilon^2}+t, z=-\frac{L}{\varepsilon^2}))_{t\in(-\infty,+\infty)}$ converges in the space of the continuous and bounded functions to

$$\eta_{tr}(t) = \int \hat{f}(\omega)\exp\left(i\omega\left(t-\frac{\sqrt{\gamma(0)}}{\sqrt{2}}B_L\right) - \frac{\omega^2\gamma(\omega)}{4}L - i\beta_0\omega^3 L\right)d\omega,$$

where B_L is a standard Brownian motion and γ is

$$\gamma(\omega) = \int_0^\infty \mathbb{E}[m(0)m(z)]e^{2i\omega z}dz. \tag{5.5}$$

Using convolution operators, the transmitted front can be written in a simpler form

$$\eta_{tr}(t) = f * K\left(t-\frac{\sqrt{\gamma(0)}}{\sqrt{2}}B_L\right), \tag{5.6}$$

which means that a random Gaussian centering appears through the Brownian motion B_L, while the pulse shape spreads in a deterministic way through the convolution by the kernel K,

$$K(t) = K_r * K_d(t).$$

Here K_d is the scaled Airy function [1]

$$K_d(t) = \frac{1}{(3\beta_0 L)^{1/3}}\mathrm{Ai}\left(-\frac{t}{(3\beta_0 L)^{1/3}}\right),$$

and the Fourier transform of K_r is

$$\hat{K}_r(\omega) = \exp\left(-\frac{\omega^2\gamma(\omega)L}{4}\right).$$

Note that the kernel K depends both on randomness (through the function γ) and on dispersion (through the parameter β_0). This stochastic formulation is in agreement with the formulation presented in [18] for small β and was validated numerically with the same code used in this paper.

Observe that a dispersion parameter $\beta = O(1)$ or even $O(\varepsilon^p)$ with $p < 2$ leaves a fast phase in the integral representation of the transmitted front, as can be seen in (5.4). This implies a dramatic spreading of the pulse for a propagation distance of order $1/\varepsilon^2$, so that no coherent front pulse can be observed at the output $z = -L/\varepsilon^2$. In that case we are led to perform time reversal using the coda of the transmitted wave containing the incoherent fluctuations.

5.3. Time reversal in transmission (TRT). We now come back to the case of a dispersion parameter β of order 1. The time reversal procedure consists of recording the transmitted coda signal at $z = -L/\varepsilon^2$ over the time interval $[(L + t_0)/\varepsilon^2, (L + t_1)/\varepsilon^2]$. A piece of the recorded signal is cut using a cut-off function $s \mapsto G_{t_0,t_1}(\varepsilon^2 s - L)$, where the support of G_{t_0,t_1} is included in $[t_0, t_1]$:

$$\eta^\varepsilon_{tr,cut}\left(\frac{t}{\varepsilon^2}\right) = \eta^\varepsilon_{tr}\left(\frac{L+t}{\varepsilon^2}, z = -\frac{L}{\varepsilon^2}\right) G_{t_0,t_1}(t).$$

One then time reverses that piece of signal and sends it back into the same medium. As in section 5.1 one usually (only) records the elevation η_{tr}. Since the velocity field is not recorded, one actually generates the time reversed elevation field with an amplification by two, which in turn generates two counter-propagating waves with equal energies. Numerically we can record both the wave elevation and the velocity field. We will present examples comparing these two cases and show that the refocused pulse is the same. The elevation of the wave sent back is given by

$$\begin{aligned}\eta^\varepsilon_{inc(TRT)}\left(\frac{t}{\varepsilon^2}, z = -\frac{L}{\varepsilon^2}\right) &= \eta^\varepsilon_{tr}\left(\frac{L+t_1-t}{\varepsilon^2}, z = -\frac{L}{\varepsilon^2}\right) G_{t_0,t_1}(t_1 - t)\\ &= \frac{1}{\varepsilon^2}\int\int \exp\left(\frac{i\omega(t_1-t)}{\varepsilon^2}\right)\hat\eta^\varepsilon_{tr}(\omega')\hat G_{t_0,t_1}\left(\frac{\omega-\omega'}{\varepsilon^2}\right)d\omega' d\omega,\end{aligned}$$

where $\hat\eta^\varepsilon_{tr}$ is the Fourier transform of the shifted received signal $t \mapsto \eta^\varepsilon_{tr}(\frac{L+t}{\varepsilon^2}, z = -\frac{L}{\varepsilon^2})$:

$$\hat\eta^\varepsilon_{tr}(\omega) = e^{i(\omega - k(\omega))\frac{L}{\varepsilon^2}}\hat f(\omega) T^\varepsilon(\omega, -L, 0).$$

Also $\eta^\varepsilon_{inc(TRT)}$ reads as

$$\eta^\varepsilon_{inc(TRT)}\left(\frac{t}{\varepsilon^2}, z = -\frac{L}{\varepsilon^2}\right) = \frac{1}{\varepsilon^2}\int\int \exp\left(\frac{i\omega(t-t_1)}{\varepsilon^2}\right)\overline{\hat\eta^\varepsilon_{tr}}(\omega')\overline{\hat G}_{t_0,t_1}\left(\frac{\omega-\omega'}{\varepsilon^2}\right)d\omega' d\omega.$$

Let us denote by $\tilde R^\varepsilon$ and $\tilde T^\varepsilon$ the reflection and transmission coefficients for the experiment corresponding to a right-going input wave incoming from the left (see Figure 5.1). Using the propagator Y^ε defined in section 4.3, $\tilde R^\varepsilon$ and $\tilde T^\varepsilon$ obey the relation

$$Y^\varepsilon(\omega, -L, 0)\begin{pmatrix} 1 \\ \tilde R^\varepsilon(\omega, -L, 0)\end{pmatrix} = \begin{pmatrix} \tilde T^\varepsilon(\omega, -L, 0) \\ 0\end{pmatrix}.$$

In terms of the propagator entries they are given by

$$\tilde R^\varepsilon(\omega, -L, 0) = -\frac{b^\varepsilon_1}{a^\varepsilon_1}(\omega, 0), \qquad \tilde T^\varepsilon(\omega, -L, 0) = \frac{1}{a^\varepsilon_1}(\omega, 0),$$

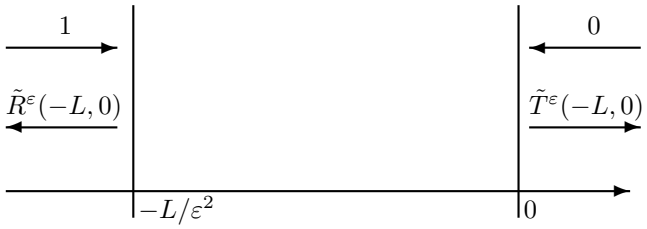


FIG. 5.1. *Adjoint reflection and transmission coefficients for time reversal.*

which shows that

$$\tilde{T}^\varepsilon(\omega,-L,0)=T^\varepsilon(\omega,-L,0).$$

Accordingly, the new incoming signal repropagates into the same medium and generates a new transmitted signal, which we observe at the time t_2/ε^2+t, that is, around the time t_2/ε^2 in the scale of the initial pulse $f(t)$. In terms of the transmission coefficients, the observed transmitted elevation signal is given by

$$\eta^\varepsilon_{tr(TRT)}\left(\frac{t_2}{\varepsilon^2}+t,z=0\right)=\int\hat{\eta}^\varepsilon_{inc(TRT)}(\omega)T^\varepsilon(\omega,-L,0)e^{\frac{i\omega t_2}{\varepsilon^2}+i\omega t}e^{-ik(\omega)\frac{L}{\varepsilon^2}}d\omega.$$

Substituting the expression of $\hat{\eta}^\varepsilon_{inc(TRT)}$ into this equation yields the following representation of the new transmitted signal:

$$\eta^\varepsilon_{tr(TRT)}\left(\frac{t_2}{\varepsilon^2}+t,z=0\right)=\frac{1}{\varepsilon^2}\int\int e^{i\omega t}e^{\frac{i\omega(t_2-t_1-L)}{\varepsilon^2}}\overline{\hat{f}}(\omega')\overline{\hat{G}}_{t_0,t_1}\left(\frac{\omega-\omega'}{\varepsilon^2}\right)$$
$$\times e^{i(k(\omega')-k(\omega))\frac{L}{\varepsilon^2}}e^{-i(\omega'-\omega)\frac{L}{\varepsilon^2}}T^\varepsilon(\omega,-L,0)\overline{T^\varepsilon}(\omega',-L,0)d\omega'd\omega.$$

After the change of variable $\omega'=\omega-\varepsilon^2h$, the representation becomes

$$\eta^\varepsilon_{tr(TRT)}\left(\frac{t_2}{\varepsilon^2}+t,z=0\right)=\int\int e^{i\omega t}e^{\frac{i\omega(t_2-t_1-L)}{\varepsilon^2}}\overline{\hat{f}}(\omega-\varepsilon^2h)\overline{\hat{G}}_{t_0,t_1}(h)$$
$$\times e^{i(k(\omega-\varepsilon^2h)-k(\omega))\frac{L}{\varepsilon^2}}e^{ihL}T^\varepsilon(\omega,-L,0)\overline{T^\varepsilon}(\omega-\varepsilon^2h,-L,0)dh\;d\omega. \tag{5.7}$$

The precise asymptotics of the transmitted wave will be carried out in section 7. It is easily seen that the refocusing will only take place if $t_2=L+t_1$ due to the fast phase.

5.4. TRT in homogeneous medium. One application of TRT is source reconstruction when the medium is known. This is motivated by the problem of waveform inversion for water waves studied in [24], where the goal is to characterize the initial sea surface displacement due to tsunamigenic earthquakes. Mathematically, in the context of this paper, the source inversion problem consists of performing TRT. The repropagation of the time reversed transmitted wave is performed by solving numerically the corresponding wave equation. In the case of the time reversal experiment for a dispersive homogeneous medium, we observe a transmitted signal and would like to recover both the location and the pulse shape of the source. This implies the recompression of the dispersive oscillatory coda of the transmitted wave. Dispersion helps with the source location identification. This is in contrast with (traveling) hyperbolic waves in a homogeneous medium.

Taking $T^\varepsilon = 1$ in (5.7) gives the transmitted signal in homogeneous medium. Observe that the quantities become independent of ε, so that ε can be taken to be equal to 1. We then get

$$\eta_{tr(TRT)}(t_1 + L + t, z) = \int\int e^{i\omega t - ik(\omega)z} \overline{\hat{f}}(\omega - h) \overline{\hat{G}}_{t_0,t_1}(h) e^{i(k(\omega-h)-k(\omega))L} e^{ihL} dh \; d\omega,$$

where we look at two cases, as follows.

(a) Hyperbolic case. If $\beta = 0$, then $k(\omega) = \omega$, and so the transmitted wave is

$$\eta_{tr(TRT)}(t_1 + L + t, z) = \int\int e^{i\omega(t-z)} \overline{\hat{f}}(\omega - h) \overline{\hat{G}}_{t_0,t_1}(h) dh \; d\omega,$$

which yields a traveling wave

$$\eta_{tr(TRT)}(t_1 + L + t, z) = (G_{t_0,t_1} f)(z - t).$$

On the one hand, it is impossible to retrieve the source location from this traveling wave. On the other hand, as soon as the support of the cut-off function is larger than the pulse width, then the reconstruction of the pulse shape is perfect.

(b) Dispersive case. If $\beta \neq 0$ and $(\beta L)^{1/3}$ is much larger than the pulse width, then

$$\eta_{tr(TRT)}(t_1 + L + t, z) = K_{z,L} * f(z - t),$$

where the kernel $K_{z,L}$ is given by

$$\begin{aligned} K_{z,L}(t) &= K_z * K_L(t), \\ K_z(t) &= \frac{1}{(3\beta z)^{1/3}} \mathrm{Ai}\left(\frac{t}{(3\beta z)^{1/3}}\right), \\ \hat{K}_L(\omega) &= G_{t_0,t_1}\left(((1 + \beta k(\omega)^2)^{3/2} - 1)L\right). \end{aligned}$$

The Airy kernel K_z results from the action of dispersion on the refocused pulse around the original source location. Let us define $z_c = T_w^3/(3\beta)$, where T_w is the pulse width. If $z < -z_c$, then pulse refocusing is not yet completed and the oscillatory tail is not yet recompressed. If $z > z_c$, then the pulse starts developing the dispersive tail again. When $z \in [-z_c, z_c]$, the oscillatory tail vanishes and the kernel $K_z$is close to a Dirac mass. This shows that *dispersion enhances the resolution of the source location* since z_c decays with increasing β. However *the reconstruction of the source shape is blurred by dispersive effects* since the cut-off function G deletes a frequency band that becomes larger as β is larger.

6. Asymptotics of the refocused pulse in reflection. From now on we assume that β is of order 1. The integral representation (5.3) of the reflected signal shows that the autocorrelation function of the reflection coefficient at two nearby frequencies will play an important role.

6.1. The frequency autocorrelation function of the reflection coefficient. We shall study the symmetric version

$$U^\varepsilon_{1,1}(\omega, h, z) = R^\varepsilon\left(\omega + \frac{\varepsilon^2 h}{2}, -L, z\right) \overline{R^\varepsilon}\left(\omega - \frac{\varepsilon^2 h}{2}, -L, z\right),$$

and we shall extend the approach developed in [2, 7] to the dispersive case. It is necessary to consider a family of moments so as to get a closed system of equations. We thus introduce for $n, p \in \mathbb{N}$

$$U^{\varepsilon}_{n,p}(\omega, h, z) = \left(R^{\varepsilon}\left(\omega + \frac{\varepsilon^2 h}{2}, -L, z\right)\right)^n \left(\overline{R^{\varepsilon}}\left(\omega - \frac{\varepsilon^2 h}{2}, -L, z\right)\right)^p.$$

Denoting

$$k'(\omega) = \frac{\partial k}{\partial \omega}(\omega) = \frac{1}{(1-\beta\omega^2)^{3/2}} = (1+\beta k^2)^{3/2} \tag{6.1}$$

and using the Riccati equation (4.19) satisfied by R^{ε}, we deduce

$$\begin{aligned}\frac{\partial U^{\varepsilon}_{n,p}}{\partial z} &= 2(n-p)Q^{\varepsilon}_1 U^{\varepsilon}_{n,p} + Q^{\varepsilon}_2 e^{\frac{2ik(\omega)z}{\varepsilon^2}}\left(ne^{ik'(\omega)hz}U^{\varepsilon}_{n-1,p} - pe^{-ik'(\omega)hz}U^{\varepsilon}_{n,p+1}\right)\\ &\quad + \overline{Q^{\varepsilon}_2} e^{-\frac{2ik(\omega)z}{\varepsilon^2}}\left(pe^{ik'(\omega)hz}U^{\varepsilon}_{n,p-1} - ne^{-ik'(\omega)hz}U^{\varepsilon}_{n+1,p}\right)\end{aligned}$$

starting from

$$U^{\varepsilon}_{n,p}(\omega, h, z = -L) = \mathbf{1}_0(n)\mathbf{1}_0(p),$$

where $\mathbf{1}_0(n) = 1$ if $n = 0$ and 0 otherwise. Taking a shifted scaled Fourier transform with respect to h,

$$V^{\varepsilon}_{n,p}(\omega, \tau, z) = \frac{k'(\omega)}{2\pi}\int e^{ihk'(\omega)(\tau-(n+p)z)}U^{\varepsilon}_{n,p}(\omega, h, z)dh,$$

we get

$$\begin{aligned}\frac{\partial V^{\varepsilon}_{n,p}}{\partial z} &= -(n+p)\frac{\partial V^{\varepsilon}_{n,p}}{\partial \tau} + 2(n-p)Q^{\varepsilon}_1 V^{\varepsilon}_{n,p}\\ &\quad + Q^{\varepsilon}_2 e^{\frac{2ik(\omega)z}{\varepsilon^2}}\left(nV^{\varepsilon}_{n-1,p} - pV^{\varepsilon}_{n,p+1}\right) + \overline{Q^{\varepsilon}_2} e^{-\frac{2ik(\omega)z}{\varepsilon^2}}\left(pV^{\varepsilon}_{n,p-1} - nV^{\varepsilon}_{n+1,p}\right)\end{aligned}$$

starting from

$$V^{\varepsilon}_{n,p}(\omega, \tau, z = -L) = \delta(\tau)\mathbf{1}_0(n)\mathbf{1}_0(p).$$

Applying a diffusion-approximation theorem [2, section 3] establishes that the processes $V^{\varepsilon}_{n,p}$ converge to diffusion processes as $\varepsilon \to 0$. In particular, the expectations $\mathbb{E}[V^{\varepsilon}_{n,n}(\omega, \tau, z)]$, $n \in \mathbb{N}$, converge to $W_n(\omega, \tau, z)$, which obey the closed system of transport equations

$$\begin{aligned}&\frac{\partial W_n}{\partial z} + 2n\frac{\partial W_n}{\partial \tau} = \frac{1}{2}\alpha_\beta(\omega)k(\omega)^2 n^2\left(W_{n+1} + W_{n-1} - 2W_n\right),\\ &W_n(\omega, \tau, z = -L) = \delta(\tau)\mathbf{1}_0(n),\end{aligned} \tag{6.2}$$

where

$$\alpha_\beta(\omega) = \alpha(k(\omega))(1+\beta k(\omega)^2)^2 = \frac{\alpha(\omega/\sqrt{1-\beta\omega^2})}{(1-\beta\omega^2)^2} \tag{6.3}$$

and α is proportional to the power spectral density of the random process m:

$$\alpha(k) = \int_0^\infty \mathbb{E}[m(0)m(z)]\cos(2kz)dz. \tag{6.4}$$

Note that the limit transport equations (6.2) have the same form as those obtained in the nondispersive case in [2]. The difference is contained in the rate coefficient $\alpha_\beta(\omega)k(\omega)^2$, which is simply $\alpha_0(\omega)\omega^2$ in the hyperbolic case. We then get the limit of the autocorrelation function of the reflection coefficient:

$$\mathbb{E}\left[R^\varepsilon\left(\omega+\frac{\varepsilon^2 h}{2},-L,0\right)\overline{R^\varepsilon}\left(\omega-\frac{\varepsilon^2 h}{2},-L,0\right)\right]\overset{\varepsilon\to 0}{\longrightarrow}\int\Lambda^L_{ref}(\omega,\tau)e^{-ih\tau}d\tau, \tag{6.5}$$

$$\Lambda^L_{ref}(\omega,\tau)=k'(\omega)^{-1}W_1(\omega,k'(\omega)^{-1}\tau,0). \tag{6.6}$$

The quantity $W_1(\omega,\tau,0)$ is obtained through the system of transport equations (6.2), which we study in the next section.

6.2. Analysis of the transport equations. We can interpret the transport equation (6.2) in terms of a jump Markov process. Let us introduce the process $(N_t)_{t\geq 0}$ with state space $\mathbb{N}$ and infinitesimal generator

$$\mathcal{L}\phi(N)=\frac{1}{2}\alpha_\beta(\omega)k(\omega)^2N^2\left(\phi(N+1)+\phi(N-1)-2\phi(N)\right).$$

As in [2], we deduce

$$\int_{\tau_0}^{\tau_1}W_1(\omega,\tau,0)d\tau=\mathbb{P}_1\left(\int_0^L 2N_s ds\in[\tau_0,\tau_1],\ N_L=0\right),$$

where $\mathbb{P}_{p_0}$ stands for the probability over the distribution of the jump process starting from $N_0=p_0$. Taking $\tau_0=0$ and $\tau_1=\infty$ yields

$$\mathbb{E}\left[|R^\varepsilon|^2(\omega,-L,0)\right]\overset{\varepsilon\to 0}{\longrightarrow}\mathbb{P}_1(N_L=0).$$

It is remarkable that the generating function of the jump process can be expressed in terms of the expectation of some functional of the diffusion process $(\theta_t)_{t\geq 0}$:

$$d\theta_t=\sqrt{\alpha_\beta(\omega)}k(\omega)dB_t+\frac{1}{2}\alpha_\beta(\omega)k(\omega)^2\coth(\theta_t)dt. \tag{6.7}$$

We have

$$\mathbb{E}_{p_0}\left[z^{N_t}\right]=\mathbb{E}\left[\tanh\left(\frac{\theta_t}{2}\right)^{2p_0}\ \middle|\ \theta_0=2\ \mathrm{argtanh}(\sqrt{z})\right],$$

where $\mathbb{E}_{p_0}$ stands for the expectation with respect to the distribution of the jump process starting from $N_0=p_0$. In particular,

$$\mathbb{E}\left[|R^\varepsilon|^2(\omega,-L,0)\right]\overset{\varepsilon\to 0}{\longrightarrow}\mathbb{P}_1(N_L=0)=\mathbb{E}\left[\tanh\left(\frac{\theta_L}{2}\right)^2\ \middle|\ \theta_0=0\right].$$

As the probability density function of the diffusion process (θ_t) is known [21], we get

$$\mathbb{E}\left[|R^\varepsilon|^2(\omega,-L,0)\right]\overset{\varepsilon\to 0}{\longrightarrow}1-\frac{4}{\sqrt{\pi}}\exp\left(-\frac{L}{l_\beta(\omega)}\right)\int_0^\infty\frac{x^2e^{-x^2}}{\cosh\left(2\sqrt{L/l_\beta(\omega)}x\right)}dx,$$

where the localization length $l_\beta(\omega)$ of the mean transmittance is affected by the dispersion

$$l_\beta(\omega) = \frac{8}{\alpha_\beta(k(\omega))k(\omega)^2} = \frac{8(1-\beta\omega^2)^3}{\alpha(\omega/\sqrt{1-\beta\omega^2}))\omega^2}. \tag{6.8}$$

If the power spectral density of the process m can be considered as constant $\alpha(k) \equiv \alpha_0$, that is to say, when the correlation length of the medium is smaller than the typical wavelength of the pulse, then the above expression of the localization length shows that *dispersion enhances localization effects*. The decay of the localization length as a function of frequency is faster in the dispersive case than in the hyperbolic case. This has been observed numerically in [18].

6.3. The refocused pulse. Choosing $t_2 = t_1$ in (5.3) shows that the refocused pulse at $z = 0$ is given by the integral representation

$$\begin{aligned}\eta^\varepsilon_{ref(TRR)}\left(\frac{t_1}{\varepsilon^2}+t, z=0\right) = \int\int e^{i\omega t}\overline{\hat{f}}(\omega-\varepsilon^2 h)\overline{\hat{G}}_{t_1}(h) \\ \times R^\varepsilon(\omega,-L,0)\overline{R^\varepsilon}(\omega-\varepsilon^2 h,-L,0)dh\, d\omega.\end{aligned} \tag{6.9}$$

The main result of this section is the *self-averaging property* of the refocused pulse. This is shown in the following theorem, which gives the convergence of the refocused pulse to a deterministic shape.

THEOREM 6.1. *For any $T > 0$, $\delta > 0$,*

$$\mathbb{P}\left(\sup_{t\in[-T,T]}\left|\eta^\varepsilon_{ref(TRR)}\left(\frac{t_1}{\varepsilon^2}+t, z=0\right) - \eta_{ref(TRR)}(t)\right| > \delta\right) \overset{\varepsilon\to 0}{\longrightarrow} 0,$$

where $\eta_{ref(TRR)}$ is the deterministic pulse shape:

$$\eta_{ref(TRR)}(t) = \left(f(-\,\cdot) * K_{TRR}(\cdot)\right)(t). \tag{6.10}$$

The Fourier transform of the kernel K_{TRR} is the convolution of the time-inverted cut-off function G_{t_1} with the density $\tau \mapsto \Lambda^L_{ref}(\omega,\tau)$ evaluated at 0:

$$\hat{K}_{TRR}(\omega) = \left(G_{t_1}(-\,\cdot) * \Lambda^L_{ref}(\omega,\cdot)\right)(0) = \int G_{t_1}(\tau)\Lambda^L_{ref}(\omega,\tau)d\tau. \tag{6.11}$$

Proof. The first step consists of proving the tightness (i.e., the relative compactness) in the space of continuous trajectories (equipped with the topology associated to the sup norm over the compact subsets) of the family of continuous processes

$$\left(\left(\eta^\varepsilon_{ref(TRR)}\left(\frac{t_1}{\varepsilon^2}+t, z=0\right)\right)_{-\infty<t<\infty}\right)_{\varepsilon>0}.$$

From (6.9) and the uniform bound $|R^\varepsilon| \le 1$, it is easily seen that the quantity $|\eta^\varepsilon_{ref(TRR)}(t_1/\varepsilon^2 + t, z = 0)|$ is uniformly bounded by

$$\int |\hat{G}(h)|dh \times \int |\hat{f}(\omega)|d\omega,$$

which we assume finite. The modulus of continuity

$$\Omega^\varepsilon(s) = \sup_{|s_1-s_2|\le s} \left| \eta^\varepsilon_{ref(TRR)}\left(\frac{t_1}{\varepsilon^2}+s_1, z=0\right) - \eta^\varepsilon_{ref(TRR)}\left(\frac{t_1}{\varepsilon^2}+s_2, z=0\right)\right|$$

is bounded by

$$\Omega^\varepsilon(s) \le \int |\hat{G}(h)|dh \times \int \sup_{|s_1-s_2|\le s} |e^{i\omega(s_2-s_1)}-1||\hat{f}(\omega)|d\omega,$$

which goes to zero as $s \to 0$ uniformly with respect to ε, by Lebesgue's theorem, and ensures tightness.

Taking the expectation in (6.9) and using (6.5), we get the convergence of the first moment:

$$\begin{aligned}\lim_{\varepsilon\to 0} \mathbb{E}\left[\eta^\varepsilon_{ref(TRR)}\left(\frac{t_1}{\varepsilon^2}+t, z=0\right)\right] &= \int\int e^{i\omega t}\overline{\hat{f}}(\omega)\overline{\hat{G}}_{t_1}(h)\int e^{-ih\tau}\Lambda^L_{ref}(\omega,\tau)d\tau dh d\omega \\ &= \int\int e^{i\omega t}\overline{\hat{f}}(\omega)G_{t_1}(\tau)\Lambda^L_{ref}(\omega,\tau)d\tau d\omega \\ &= \left(f(-\,.) * K_{TRR}(.)\right)(t) = \eta_{ref(TRR)}(t).\end{aligned}$$

In order to prove the convergence in probability of $\eta^\varepsilon_{ref(TRR)}(t_1/\varepsilon^2+t, z=0)$ to the deterministic refocused pulse $\eta_{ref(TRR)}$, we compute its second moment and show that it converges to the square of the first moment obtained above. This computation has been done in the acoustic case [9] using the moment analysis of the reflected signal established in [7]. The second moment involves the moment of the product of the reflection coefficients at four frequencies. The presence of the cut-off function G_{t_1} used in time reversal automatically pairs the frequencies. The moment analysis then shows that the reflection coefficients for the two pairs become independent, which proves the result. The same techniques apply to the dispersive case since the Riccati equation (4.19) for the reflection coefficient has the same form as in the acoustic case. □

As for acoustic waves, the case of a large slab (L large) leads to explicit formulas for the refocused pulse. This is developed in the following section.

6.4. Large slab. For acoustic waves the hyperbolicity of the equations makes the reflected quantities of interest independent of L for L large enough. This leads to explicit formulas for the power spectral density Λ^L_{ref}. In our context of dispersive waves, the velocities of the waves are still bounded as we consider a pulse with compactly supported spectrum. For this reason, the power spectral density also becomes independent of L for L large enough. Applying the same approach as in [2] (where the case of acoustic waves was addressed), we get that the function Λ^L_{ref} converges as L grows to infinity to the limit density

$$\Lambda^\infty_{ref}(\omega,\tau) = \frac{\kappa_\beta(\omega)\omega^2}{\left(1+\kappa_\beta(\omega)\omega^2\tau\right)^2}, \tag{6.12}$$

where

$$\kappa_\beta(\omega) = \frac{\alpha_\beta(\omega)k(\omega)^2}{4\omega^2 k'(\omega)} = \frac{\alpha(\omega/\sqrt{1-\beta\omega^2})}{4\left(1-\beta\omega^2\right)^{3/2}}.$$

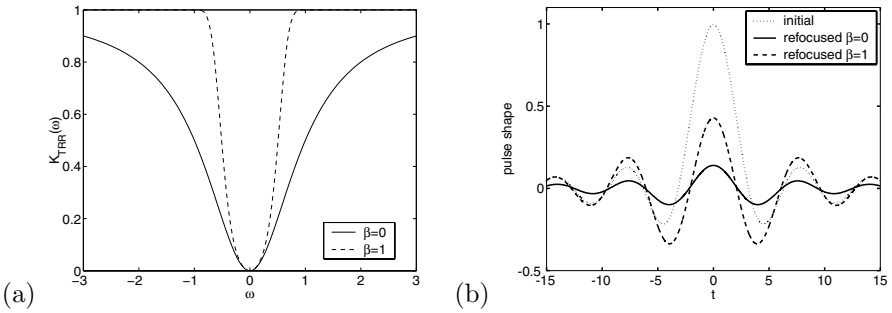


FIG. 6.1. *Fourier transform of the convolution kernel* K_{TRR} (a) *and refocused pulse* (b). *We consider a square cut-off function* $G(t) = \mathbf{1}_{[0,t_1]}(t)$, *and we assume* $\alpha(\omega) \equiv 4$, $t_1 = 1$. *The initial pulse has sinc shape* $f(t) = \sin(t)/t$, *and its spectrum is* $\hat{f}(\omega) = (1/2)\mathbf{1}_{[-1,1]}(\omega)$. *(Note that* $\beta = 1$ *corresponds to a very dispersive configuration.)*

The deterministic refocused pulse is then given by (6.10) with the explicit Λ^{∞}_{ref} derived in this section. Taking, for instance, a square cut-off function $G_{t_1}(t) = \mathbf{1}_{[0,t_1]}(t)$, the kernel K_{TRR} reads as a high-band filter because its Fourier transform is

$$\hat{K}_{TRR}(\omega) = \frac{\kappa_\beta(\omega)\omega^2 t_1}{1 + \kappa_\beta(\omega)\omega^2 t_1}. \tag{6.13}$$

An example is presented in Figure 6.1. The cut-off frequency of the filter $\hat{K}_{TRR}$ decays with increasing dispersion parameter β. This shows that time reversal focusing in reflection is more efficient in the dispersive case than in the hyperbolic case. This is consistent with the observation that localization effects are enhanced in the presence of dispersion (see (6.8)).

7. Asymptotics of refocused pulse in transmission.

7.1. The frequency autocorrelation function of the transmission coefficient. We study here the autocorrelation function of the transmission coefficient at two nearby frequencies. We first define a new family of processes indexed by $n, p \in \mathbb{N}$,

$$\tilde{U}^{\varepsilon}_{n,p}(\omega, h, z) = U^{\varepsilon}_{n,p}(\omega, h, z) T^{\varepsilon}\left(\omega + \frac{\varepsilon^2 h}{2}, -L, z\right) \overline{T^{\varepsilon}}\left(\omega - \frac{\varepsilon^2 h}{2}, -L, z\right),$$

which satisfy

$$\begin{aligned}\frac{\partial \tilde{U}^{\varepsilon}_{n,p}}{\partial z} &= 2(n-p)Q^{\varepsilon}_1 \tilde{U}^{\varepsilon}_{n,p} + Q^{\varepsilon}_2 e^{\frac{2ik(\omega)z}{\varepsilon^2}}\left(n e^{ik'(\omega)hz}\tilde{U}^{\varepsilon}_{n-1,p} - (p+1)e^{-ik'(\omega)hz}\tilde{U}^{\varepsilon}_{n,p+1}\right) \\ &\quad + \overline{Q^{\varepsilon}_2} e^{-\frac{2ik(\omega)z}{\varepsilon^2}}\left(p e^{ik'(\omega)hz}\tilde{U}^{\varepsilon}_{n,p-1} - (n+1)e^{-ik'(\omega)hz}\tilde{U}^{\varepsilon}_{n+1,p}\right)\end{aligned}$$

starting from

$$\tilde{U}^{\varepsilon}_{n,p}(\omega, h, z = -L) = \mathbf{1}_0(n)\mathbf{1}_0(p).$$

Taking a shifted scaled Fourier transform with respect to h,

$$\tilde{V}^{\varepsilon}_{n,p}(\omega, \tau, z) = \frac{k'(\omega)}{2\pi}\int e^{ihk'(\omega)(\tau - (n+p)z)}\tilde{U}^{\varepsilon}_{n,p}(\omega, h, z)dh,$$

we get

$$\frac{\partial \tilde{V}^{\varepsilon}_{n,p}}{\partial z} = -(n+p)\frac{\partial \tilde{V}^{\varepsilon}_{n,p}}{\partial \tau} + 2(n-p)Q^{\varepsilon}_1 \tilde{V}^{\varepsilon}_{n,p}$$
$$+ Q^{\varepsilon}_2 e^{\frac{2ik(\omega)z}{\varepsilon^2}} \left(n\tilde{V}^{\varepsilon}_{n-1,p} - (p+1)\tilde{V}^{\varepsilon}_{n,p+1} \right) + \overline{Q^{\varepsilon}_2} e^{-\frac{2ik(\omega)z}{\varepsilon^2}} \left(p\tilde{V}^{\varepsilon}_{n,p-1} - (n+1)\tilde{V}^{\varepsilon}_{n+1,p} \right)$$

starting from

$$\tilde{V}^{\varepsilon}_{n,p}(\omega, \tau, z = -L) = \delta(\tau)\mathbf{1}_0(n)\mathbf{1}_0(p).$$

Applying a diffusion-approximation theorem [2, section 3.14] establishes that the processes $\tilde{V}^{\varepsilon}_{n,p}$ converge to diffusion processes as $\varepsilon \to 0$. In particular, the expectations $\mathbb{E}[\tilde{V}^{\varepsilon}_{n,n}(\omega, \tau, z)]$ converge to $\tilde{W}_n(\omega, \tau, z)$, which obey the closed system of transport equations

$$\frac{\partial \tilde{W}_n}{\partial z} + 2n\frac{\partial \tilde{W}_n}{\partial \tau} = \frac{1}{2}\alpha_\beta(\omega)k(\omega)^2 \left((n+1)^2\tilde{W}_{n+1} + n^2\tilde{W}_{n-1} - ((n+1)^2 + n^2)\tilde{W}_n \right),$$
$$\tilde{W}_n(\omega, \tau, z = -L) = \delta(\tau)\mathbf{1}_0(n).$$

We then get the limit of the autocorrelation function of the transmission coefficient:

$$\mathbb{E}\left[T^{\varepsilon}\left(\omega + \frac{\varepsilon^2 h}{2}, -L, 0 \right) \overline{T^{\varepsilon}}\left(\omega - \frac{\varepsilon^2 h}{2}, -L, 0 \right) \right] \overset{\varepsilon \to 0}{\longrightarrow} \int \Lambda^L_{tr}(\omega, \tau) e^{-ih\tau} d\tau, \tag{7.1}$$

$$\Lambda^L_{tr}(\omega, \tau) = k'(\omega)^{-1} \tilde{W}_0(\omega, k'(\omega)^{-1}\tau, 0). \tag{7.2}$$

7.2. Analysis of the transport equations. We can interpret the transport equation in terms of a jump Markov process as in section 6.2. Let us introduce the process $(\tilde{N}_t)_{t \geq 0}$ with state space $\mathbb{N}$ and infinitesimal generator:

$$\tilde{\mathcal{L}}\phi(\tilde{N}) = \frac{1}{2}\alpha_\beta(\omega)k(\omega)^2 \left((\tilde{N}+1)^2(\phi(\tilde{N}+1) - \phi(\tilde{N})) + \tilde{N}^2(\phi(\tilde{N}-1) - \phi(\tilde{N})) \right).$$

Note that $\tilde{\mathcal{L}}$ is the adjoint of the generator $\mathcal{L}$ of the process $(N_t)_{t \geq 0}$, which means that $(\tilde{N}_t)_{t \geq 0}$ is the time reversed process of $(N_t)_{t \geq 0}$. We have

$$\int_{\tau_0}^{\tau_1} \tilde{W}_0(\omega, d\tau, 0) = \tilde{\mathbb{P}}_0 \left(\int_0^L 2\tilde{N}_s ds \in [\tau_0, \tau_1],\ \tilde{N}_L = 0 \right), \tag{7.3}$$

where $\tilde{\mathbb{P}}_{p_0}$ stands for the probability over the distribution of the jump process starting from $\tilde{N}_0 = p_0$. The generating function of the jump process is again expressed in terms of the expectation of some functional of the diffusion process $(\theta_t)_{t \geq 0}$ defined by (6.7):

$$\tilde{\mathbb{E}}_{p_0}\left[z^{\tilde{N}_t} \right] = \mathbb{E}\left[\left(1 - \tanh\left(\frac{\theta_t}{2} \right)^2 \right) \tanh\left(\frac{\theta_t}{2} \right)^{2p_0} \ \middle|\ \theta_0 = 2\ \mathrm{argtanh}(\sqrt{z}) \right].$$

It should be noted also that $\tilde{W}_0$ is not a density with respect to the Lebesgue measure over $\mathbb{R}^+$ (while W_1 is a density, as seen in section 6.2). It consists actually of the sum of a Dirac mass at 0 and a density:

$$\tilde{W}_0(\omega, d\tau, 0) = p_{\omega,d}\delta_0(d\tau) + \tilde{W}_{0,c}(\omega, d\tau, 0).$$

This expression is obtained by disintegrating the right-hand side of (7.3) over the first jump time of the process $(\tilde{N})_{t\geq 0}$. The weight of the Dirac mass is

$$p_{\omega,d} = \exp\left(-\frac{4L}{l_\beta(\omega)}\right),$$

while the absolutely continuous part is given by

$$\int_0^{\tau_1} \tilde{W}_{0,c}(\omega,\tau,0)d\tau = \int_0^L \frac{4}{l_\beta(\omega)} e^{\frac{-4(L-t)}{l_\beta(\omega)}} \tilde{\mathbb{P}}_1\left(\int_0^t 2\tilde{N}_s ds \in [0,\tau_1],\ \tilde{N}_t = 0\right) dt.$$

It seems impossible to derive a closed form expression for the density part. We can either derive expansions or perform numerical simulations based on Monte-Carlo simulations of the random jump process $(\tilde{N}_t)_{t\geq 0}$. For instance, we can expand $\tilde{W}_{0,c}$ for small τ. Indeed, if $\tau_1 \ll l_\beta(\omega)$, then

$$\tilde{\mathbb{P}}_1\left(\int_0^t 2\tilde{N}_s ds \in [0,\tau_1],\ \tilde{N}_t = 0\right) \simeq \frac{2\tau_1}{l_\beta(\omega)} \exp\left(-\frac{4t}{l_\beta(\omega)}\right),$$

so that

$$\tilde{W}_{0,c}(\omega,\tau,0) \overset{\tau \ll l_\beta(\omega)}{\simeq} \exp\left(-\frac{4L}{l_\beta(\omega)}\right) \frac{8L}{l_\beta(\omega)^2}.$$

This approximate expression will be used in the next section to give a closed form expression of the refocused pulse in a particular regime.

7.3. The refocused pulse. The following theorem expresses the self-averaging property of the refocused pulse.

THEOREM 7.1. *For any $T > 0$, $\delta > 0$,*

$$\mathbb{P}\left(\sup_{t\in[-T,T]} \left|\eta^\varepsilon_{tr(TRT)}\left(\frac{t_1+L}{\varepsilon^2} + t, z = 0\right) - \eta_{tr(TRT)}(t)\right| > \delta\right) \overset{\varepsilon\to 0}{\longrightarrow} 0,$$

where $\eta_{tr(TRT)}$ is the refocused pulse shape:

$$\eta_{tr(TRT)}(t) = \left(f(-\,\cdot) * K_{TRT}(\cdot)\right)(t). \tag{7.4}$$

The Fourier transform of the kernel is the convolution of the time-inverted cut-off function G_{t_0,t_1} with the density $\tau \mapsto \Lambda^L_{tr}(\omega,\tau)$ evaluated at $(1-k'(\omega))L$:

$$\begin{aligned}\hat{K}_{TRT}(\omega) &= \left(G_{t_0,t_1}(-\,\cdot) * \Lambda^L_{tr}(\omega,\cdot)\right)((1-k'(\omega))L) \\ &= \int G_{t_0,t_1}(\tau - (1-k'(\omega))L)\Lambda^L_{tr}(\omega,d\tau).\end{aligned} \tag{7.5}$$

Proof. The proof follows the same lines as that of Theorem 6.1 with the transport equations corresponding to the transmission problem. ∎

Homogeneous dispersive case. Assume here that randomness is absent ($\alpha_\beta(\omega) \equiv 0$). Then $\Lambda^L_{tr}(\omega,\tau) = \delta_0(\tau)$, so that

$$\hat{K}_{TRT}(\omega) = G_{t_0,t_1}((k'(\omega)-1)L),$$

which is consistent with the results of section 5.4 at $z = 0$.

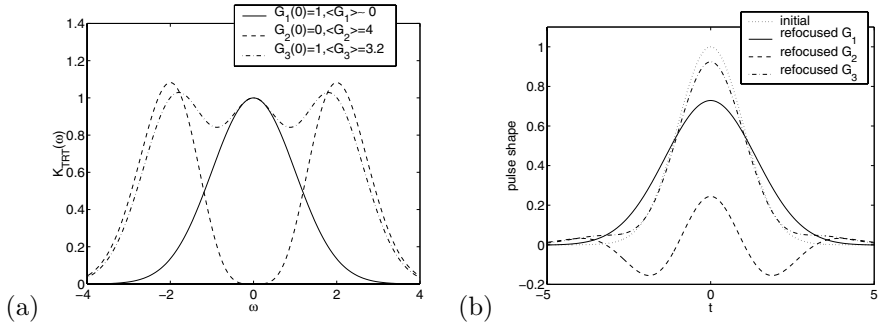


FIG. 7.1. *Fourier transform of the convolution kernel* K_{TRT} (a) *and refocused pulse* (b). *We consider the three cut-off functions* G_1, G_2, G_3 *described within the text. Here we assume* $\alpha(\omega) \equiv 1$, $L = 1$. *The initial pulse has Gaussian shape* $f(t) = \exp(-t^2/2)$.

Random nondispersive case. Assume here that $\beta = 0$. Consider an input pulse f which is such that the power spectral density of the process m can be considered as constant over the spectral range $[-\omega_{max}, \omega_{max}]$ of f: $\alpha(\omega) \equiv \alpha_0$. Finally, assume that we record a small piece of the transmitted wave in the sense that the cut-off function G_{t_0,t_1} has its support in $[t_0, t_1]$ such that $t_0 < 0$ and $t_1 > 0$ with $\alpha_0 \omega_{max}^2 t_1 \ll 1$. Then

$$\hat{K}_{TRT}(\omega) = e^{-\frac{\alpha_0 \omega^2 L}{2}} \left(G_{t_0,t_1}(0) + \frac{\alpha_0^2 \omega^4 L}{8} \langle G_{t_0,t_1} \rangle \right),$$

where $\langle G_{t_0,t_1} \rangle = \int_0^\infty G_{t_0,t_1}(t)dt$, so that

$$\eta_{tr(TRT)}(t) = G_{t_0,t_1}(0) \left(f(-\,\cdot) * K_{TRT,1}(\cdot) \right)(t) + \langle G_{t_0,t_1} \rangle \left(f(-\,\cdot) * K_{TRT,2}(\cdot) \right)(t),$$

$$\hat{K}_{TRT,1}(\omega) = e^{-\frac{\alpha_0 \omega^2 L}{2}},$$

$$\hat{K}_{TRT,2}(\omega) = \frac{\alpha_0^2 \omega^4 L}{8} e^{-\frac{\alpha_0 \omega^2 L}{2}}.$$

The convolution kernel $K_{TRT,1}$ results from the double action of the O'Doherty–Anstey theory on the front pulse in forward and backward directions. Of course this contribution completely vanishes if we do not record the front of the pulse ($G_{t_0,t_1}(0) = 0$). The convolution kernel $K_{TRT,2}$ is a filter that retains only the frequencies around $1/\sqrt{\alpha_0 L}$, those which can probe the medium without being completely reflected by the strong localization effect.

7.4. Numerical illustrations. We would like to illustrate results obtained in the previous section. We consider the hyperbolic random case discussed in subsection 7.3. In Figure 7.1 we plot the Fourier transform of the kernel K_{TRT} for three different time reversal calculations corresponding to three different cut-off functions G_{t_0,t_1} that we shall denote by G_1, G_2, and G_3. In the first calculation, we record only the front pulse and send it back into the medium $G_1(0) = 1$, $\langle G_1 \rangle \ll L$. We may think, for instance, that

$$G_1(t) = \cos^2\left(\frac{t}{t_1}\right) \mathbf{1}_{[-\pi t_1/2, \pi t_1/2]}(t),$$

with $\alpha_0 \omega_{max}^2 t_1 \ll 1$ and $t_1 \ll L$. The refocused pulse results from the action of the O'Doherty–Anstey theory, and its shape is the convolution of the initial pulse shape with the kernel $K_{TRT,1}$ (solid line, Figure 7.1).

In the second calculation, we record a small piece of the coda but not the front pulse $G_2(0) = 0$, $\langle G_2 \rangle = 4L$. We may think, for instance, that

$$G_2(t) = \frac{8L}{\pi t_1} \sin^2\left(\frac{t}{t_1}\right) \mathbf{1}_{[0,\pi t_1]}(t),$$

with $\alpha_0 \omega_{max}^2 t_1 \simeq 0.1 - 0.2$. Only medium range frequencies have been recorded, so that the refocused pulse shape is the convolution of the initial pulse shape with the kernel $K_{TRT,2}$ (dashed line, Figure 7.1).

In the third calculation, we record both the front pulse and a piece of the coda, so that $G_3(0) = 1$, $\langle G_3 \rangle \simeq 3.2L$. We may think, for instance, that

$$G_3(t) = \frac{6.4L}{\pi t_1} \sin^2\left(\frac{t + t_0}{t_1}\right) \mathbf{1}_{[-\pi t_0, \pi(t_1 - t_0)]}(t),$$

with $\alpha_0 \omega_{max}^2 t_1 \simeq 0.1 - 0.2$ and $t_0 \simeq \sqrt{\pi t_1^3/(6.4L)}$. In such a case a broad range of frequencies are recorded, and the cut-off function has been chosen in such a way that the weighted sum of the two kernels $K_{TRT,1}$ and $K_{TRT,2}$ define a kernel K_{TRT} with a large band of frequencies (dash-dotted line, Figure 7.1). As a result, the refocused pulse is close to the initial pulse.

It has been observed experimentally [10] that retransmitting part of the coda produces better refocusing than resending the front. This observation addresses spatial refocusing, while in this paper we focus our attention on time refocusing. The above illustrations show that the contributions of the coda and the front to the refocused pulse are actually complimentary. The contribution of the front is concerned with the low-frequency components of the pulse, while the contribution of the coda is concerned with the high-frequency components of the pulse. If we extrapolate this observation to three-dimensional configurations, then we can understand the quoted experimental observation in the sense that the high-frequency components are the ones that are expected to give the precise location of the source point.

7.5. TRT numerical experiments and application to source reconstruction. In this section we further illustrate TRT. A numerical method for the nonlinear terrain-following Boussinesq equation has been fully described in [19]. In this section we describe time reversal experiments in transmission by performing numerical experiments corresponding to the linearized terrain-following Boussinesq system (2.1)–(2.2). The initial wave elevation profile is incoming from the left and is given either by a Gaussian $\eta_0(z) = f(z) = \exp(-z^2/0.05)$ or by its spatial derivative $f'(z)$, as displayed in Figure 7.2. The corresponding initial velocity field is calculated in order to generate only a right-propagating mode, as presented in section 3. This is easily done by performing the inverse FFT of $\check{u} \equiv (\omega/k)\check{f}$.

7.5.1. Previous numerical results in related configurations. In a previous article by Fouque and Nachbin [15], TRR numerical experiments were conducted with a weakly nonlinear shallow water system. In particular, formula (6.10) (for the refocused pulse shape in reflection) was numerically captured in the hyperbolic ($\beta = 0$) case. The corresponding formula in [15] reads as (6.10) with K_{TRR}, as given by expression (6.13), with $\kappa_{(\beta=0)}(\omega) = \alpha(\omega)/4$. A weakly nonlinear example was also presented showing that formula (6.10) with $\beta = 0$ still holds as a good approximation. As a consequence of these early results, a complete nonlinear hyperbolic theory has been recently developed by the present authors [13]. Subsequently [14, 19] nonlinear

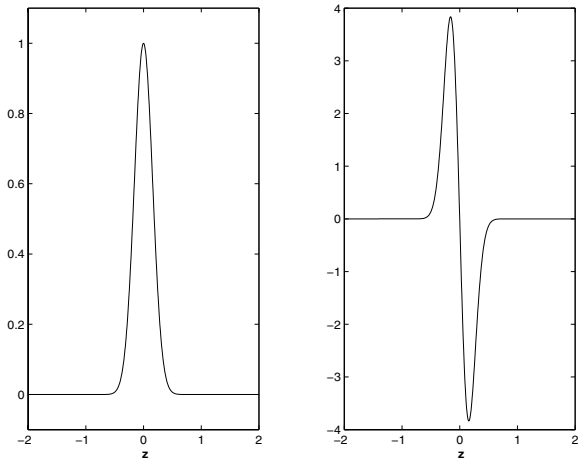


FIG. 7.2. *Initial wave elevation profiles considered in the experiments.*

experiments were further extended, and several numerical experiments for TRR were presented for both linear and weakly nonlinear dispersive waves, including solitary waves. Theory is not yet available for weakly dispersive, weakly nonlinear (solitary) waves.

Connecting the above comments and previous results with the present paper, we recall that numerical results for the transmitted front (as at the end of section 5.2) were presented by Grajales and Nachbin in [18]. The present numerical code captured quantitatively the composition of both kernels K_d and K_r, defined through (5.6). One should keep in mind that the present stochastic O'Doherty–Anstey formulation in transmission is more general than the deterministic theory given in [18], in part because it does not necessarily rely on β being small and also because it displays the *self-averaging* property. Moreover, time reversal was not addressed in [18].

Hence it is important to note that the transmission formula (5.6) plays an essential role in expression (5.7), which converges asymptotically to the TRT formula (7.4). The limiting form (7.1) of the frequency autocorrelation function was studied in section 7.1 and is characterized by the solution $\tilde{W}_0$ of a transport equation, as indicated in (7.2). In contrast to the TRR problem it is not possible to derive a closed form expression for the power spectral density Λ^L_{tr}, as mentioned at the end of section 7.2. One can derive expansions or perform Monte-Carlo simulations with the corresponding jump process. On the other hand, in the TRR problem this was made possible with the large slab hypothesis (section 6.4), leading to a closed form expression for Λ^∞_{ref}. Thus extracting quantitative information from expression (7.4) is a complex task, particularly due to the difficulty of computing Λ^L_{tr}.

Our strategy for presenting numerical results that address the theoretical expression for TRT is as follows. In section 7.3 a dispersive regime was identified where TRT can be easily checked: the homogeneous dispersive case. It has never been verified that the oscillatory effect of the Airy kernel can be completely recompressed, even for large values of β where we end up completely losing track of the initial pulse shape. Phases are scrambled due to dispersion but recompressed (reorganized) through time reversal. This will be shown below.

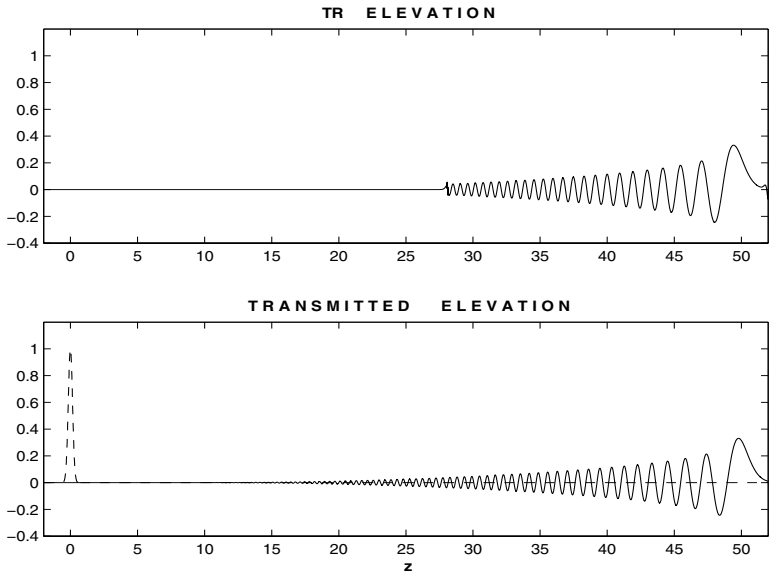


FIG. 7.3. *Bottom graph: the complete transmitted wave elevation. The initial condition is given by a dashed line. Dispersive propagation ($\beta = 0.01$) over a flat bottom. Top graph: cut-off wave elevation profile to be time reversed and sent back towards the origin.*

The next step is to add randomness. In forward transmission the dispersive O'Doherty–Anstey attenuation mechanism has been quantitatively validated in [18], for a specific realization. Note that these are two separate ways of addressing the two main mechanisms encoded in Λ_{tr}^{L}: the dispersive and the incoherent coda production and recompression. Finally, in the absence of a closed form expression for Λ_{tr}^{L}, we proceed to qualitatively verifying the combined effect for the dispersive TRT in a random environment.

7.5.2. New experiments for dispersive TRT. The new experiments of interest are in the TRT regime illustrating how, in particular, it can be applied to source reconstruction (i.e., waveform inversion).

We first consider the homogeneous dispersive case discussed above. In this problem a Gaussian pulse will be gradually transformed into an Airy function (cf. section 5.4, case (b)). An oscillatory tail develops behind the wavefront due to dispersion, as displayed in Figure 7.3. TRT will recompress the oscillatory tail, and the initial waveform is obtained as indicated in the sequence of Figure 7.4. In these experiments we used the Gaussian pulse (of approximately unit width) for the right-propagating initial elevation, together with its consistent (right-going) dispersive velocity profile ($\beta = 0.01$). Both the wave elevation η and the wave velocity u were recorded for time reversion. Hence no right-going mode was produced in the time reversed experiment.

Dispersion is then increased to a level where we will completely lose track of the initial pulse shape. Let $\beta = 0.1$, and consider the derivative of a Gaussian for the initial profile $\eta_0(z)$. In Figure 7.5 we present the forward experiment (in the top graph), having only a right-going mode. Time evolves from bottom trace to the top. The final trace (at the top) shows that we have completely lost track of the initial profile highlighted in the bottom trace. Both η and u are recorded and time reversed. Hence

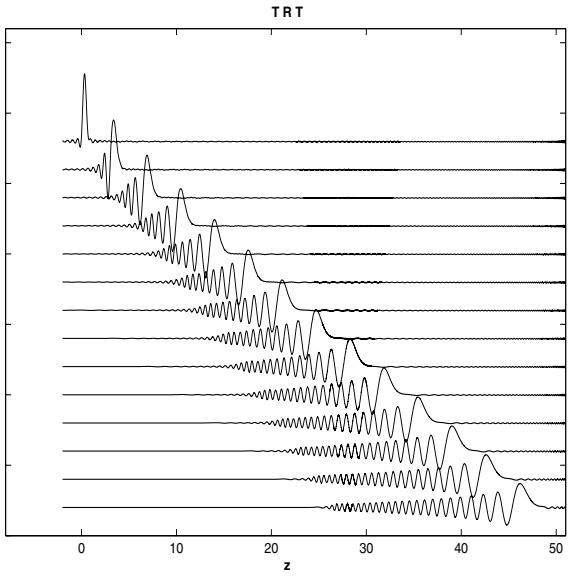


FIG. 7.4. *Time reversal in transmission (TRT) over a flat bottom. The initial profile was a Gaussian at the origin. The time reversed profile is the trace at the bottom. Time evolves from bottom to top at time increments of* 3.6 *units. Complete refocusing is observed in the top trace. The dispersion level is* $\beta = 0.01$.

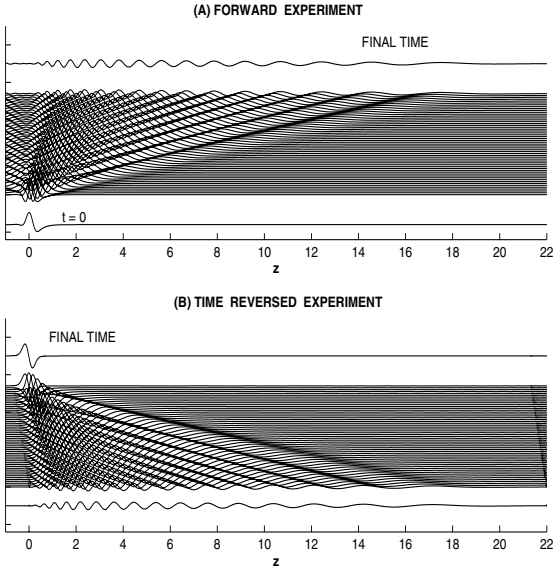


FIG. 7.5. *TRT over a flat bottom. The initial profile is a derivative of a Gaussian at the origin (bottom trace of graph* (A)*). Time evolves from bottom to top. Full recompression is observed in graph* (B)*. The dispersion level has been increased to* $\beta = 0.1$.

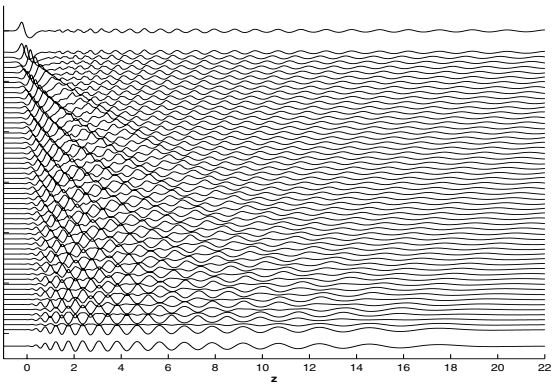


FIG. 7.6. *TRT over a flat bottom. Time evolves from bottom to top. The TR wave elevation (bottom trace) was amplified by a factor of two, while the velocity field u was not used for TR.*

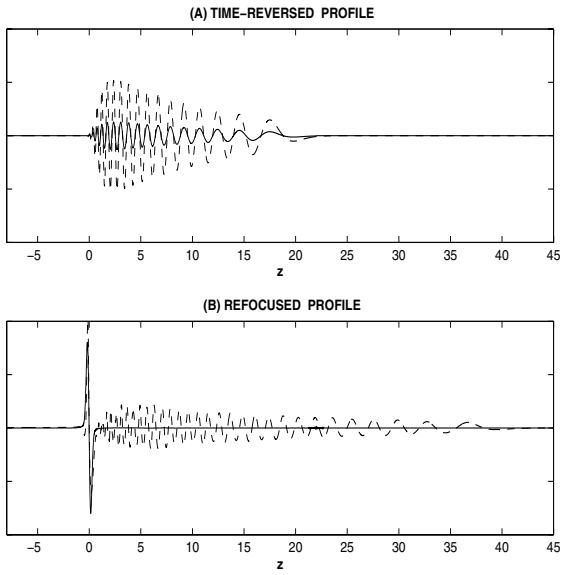


FIG. 7.7. (A) *Solid line: the recorded η; dashed line: the amplified η for time reversion.* (B) *Refocused pulse for the amplified experiment of Figure* 7.6 *(dashed line) and for the experiment in Figure* 7.5 *(solid line).*

a left-propagating mode is generated for the time reversed experiment. In the bottom graph of Figure 7.5 we clearly see the full recompression as predicted in section 5.4.

Next we consider the case where we record only the wave elevation η. For the time reversal experiment we re-emit this elevation field with a two-fold amplification. The corresponding dynamics is presented in Figure 7.6. We clearly see that, as recompression takes place along the left-propagating mode, there is a small dispersive wave propagating to the right. In Figure 7.7(A) we have the "doubled" time reversed profile compared to the recorded profile. In Figure 7.7(B) we see that the refocused pulse is the same for both experiments considered with $\beta = 0.1$. The oscillatory coda seen in

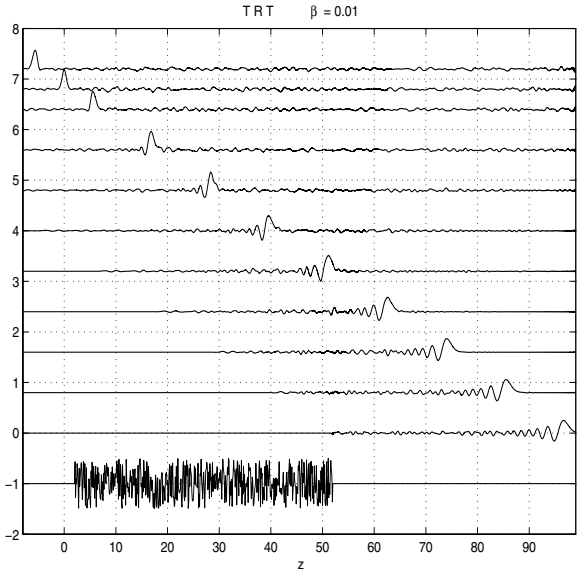


FIG. 7.8. *TRT over a random topography. The fluctuation level is* 50% *and the correlation length is* $\varepsilon = 0.1$. *The realization of the topography used for the simulation is given at the bottom. Just above the topography we have the wave elevation profile used for time reversion. Then time evolves from bottom to top in increments of* 11.52 *time units. The expected refocusing time takes place at half the time increment used and is therefore graphed accordingly.*

Figure 7.7(B) is due to the right-propagating mode in the "amplified" experiment.

We now repeat both TRT experiments in the presence of a random topography expressed through the coefficient $M(z)$. In these experiments both η and u are used in the time reversed data. In Figure 7.8 a realization of the random topography is given at the bottom of the graph, together with the transmitted wave elevation which will be time reversed and sent back into the random medium. Note that to the left of the (transmitted) oscillatory coda we have (small) incoherent radiation. At the correct time the deterministic front, coda, and random radiation recompress to give rise to (a reduced version of) the original waveform, namely a Gaussian. The correct time is exactly the time for the wave to reach the origin ($t = 97.92$). This was the time up to which the time reversed signal was originally recorded. Note, however, that the resolution of the source location is rather poor. The three upper curves in Figure 7.8 show almost the same waveform. This is of course an expected consequence of the hyperbolicity of the equations.

Our final illustration of TRT considers strong dispersive effects. In the previous example we had $\beta = 0.01$. Now we consider a value ten times larger. We now adopt the Gaussian's derivative as the initial wave elevation profile. This function has more energy on higher Fourier modes than the Gaussian. Thus the effect of dispersion will be even more noticeable. In Figure 7.9 we see the topography realization at the very bottom of the figure. Above the topography we find the transmitted wave elevation profile to be used in the time reversal experiment. This profile was recorded after 95.4 time units. The corresponding velocity profile is reversed. From bottom to top, the next three curves correspond to times $t = 91.8, 95.4$ (the expected refocusing time), and 99 time units. Only at time $t = 95.4$ do we have the original initial profile. At neigh-

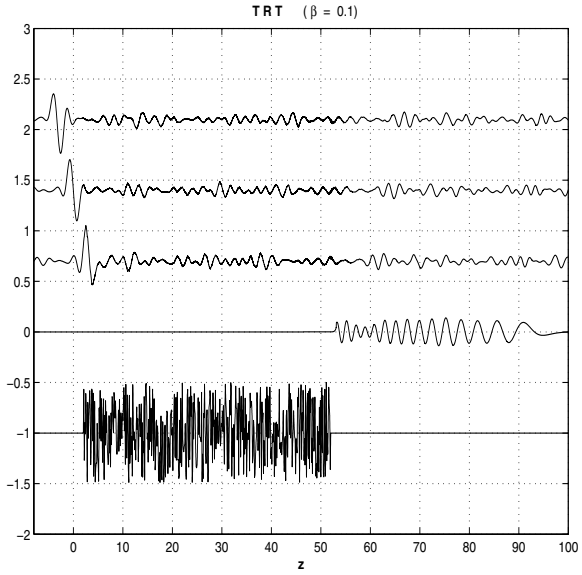


FIG. 7.9. *TRT over a random topography. The dispersion level has been increased* 10 *times* ($\beta = 0.1$). *The trace at the bottom represents the time reversed wave elevation profile. The three following curves (from bottom to top) correspond to times* $t = 91.8$, 95.4 *(the expected refocusing time), and* 99 *time units. The topography's fluctuation level is* 50%, *and the correlation length is* $\varepsilon = 0.1$.

boring times we see the effect of dispersion. At $t = 91.8$ we see that the oscillatory coda (here ahead of the left-propagating pulse) is still being recompressed. At time $t = 99$ the pulse starts developing the usual dispersive coda behind it. The wave source is located at the origin with a much higher accuracy than in the hyperbolic case. The source location is precisely the point between coda recompression and coda generation. Note that from the TR initial profile (at the bottom of Figure 7.9) it is very difficult to predict the source's waveform, while TRT has naturally performed the waveform inversion. We are currently working on the extension of these results and applications to higher dimensions. A good numerical model and bathymetric information can be invaluable tools for performing the time reversed dynamics and waveform inversion.

8. Conclusion. In this paper we have addressed the time reversal for waves governed by a random dispersive Boussinesq system. We have demonstrated that source location by time reversal is more effective in the dispersive case than in the hyperbolic case, because the source location is precisely the point between coda recompression and coda generation. Our analysis also shows that dispersion enhances localization effects in random medium. As a result, time reversal focusing in reflection (resp., in transmission) is more efficient (resp., less efficient) in the dispersive case than in the hyperbolic case, as indicated in Figure 6.1(a). Extension to more general dispersion relations is straightforward. The only but important hypothesis is that the addressed dispersion relation $k(\omega)$ should be an odd function so that it preserves time reversibility. These statements can also be generalized to some extent to three-dimensional configurations. In 3D configurations the pulse refocuses in time and in space [12, 6]. Accordingly, even in absence of dispersion source localization is possible, as it is given

by the point where the refocused wave reaches its climax. However, we conjecture that dispersion improves the resolution of the source location, as the pulse spreading is enhanced when propagating away from the original source location. Furthermore, the spectral phase modulations are larger in the presence of dispersion, so that only close wavenumbers are phase-matched. We can thus expect that dispersion enhances the statistical stability as well as the super-resolution in spatial refocusing described in [3, 6, 23], and in time refocusing as described in this paper.

REFERENCES

[1] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] M. Asch, W. Kohler, G. Papanicolaou, M. Postel, and B. White, *Frequency content of randomly scattered signals*, SIAM Rev., 33 (1991), pp. 519–625.

[3] G. Bal, G. Papanicolaou, and L. Ryzhik, *Self-averaging in time reversal for the parabolic wave equation*, Stoch. Dyn., 2 (2002), pp. 507–531.

[4] G. Bal, T. Tomorowski, and L. Ryzhik, *Self-averaging of Wigner transforms in random media*, Comm. Math. Phys., 242 (2003), pp. 81–135.

[5] G. Bal and L. Ryzhik, *Time reversal and refocusing in random media*, SIAM J. Appl. Math., 63 (2003), pp. 1475–1498.

[6] P. Blomgren, G. Papanicolaou, and H. Zhao, *Super-resolution in time-reversal acoustics*, J. Acoust. Soc. Am., 111 (2002), pp. 230–248.

[7] R. Burridge, G. Papanicolaou, and B. White, *Statistics for pulse reflection from a randomly layered medium*, SIAM J. Appl. Math., 47 (1987), pp. 146–168.

[8] J. F. Clouet and J. P. Fouque, *Spreading of a pulse traveling in random media*, Ann. Appl. Probab., 4 (1994), pp. 1083–1097.

[9] J. F. Clouet and J. P. Fouque, *A time-reversal method for an acoustical pulse propagating in randomly layered media*, Wave Motion, 25 (1997), pp. 361–368.

[10] A. Derode, P. Roux, and M. Fink, *Robust acoustic time reversal with high-order multiple scattering*, Phys. Rev. Lett., 75 (1995), pp. 4206–4209.

[11] M. Fink, *Time reversal mirrors*, J. Phys. D: Appl. Phys., 26 (1993), pp. 1333–1350.

[12] M. Fink, *Time reversed acoustics*, Scientific American, November (1999), pp. 91–97.

[13] J.-P. Fouque, J. Garnier, and A. Nachbin, *Shock structure due to stochastic forcing and the time reversal of nonlinear waves*, Phys. D., to appear.

[14] J.-P. Fouque, J. Garnier, J. C. Muñoz Grajales, and A. Nachbin, *Time reversing solitary waves*, Phys. Rev. Lett, 92 (2004), paper 094502.

[15] J.-P. Fouque and A. Nachbin, *Time-reversed refocusing of surface water waves*, Multiscale Model. Simul., 1 (2003), pp. 609–629.

[16] J.-P. Fouque and K. Solna, *Time-reversal aperture enhancement*, Multiscale Model. Simul., 1 (2003), pp. 239–259.

[17] P. Lewicki, R. Burridge, and G. Papanicolaou, *Pulse stabilization in a strongly heterogeneous medium*, Wave Motion, 20 (1994), pp. 177–195.

[18] J. C. Muñoz Grajales and A. Nachbin, *Dispersive wave attenuation due to orographic forcing*, SIAM J. Appl. Math., to appear.

[19] J. C. Muñoz Grajales and A. Nachbin, *Stiff microscale forcing and solitary wave refocusing*, Multiscale Model. Simul., to appear.

[20] A. Nachbin, *A terrain-following Boussinesq system*, SIAM J. Appl. Math., 63 (2003), pp. 905–922.

[21] G. C. Papanicolaou, *Wave propagation in a one-dimensional random medium*, SIAM J. Appl. Math., 21 (1971), pp. 13–18.

[22] G. Papanicolaou, *Asymptotic analysis of stochastic equations*, in MAA Stud. in Math. 18, M. Rosenblatt, ed., Mathematical Association of America, Washington, DC, 1978, pp. 111–179.

[23] G. Papanicolaou, L. Ryzhik, and K. Solna, *Statistical stability in time reversal*, SIAM J. Appl. Math., to appear.

[24] C. Pires and P. M. A. Miranda, *Tsunami waveform inversion by adjoint methods*, J. Geophys. Res., 106 (2001), pp. 19773–19796.

[25] K. Solna and G. Papanicolaou, *Ray theory for a locally layered medium*, Waves in Random Media, 10 (2000), pp. 151–198.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 1839–1863

ANALYSIS OF STRESS-DRIVEN GRAIN BOUNDARY DIFFUSION. PART I*

JON WILKENING†, LEN BORUCKI‡, AND J. A. SETHIAN§

Abstract. The stress-driven grain boundary diffusion problem is a continuum model of mass transport phenomena in microelectronic circuits due to high current densities (electromigration) and gradients in normal stress along grain boundaries. The model involves coupling many different equations and phenomena, and difficulties such as nonlocality, complex geometry, and singularities in the stress tensor have left open such mathematical questions as existence of solutions and compatibility of boundary conditions. In this paper and its companion, we address these issues and establish a firm mathematical foundation for this problem.

We use techniques from semigroup theory to prove that the problem is well posed and that the stress field relaxes to a steady state distribution which, in the nondegenerate case, balances the electromigration force along grain boundaries. Our analysis shows that while the role of electromigration is important, it is the interplay among grain growth, stress generation, and mass transport that is responsible for the diffusive nature of the problem. Electromigration acts as a passive driving force that determines the steady state stress distribution, but it is not responsible for the dynamics that drive the system to steady state.

We also show that stress singularities may develop near grain boundary junctions; however, stress components directly involved in the diffusion process remain finite for all time. Thus, we have identified a mechanism by which large "hidden" stresses may develop that are not directly involved in the diffusion process but may play a role in void nucleation and stress-induced damage.

Key words. grain boundary, diffusion, electromigration, elasticity, semigroups

AMS subject classifications. 35Q72, 47D03, 74F99

DOI. 10.1137/S0036139903438235

1. Introduction.

A microelectronic circuit consists of a silicon substrate with doped regions that function as circuit elements (transistors, diodes, resistors, and capacitors), metal lines and vias (interconnects) that connect the circuit elements together, intermetallic dielectric material that keeps the interconnects in place and insulated from each other, various oxide layers and diffusion barriers that are primarily needed in the manufacturing stage to control the doping process and keep the metal from diffusing into the silicon, and passivation to keep all the components in place and protected [28, 33].

A typical interconnect line might be an alloy of Al-0.5%Cu, have dimensions of $0.5 \times 0.5 \times 300$ microns, and carry a current density of 20 $m\text{A}/\mu m^2$. As electrons

*Received by the editors December 6, 2002; accepted for publication (in revised form) November 26, 2003; published electronically August 4, 2004.
http://www.siam.org/journals/siap/64-6/43823.html

†Courant Institute of Mathematical Sciences, New York, NY 10012 (wilken@cims.nyu.edu). The research of this author was supported in part by a Department of Energy Computational Science Graduate Student Fellowship while the author was at U.C. Berkeley; by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract DE-AC03-76SF00098; and by the National Science Foundation through grant DMS-0101439.

‡Motorola, Inc., Tempe, AZ 85284 (Len.Borucki@intelligentplanar.com). The research of this author was supported in part by the Division of Mathematical Sciences of the National Science Foundation, University-Industry Program.

§Department of Mathematics and Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94721 (sethian@math.berkeley.edu). The research of this author was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract DE-AC03-76SF00098 and by the Division of Mathematical Sciences of the National Science Foundation.

flow through the line, they are scattered by imperfections in the crystal lattice of the metal and impart momentum to the ion cores. This "electron wind" force is stronger than the opposing direct force of the electric field, so ions are transported in the same direction as the flow of electrons. This process is known as electromigration; it is a dominant failure mechanism in microelectronic devices.

Grain boundaries, void surfaces, and passivation interfaces are fast diffusion paths along which the diffusion constant typically is seven to eight orders of magnitude higher than in the grains; therefore, most of the mass transport occurs at these locations. The inhomogeneous redistribution of atoms leads to the development of stresses in the line. Stress gradients along grain boundaries and surface tension at void surfaces both contribute to the flux of atoms, usually opposing the electromigration term and increasing the lifetime of the line. Significant residual stresses left over from thermal contraction during the manufacturing process also affect the formation of voids and the transport of atoms.

As microelectronic circuits become smaller and current densities become higher, failure due to electromigration damage becomes an ever increasing problem in the design of circuits. Many theoretical models have been proposed to explain the role of various combinations of electromigration, stress gradients, diffusion, temperature, anisotropy, surface tension, and hillock formation on the mass transport of atoms in the bulk grains, along void surfaces, along grain boundaries, and at passivation interfaces. A useful reference written from the engineering perspective is the review article by Ho and Kwok [14]; see also [27]. The concept of the electron wind force was formulated by Fiks [9] and by Huntington and Grone [15]. In his experimental work in the 1970s, Blech [2] studied the behavior of thin films of aluminum and titanium-nickel when large currents were passed through them and demonstrated the existence of a threshold current density below which no damage occurs, which varies inversely with the stripe length. Shortly thereafter, Blech and Herring [3] offered the explanation that stress gradients were developing along grain boundaries in the sample to counter the electron wind force, but they could be sustained only up to a critical threshold. Once this threshold was reached, there was no physical mechanism to stop the transport of material, and the stripe eroded at one end and formed hillocks at the other.

Recent models of these phenomena were described by Mullins [21], Cocks and Gill [6], Korhonen et al. [18], Sarychev et al. [25], and Kirchheim [17]. The Mullins paper presents a nice overview of mass transport along surfaces and grain boundaries and discusses cobble creep and grain boundary grooving. The paper by Cocks and Gill gives a variational approach to the dynamics of grain boundary motion associated with decreasing grain boundary area; they did not include stress in their model. The papers [25] and [17] deal primarily with electromigration, stress-driven diffusion, and vacancy generation in the grains, while [18] focuses on electromigration and grain boundary diffusion. The latter three papers use a statistical argument about the orientation of the grain boundaries in order to model the stress as a scalar variable instead of a tensor; one should keep in mind, however, that for any particular sample, the grain boundaries have a specific geometry, and singularities can occur in the stress field that are ignored with this simplifying assumption.

Bower, Craft, Fridline, and collaborators use an advancing front algorithm to generate a sequence of adaptive, evolving finite element meshes to study grain growth, void evolution, hillock formation, and grain boundary sliding for possibly anisotropic materials responding to stress, surface tension, thermal expansion, and electromi-

gration; see, e.g., [4, 10]. They use interesting semi-implicit techniques to overcome timestep limitations due to the stiffness of the equations, and they use Lagrange multipliers to determine the normal stress along grain boundaries.

An alternative approach based on the theory described in this paper is presented in [30, 26], where a new singularity capturing the least squares finite element method is developed to study the effect of singularities in the stress field on the grain boundary diffusion process. Further references to the literature on numerical methods for grain growth and void evolution may be found there.

Mathematical analysis of the partial differential equations involved is largely absent in the electromigration literature. There are several reasons for this. First, there is no universal agreement in the electromigration community on exactly how all the phenomena fit together, especially at junctions where grain boundaries meet voids or other grain boundaries, and the process of void nucleation is far from understood. Second, the problem is very complicated, with many different (stiff) phenomena coupled together in a nonlocal, nonlinear way. Growth rates depend on taking derivatives of stress components along grain boundaries and curvatures along surfaces. Boundary conditions specify the gradient of the normal stress at junctions where the stress field is singular. Many of the equations couple the displacement field to the stress field, and it is difficult to visualize how this constrains the evolution of the system. Both displacement and flux boundary conditions are specified at junctions where grain boundaries meet the outer walls; in simpler problems such as the heat equation, this would overspecify the boundary conditions. As a result of these and similar difficulties, one typically has a long list of equations reflecting various principles such as mass conservation and chemical potential continuity that one would like to use as a model. But occasionally, incomplete physical reasoning can lead to mathematically ill-posed problems; therefore, it is important to develop a rigorous justification for the collection of equations to ensure a self-consistent model.

The goal of this paper is to provide a mathematical framework in which we can analyze a modest subset of the phenomena mentioned above. We assume there are no voids in the structure, and we work within the framework of linear elasticity (small strain and small grain growth). This may be thought of as the linearization of a nonlinear grain boundary migration theory. Most of the difficult problems mentioned above persist in this setting. The equations are nonlocal and couple together many different stiff phenomena that relate rates of change of displacement jumps to spatial derivatives of the normal stress. The boundary conditions specify the gradient of the normal stress at locations where the stress tensor becomes singular. And the geometry of the problem involves the complicated branching structure of a grain boundary network which does not have a natural ordering or orientation of its segments. The same approach is taken in [30].

In section 2, we exhibit the equations in dimensionless form and briefly describe the physical considerations that lead to these equations. Our main contribution here is to model the net grain growth along the grain boundary Γ as the jump in a normal component of the displacement across Γ, thinking of it as a scalar function g defined on Γ. This is identical to what Bower and Craft did in [4], except that they viewed each side of the grain boundary as a moving interface (in parallel with their treatment of void surfaces) and did not single out g as important. They derived an equation for the change Δg (denoted Δu_n in their paper) in a timestep but used it only to update the displacements $\mathbf{u}$ on each side of the grain boundary. The advantage of treating g as a time evolving *function* defined on the grain boundary (which is

fixed in the reference configuration) is that we are able to recast the problem as an ordinary differential equation on a Hilbert space and to apply techniques of semigroup theory to prove the equations are well posed. The difficulties due to nonlocality, the existence of singularities, and the complicated nature of the boundary conditions are all absorbed into two unbounded operators L and S, which turn out to possess many nice properties, such as self-adjointness, discrete spectra, and positivity (or negativity). An overview of this procedure is given in section 3, and full details are presented in section 6.

In section 4, we define the grain boundary normal stress problem and the operator S, which is a type of Dirichlet-to-Neumann map that maps the displacement jump g across Γ to the normal stress $\eta = \mathbf{n} \cdot \sigma \mathbf{n}$ on Γ. We state the important properties of S (which are proved in the companion paper [31]) and identify a new class of grain boundaries, which we call *degenerate*. Throughout this paper, we assume the grain boundary network is nondegenerate in order to simplify the presentation. The degenerate case is dealt with in [31].

In section 5, we analyze the operator L $(= -\frac{\partial^2}{\partial s^2})$ on the grain boundary network. The most important properties of this operator are that it is positive and self-adjoint, its domain $\mathcal{D}(L)$ consists only of functions that satisfy continuity and flux boundary conditions (useful for proving that the normal stress η has these properties), and the domain of $L^{\frac{1}{2}}$ is precisely $H^1(\Gamma)$ (useful in specifying the Hilbert space in which η evolves). The boundary conditions of chemical potential continuity and flux balance at junctions turn out to be exactly what are needed for these results to hold.

In section 6, we show that the equation governing the evolution of normal stress (namely, $\eta_t = SL\eta$) generates an analytic semigroup $\{E_t : t \geq 0\}$ of bounded linear operators on $H^1(\Gamma)$. The primary difficulties that arise have to do with the fact that L (and in the degenerate case, S) has a nontrivial finite dimensional kernel which must be dealt with using projections. We also discuss the role of electromigration as a passive driving force, the enforcement of boundary conditions, and the development of stress singularities near corners and grain boundary junctions.

Finally, in the Appendix, we study an infinite interconnect line with a single grain boundary running through its center, which provides insight into the nature of the diffusion process without the complication of boundary conditions and singularities.

2. Problem statement. In this section, we describe a two dimensional continuum model of electromigration and stress-driven grain boundary diffusion in the linear regime of small strain, small grain growth elasticity. The reference configuration (including the location of the grain boundary in the reference configuration) remains fixed, while the stress and displacement fields defined on this domain evolve in time.

A grain is a region where the atoms are aligned in a regular lattice. A grain boundary is an interface between two grains where the lattice structure becomes disorganized as the lattice alignment changes from one side to the other. In our (continuum) model, we ignore details of lattice alignment and assume all grain boundaries have equivalent properties.

The grain boundaries are assumed to be fast diffusion paths along which atoms are transported much more easily than in the bulk grains. At each point on the grain boundary, we have a flux J of atoms traveling along the grain boundary. J has units of surface flux ($\mathrm{cm}^{-1}\mathrm{s}^{-1}$), where we consider our two dimensional domain to have a thickness δ in the third dimension. If a portion of the grain boundary has more atoms flowing into it than out, the atoms incorporate themselves into the lattice of the adjacent grains and cause the grains to move apart to make room for the new

atoms. At the same time there will be a net flux of atoms out of other regions of the grain boundary, where atoms are removed from the lattice of each grain and the grains move together so as not to leave a gap.

Although our analysis of this problem does not rely on the grain boundaries being straight, we have omitted curvature-driven grain boundary motion from our model. For this reason, we assume that the grain boundaries are initially straight and that as grain growth occurs, the appropriate fraction of atoms attaches to each side of the grain boundary so that Γ remains fixed in the spatial (stressed) configuration. We adopt an Eulerian viewpoint where the reference configuration on which the stress and displacement fields are defined is the spatial configuration, and the natural (unstressed) shape of each grain changes in time as material is added to (or removed from) its boundary. For a given deformation φ mapping the natural state to the stressed state, the displacement is defined as $\mathbf{u}(x) = x - \varphi^{-1}(x)$ instead of $\varphi(x) - x$. The linearized equations of elasticity [20, 5] are the same in the material and Eulerian viewpoints.

We assume the interconnect line consists of several disjoint bounded polygonal grains Ω_k, and we denote their union (an open set) by $\Omega = \bigcup_k \Omega_k$. $\overline{\Omega}$ is assumed to be connected; see Figure 2.1. We denote the outer boundary (the "walls") of the domain by $\Gamma_0 = \partial(\overline{\Omega})$ and the grain boundary network by $\Gamma = (\partial\Omega \setminus \Gamma_0)^-$. Γ consists of N closed line segments $\Gamma = \bigcup_{j=1}^{N} \Gamma_j$. Each segment is given an arbitrary orientation (a unit tangent vector $\mathbf{t}_j$) and an arclength parameter s which increases in the $\mathbf{t}_j$ direction. The unit normal $\mathbf{n}_j$ points from right to left facing along $\mathbf{t}_j$. We do not impose the Young condition, requiring that grain boundaries meet at 120° angles, since it is not required for well posedness unless curvature is included as a driving force.

The net grain growth g is defined on Γ as the jump in normal component of displacement across the grain boundary:

$$g(x) := [\mathbf{u}(x^+) - \mathbf{u}(x^-)] \cdot \mathbf{n}_j \qquad (x \in \Gamma_j). \tag{2.1}$$

It represents the distance the original grains have separated to accommodate the new material that occupies that space; see Figure 2.2. In the Eulerian picture, $g(x) = [\varphi^{-1}(x^-) - \varphi^{-1}(x^+)] \cdot \mathbf{n}$ is the amount that opposite sides of the grain boundary at x would overlap if the grains were allowed to pass through each other to achieve their stress-free shapes. This overlap corresponds to new material added during the diffusion process.

In Figure 2.3, we list the equations and boundary conditions in nondimensional form. We choose an arbitrary length scale L ($\sim 1\mu$m) and define the timescale $t_0 = \frac{kTL^3}{\nu_b D_b \Omega_a^2 \mu}$, where k is the Boltzmann constant, T is temperature, D_b is the diffusion constant for grain boundary surface diffusion at temperature T, ν_b is the number of participating atoms per unit area, Ω_a is the volume of an atom in the atomic lattice, and μ is the shear modulus. See [21] for typical values of these parameters. We then define the dimensionless variables

$$\tilde{x} = \frac{x}{L}, \quad \tilde{t} = \frac{t}{t_0}, \quad \tilde{\mathbf{u}} = \frac{\mathbf{u}}{L}, \quad \tilde{\sigma} = \frac{\sigma}{\mu}, \quad \tilde{\psi} = \frac{|Z^*|e}{\Omega_a \mu}\psi, \quad \tilde{J} = \frac{\Omega_a t_0}{L^2} J, \quad \text{etc.}, \tag{2.2}$$

and rewrite the equations (see [4, 30]) in terms of these variables (dropping the tildes). Z^*e is a phenomenological effective charge for an ion in the lattice; $e = |e|$ is the elementary electric charge; and for a good conductor [21, 29], $Z^* \approx -5$. This means the electron wind force is stronger than the opposing direct force of the electric field.

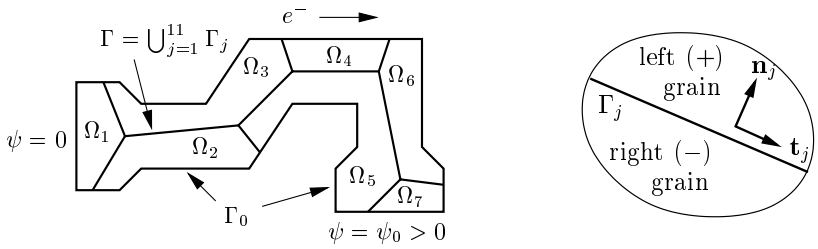


FIG. 2.1. *Left: geometry of an interconnect line. Right: arbitrarily assigned orientation of grain boundary segment determines tangential and normal directions, left and right grain labels, etc. The normal vector* $\mathbf{n}_j$ *points from right to left when facing along the* $\mathbf{t}_j$ *direction.*

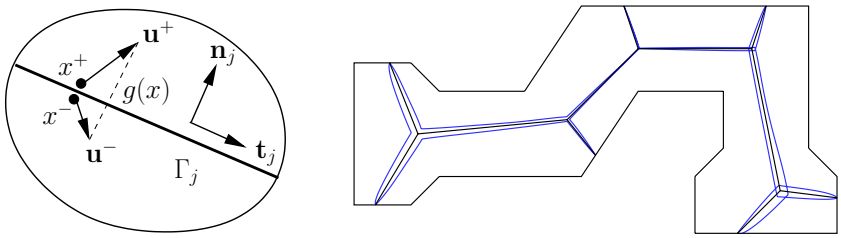
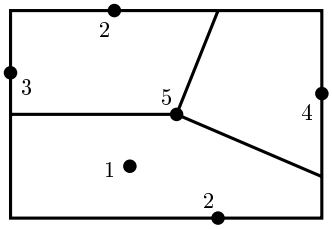


FIG. 2.2. *Left:* $g(x)$ *is the jump in normal component of displacement across* Γ *at* x*. The sign of* g *is independent of the orientation chosen for the segment. Right: exaggerated view of the natural state of each grain obtained by plotting* $x - C\mathbf{u}(x^{\pm})$*,* $x \in \Gamma$*, with a suitable* $C > 0$*. Grains must be zipped together (*$g < 0$*) on the left and pushed apart (*$g > 0$*) on the right in order to fit together.*

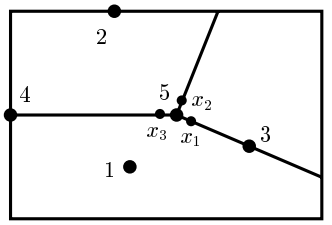
Electric Potential

1. $\nabla^2 \psi = 0$
2. $\partial_n \psi = 0$
3. $\psi = 0$
4. $\psi = \psi_0$
5. Grain boundaries are invisible to ψ



Elasticity and Grain Growth

1a. $\mu \Delta \mathbf{u} + (\lambda + \mu) \nabla (\nabla \cdot \mathbf{u}) = 0$
2a. $\mathbf{u} = 0$
3a. $\mathbf{u}(x^+) - \mathbf{u}(x^-) = g(x)\mathbf{n}$
3b. $\sigma(x^+) = \sigma(x^-)$
3c. $\mathbf{n} \cdot \sigma(x)\mathbf{n} = \eta(x)$
3d. $\partial_t g = -\partial_s^2(\eta + \psi)$ $\begin{cases} J = \partial_s(\eta + \psi) \\ g_t + J_s = 0 \end{cases}$
4a. $g = 0$
4b. $\partial_s(\eta + \psi) = 0$
5a. $g(x_1)\mathbf{n}_1 + g(x_2)\mathbf{n}_2 + g(x_3)\mathbf{n}_3 = 0$
5b. $\eta(x_1) = \eta(x_2) = \eta(x_3)$
5c. $\partial_{s_1}(\eta + \psi) + \partial_{s_2}(\eta + \psi) + \partial_{s_3}(\eta + \psi) = 0$



Initial Condition: $g \equiv 0$

FIG. 2.3. *Summary of equations and boundary conditions. Segments are assumed (in this figure only) to be parameterized away from the triple junction to avoid minus signs.*

The electric potential ψ is found by solving the Laplace equation under the assumption that the grain boundaries do not significantly affect the flow of current in the line. The displacement field $\mathbf{u}$ is found by solving the Lamé equations of linearized elasticity (assuming plane strain). The stress tensor satisfies Hooke's law,

$$\sigma = 2\mu\epsilon + \lambda \operatorname{tr}(\epsilon) I, \tag{2.3}$$

where λ and μ are the Lamé coefficients, $\epsilon_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i)$ are the components of the strain tensor, and $\operatorname{tr}(\cdot)$ is the trace operator; see [20, 5].

Referring to Figure 2.3, 2a enforces the requirement that the displacement is zero at the outer walls (passivation). In 3a, we assume that the grains do not slide tangentially relative to each other, and we define the displacement jump g. In 3b, we enforce the local balance of forces (tractions) across the grain boundary, which together with the no-sliding assumption implies that all components of the stress tensor are continuous across grain boundaries. In 3c, we define the normal stress η on the grain boundary, which is well defined by 3b.

Equation 3d in Figure 2.3 is the main evolution equation, which gives the grain growth rate in terms of the normal stress and the electrostatic potential. This equation is a consequence of the continuity equation, the Einstein–Nernst equation, the Blech–Herring model of the chemical potential of an atom in a grain boundary of a stressed solid [3, 30], and electromigration:

$$\partial_t g + \Omega_a \partial_s J = 0 \qquad \text{(continuity equation)}, \tag{2.4}$$

$$J = -\tfrac{\nu_b D_b}{kT} \partial_s \mu_b \qquad \text{(Einstein–Nernst, } \mu_b = \text{chemical potential)}, \tag{2.5}$$

$$\mu_b = \mu_0 - \Omega_a \sigma_{nn} \qquad \text{(Blech–Herring, } \mu_0 = \text{const)}, \tag{2.6}$$

$$\partial_s \mu_b \to \partial_s \mu_b + Z^* e\, \partial_s \psi \qquad \text{(electron wind force, } Z^* e < 0), \tag{2.7}$$

$$J = \tfrac{\nu_b D_b}{kT} (\Omega_a \partial_s \sigma_{nn} + |Z^* e|\, \partial_s \psi) \quad \text{(flux before nondimensionalizing)}. \tag{2.8}$$

Note that qualitatively, atoms are transported from regions of compression to regions of tension and travel against the electric field $\mathbf{E} = -\nabla\psi$ in the same direction that electrons flow. Here, $\partial_s \psi$ is the derivative of $\psi\big|_\Gamma$ with respect to arc length, so $-\partial_s \psi = \mathbf{E} \cdot \mathbf{t}$ is the component of the electric field along the grain boundary.

Equation 4a in Figure 2.3 follows from 2a and 3a but is worth recording as a boundary condition on g. Equation 4b enforces zero flux at gb-wall junctions: atoms are not allowed to flow in or out of the network where the grain boundary meets passivation, so global mass conservation should hold. Equation 5a is a compatibility requirement following from 3a: if we start in one grain and follow the jump in displacement around a triple junction, we have to end up with the original displacement when we return. (The point x_i here is infinitesimally close to the triple junction on segment i.) Finally, equations 5b and 5c enforce chemical potential continuity and flux balance at triple junctions, respectively.

3. Strategy. Thus far, each equation represents either a definition (of g or η) or some physical requirement such as chemical potential continuity or mass conservation. The next task is to find a way to organize them so that mathematical questions such as well posedness can be addressed. One major challenge is to identify the role played by singularities in the stress field near junctions and to understand the sense in which 4b, 5b, and 5c of Figure 2.3 can be expected to hold in light of these singularities. Another goal is to find a way to untangle the equations and boundary conditions in order to handle the nonlocal nature of expressions relating the displacement jump g

to the normal stress η; placing local constraints on one imposes (rather awkward) global constraints on the other via the Lamé equations. It is not immediately obvious that the evolution 3d is compatible with conditions 4a–5c.

Our approach is to recast the problem as an ordinary differential equation on a Hilbert space, writing the equation in terms of the normal stress η and absorbing all the boundary conditions into the operators. In this way we take advantage of linearity and gain insight into the role played by each of the boundary conditions in the well posedness of the problem.

3.1. A type of Dirichlet-to-Neumann map. Equations 1a–3c in Figure 2.3 can be thought of as providing a mapping between the jump in displacement g and the normal stress η. If we are given one (in some appropriate space), the other can be determined by solving the elasticity equations. There is a duality between g and η embodied in the energy relation

$$E = -\frac{1}{2}\int_\Gamma \eta g \, ds, \tag{3.1}$$

relating the elastic energy stored in the grains to the work done at the grain boundaries to accommodate the accumulation or depletion of atoms there. See the companion paper [31] for further details.

We denote the operator that maps a given grain growth function g on Γ to the corresponding normal stress η by

$$S : \mathcal{D}(S) \to L^2(\Gamma) : g \mapsto \eta \qquad \text{(grain growth to normal stress map)}. \tag{3.2}$$

Although S is unbounded, it turns out to be self-adjoint and negative (essentially due to (3.1)), and its domain is dense in $L^2(\Gamma)$. Moreover, S has a compact pseudoinverse B such that SB is the identity (nondegenerate case) or differs from the identity by a finite rank projection (degenerate case). These properties are discussed in section 4 and proved in [31].

3.2. The second derivative operator. We define the operator $L : \mathcal{D}(L) \to L^2(\Gamma)$ to be the negative of the second derivative operator with respect to arc length on each grain boundary segment. If η is twice continuously differentiable on each Γ_j and satisfies certain boundary conditions at the junctions, then $\eta \in \mathcal{D}(L)$ and the restriction of $L\eta$ to the interior of Γ_j is given by

$$L\eta(x) = -\frac{\partial^2 \eta}{\partial s^2} \qquad (x \in \Gamma_j^o). \tag{3.3}$$

In section 5 we will show that boundary conditions 4b, 5b, and 5c from Figure 2.3 enforcing mass conservation and chemical potential continuity are exactly what are needed for some of the most useful properties of L on the unit interval to carry over to the more complicated branching structure of a grain boundary network. In particular, L is self-adjoint, positive, and densely defined, and its kernel is finite dimensional. Moreover, if $L^{\frac{1}{2}}$ is modified by a finite rank projection to remove its kernel, it becomes an isomorphism from $H^1(\Gamma)$ onto $L^2(\Gamma)$, which is important in our proof of well posedness.

3.3. An ordinary differential equation on a Hilbert space. The evolution of the jump in displacement g is governed by equation 3d of Figure 2.3, namely,

$$g_t = -\partial_s^2(\eta + \psi) = L(Sg + \psi). \tag{3.4}$$

Applying S to (3.4), we obtain

$$\eta_t = SL(\eta + \psi). \tag{3.5}$$

The term ψ is acting as a passive driving force in (3.4) and (3.5). In general, if the equation

$$\frac{dx}{dt} = Ax, \qquad x(0) = x_0, \tag{3.6}$$

generates a strongly continuous semigroup $\{E_t : t \geq 0\}$ of bounded linear operators on a Banach space X [13, 32, 16, 1], then for $f \in X$ the solution to the equation

$$\dot{x} = A(x + f), \qquad x(0) = x_0, \tag{3.7}$$

is given by

$$x(t) = E_t(x_0 + f) - f \qquad (t \geq 0). \tag{3.8}$$

In section 6, we will show that the equation

$$\eta_t = SL\eta, \qquad \eta(x, t = 0) = \eta_0(x), \tag{3.9}$$

generates such a semigroup in $H^1(\Gamma)$, so the evolution of the normal stress η with initial condition η_0 (usually taken to be zero) is given by

$$\eta(t) = E_t(\eta_0 + \psi) - \psi. \tag{3.10}$$

In the nondegenerate case, the evolution of g may be obtained directly from that of η via $g(t) = B\eta(t)$, but in the degenerate case this is not so. In both cases, the normal stress η evolves to a steady state stress distribution, but in the degenerate case the displacement jump g can grow linearly without bound along certain growth modes which are not suppressed by the stress-driven diffusion mechanism. The picture is suggestive of continental drift in plate tectonics, but the underlying physical mechanism is entirely different; see [31].

4. The grain boundary normal stress problem. In this section we describe the interface boundary conditions that are imposed when solving the Lamé equations, introduce the notion of degeneracy of a grain boundary network, and define the operators S and B. Rigorous proofs of the key properties of S and B are presented in the companion paper [31].

DEFINITION 4.1 (grain boundary normal stress). *Given a function $\eta \in L^2(\Gamma)$, find the displacement field $\mathbf{u} \in H^1(\Omega)^2$ satisfying $\mu\Delta\mathbf{u} + (\lambda + \mu)\nabla(\nabla \cdot \mathbf{u}) = 0$ in the interior of each grain subject to the boundary conditions*

$$\mathbf{u}(x) = \mathbf{0} \qquad (x \in \Gamma_0), \tag{4.1}$$

$$[\mathbf{u}(x^+) - \mathbf{u}(x^-)] \cdot \mathbf{t}_j = 0 \qquad (x \in \Gamma_j), \tag{4.2}$$

$$[\sigma(x^+) - \sigma(x^-)]\mathbf{n}_j = \mathbf{0} \qquad (x \in \Gamma_j), \tag{4.3}$$

$$\mathbf{n}_j \cdot \sigma(x)\mathbf{n}_j = \eta(x) \qquad (x \in \Gamma_j). \tag{4.4}$$

Here $H^1(\Omega)$ denotes the set $\{w \in L^2(\Omega) : w|_{\Omega_k} \in H^1(\Omega_k)\}$, so the jump in normal component of displacement $[\mathbf{u}(x^+) - \mathbf{u}(x^-)] \cdot \mathbf{n}_j$ is permitted to be nonzero for $x \in \Gamma_j$.

To make sense of the boundary condition (4.4) for a general $\eta \in L^2(\Gamma)$, a suitable notion of weak solution must be defined. This is done in [31], where it is shown that for certain "degenerate" grain boundary networks, additional compatibility conditions must be satisfied by η for a solution $\mathbf{u}$ to exist, and when it does exist, it is not unique. This situation may be characterized as follows.

DEFINITION 4.2. *A grain boundary network is said to be* degenerate *if there exists a nonzero displacement field* $\mathbf{u}$ *consisting of infinitesimal rigid body motions*

$$u_1\big|_{\Omega_k} = a_k - c_k y, \quad u_2\big|_{\Omega_k} = b_k + c_k x \quad (a_k, b_k, c_k \text{ constants}) \tag{4.5}$$

such that the boundary conditions (4.1)–(4.4) *hold with* $\eta \equiv 0$. *The jump in normal component of displacement across grain boundaries is permitted to be nonzero.*

In other words, degeneracy occurs when stress-free infinitesimal rigid body motions exist (grain by grain) that are zero at the outer walls and satisfy a no-sliding condition across grain boundaries. An algorithm for finding the degeneracies of any grain boundary network is presented in [31], where it is shown that degeneracy is a consequence of pathologies such as junction angles greater than 180° or a large number of quadruple (or higher) order junctions. In this paper we explicitly assume the grain boundary network is nondegenerate, leaving the general case to [31]. This substantially simplifies our proof of well posedness while retaining the key ideas.

DEFINITION 4.3. *When the grain boundary network is nondegenerate, the normal stress to grain growth operator* $B : L^2(\Gamma) \to L^2(\Gamma) : \eta \mapsto g$ *is defined via*

$$(B\eta)(x) = [\mathbf{u}(x^+) - \mathbf{u}(x^-)] \cdot \mathbf{n}_j \qquad (x \in \Gamma_j), \tag{4.6}$$

where $\mathbf{u}$ *is the (unique) solution to the grain boundary normal stress problem corresponding to* η.

Remark 4.4. As discussed in Figure 2.2, this definition is independent of the orientation chosen for each segment Γ_j.

DEFINITION 4.5. *When the grain boundary network is nondegenerate, the operator* $S : \mathcal{D}(S) \to L^2(\Gamma) : g \mapsto \eta$ *is defined as the inverse of* B. *In other words,* $\mathcal{D}(S) := \text{range}(B)$ *and for* $y \in \mathcal{D}(S)$, Sy *is the unique* x *such that* $Bx = y$.

THEOREM 4.6. B *is self-adjoint, negative, and compact.* S *is self-adjoint, negative, closed, and densely defined. In the nondegenerate case considered here,* B *is also injective and has dense range.*

Proof. See [31] for the proof. ∎

Remark 4.7. The domain $\mathcal{D}(S)$ is quite complicated due to the variety of ways self-similar solutions of the Lamé equations can behave near grain boundary junctions; see [26, 30]. In particular, even for smooth functions η that are continuous at junctions, $g = B\eta$ generally will be discontinuous at junctions and exhibit infinite slopes. As a result, it would be very difficult to define S directly by setting up a boundary value problem specifying g along Γ such that the resulting normal stress η is always meaningful in the trace sense. By defining S as we have, we can derive its properties by studying the compact operator B, which is well defined for all $\eta \in L^2(\Gamma)$. Moreover, this approach allows us to explain how it is possible to impose boundary conditions involving the normal stress at junctions where the stress tensor develops singularities: the components directly involved in the diffusion process remain finite and well behaved for all time (and satisfy the boundary conditions) while other components of the stress tensor blow up; see section 6.4.

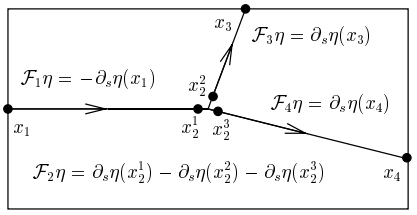


FIG. 5.1. *The flux $\mathcal{F}_i\eta$ into junction i is a sum over incident segments j of slopes $\pm\partial_s\eta$.*

5. The Poisson equation on the grain boundary network. In this section we show that the operator L is self-adjoint and positive, that there is a compact pseudoinverse G for L such that $I - LG$ is a finite rank projection onto $\ker(L)$ in $L^2(\Gamma)$, and that $A^{\frac{1}{2}} := (L + I - LG)^{\frac{1}{2}}$ is an isomorphism from $H^1(\Gamma)$ onto $L^2(\Gamma)$. We also characterize $\mathcal{D}(L)$ and $\mathcal{D}(L^{\frac{1}{2}})$ and identify the kernel of L as the set of functions that are constant on each connected component of Γ. Although these facts are well known in the case of the Neumann problem on the unit interval, several nonobvious tricks must be used to prove them for a network, and notation must be introduced that can cleanly handle the lack of a natural ordering and orientation for the grain boundary segments. In particular, we point out that not every function on a network with loops has a continuous antiderivative (unless its integral around each loop is zero).

5.1. Boundary conditions. As discussed in section 3.2, the operator L is the negative of the second derivative operator with respect to arc length on each grain boundary segment. If η is twice continuously differentiable on each segment and satisfies the boundary conditions

$$\begin{array}{ll} 1. & \eta \text{ is continuous at } x_i \\ 2. & \mathcal{F}_i\eta = 0 \end{array} \qquad (i \text{ any junction label}), \tag{5.1}$$

then the restriction of $L\eta$ to the interior of Γ_j is given by

$$L\eta(x) = -\frac{\partial^2\eta}{\partial s^2} \qquad (x \in \Gamma_j^o). \tag{5.2}$$

Here $\mathcal{F}_i$ is a flux operator for junction i, defined by

$$\mathcal{F}_i\eta = (-1)^{k_i}\partial_s\eta(x_i) \qquad (x_i \text{ a gb-wall junction}), \tag{5.3}$$

$$\mathcal{F}_i\eta = \sum_{j=1}^{p_i}(-1)^{k_i^j}\partial_s\eta(x_i^j) \quad (x_i \text{ a gb junction of order } p_i), \tag{5.4}$$

where x_i^j is infinitesimally close to junction x_i on segment j and k_i^j is 0 or 1 depending on whether segment j is parameterized toward or away from x_i; see Figure 5.1. At junctions where a grain boundary meets an outer wall, only the second condition in (5.1) is imposed since the first condition is automatic.

5.2. Integration by parts on the network. Let $C(\Gamma)$ denote the space of continuous functions on Γ, and let $\widetilde{C}(\Gamma)$ denote the space of functions f continuous on the interiors of the Γ_j with well-defined limits $f(x_i^j)$ at the endpoints x_i of Γ_j but with possibly different limiting values at x_i when approached from different segments.

Differentiation is defined segment by segment, where we recall that each segment is given an arbitrary orientation along which the arc length parameter increases. We define

$$\widetilde{C}^r(\Gamma) = \{f : f^{(k)} \in \widetilde{C}(\Gamma),\ 0 \le k \le r\}. \tag{5.5}$$

For $f \in \widetilde{C}(\Gamma)$, we write

$$\left[f\right]_\Gamma = \sum_j \left[f(x^j_{i_1(j)}) - f(x^j_{i_0(j)})\right], \tag{5.6}$$

where $i_0(j)$ and $i_1(j)$ are the junction indices of the initial and final endpoints of segment j. We note that although both sides of (5.7) and (5.8) depend on the orientation chosen for each segment, the identities

$$\left[f'g\right]_\Gamma = \sum_i g(x_i)\mathcal{F}_i f \qquad \left(f \in \widetilde{C}^1(\Gamma)\ g \in C(\Gamma)\right) \tag{5.7}$$

and

$$\int_\Gamma f'g\,ds = \left[fg\right]_\Gamma - \int_\Gamma g'f\,ds \qquad \left(f, g \in \widetilde{C}^1(\Gamma)\right) \tag{5.8}$$

are valid for any particular choice. It follows that

$$(Lu, v) = (u, Lv) \qquad \left(u, v \in \widetilde{C}^2(\Gamma) \text{ and satisfy (5.1)}\right) \tag{5.9}$$

holds independently of the orientations chosen for each segment; i.e., the boundary value problem $Lu = \lambda u$ subject to the boundary conditions (5.1) is self-adjoint.

5.3. Construction of a Green's function. In this section we construct a Green's function $G_l(x, y)$ for the operator $L + l^2$, where l is any positive real number. Since L has a nontrivial kernel, there is no Green's function when $l = 0$. We begin by defining an auxiliary function

$$K_l(x, y) = \begin{cases} 0, & x, y \text{ on different segments,} \\ k_l(a_j, s_j(x), s_j(y)), & x, y \in \Gamma_j, \end{cases} \tag{5.10}$$

where $a_j = |\Gamma_j|$ is the length of Γ_j, $s_j(x) = |x - x_{i_0(j)}|$ is the value of the arc length parameter along Γ_j at x, and $k_l(a, x, y)$ is the Green's function for $-\partial_x^2 + l^2$ on the interval $(0, a)$ with Dirichlet boundary conditions; see Figure 5.2. For $(0 \le x \le y \le a)$, we have

$$k_l(a, x, y) = k_l(a, y, x) = \left(\cosh ly \frac{\sinh la}{l} - \cosh la \frac{\sinh ly}{l}\right) \frac{\sinh lx}{\sinh la}. \tag{5.11}$$

To get G_l from K_l, we have to fix the flux boundary conditions at each junction. Let n be the number of junctions, and define the linear operator $T_l : \mathbb{R}^n \to \mathbb{R}^n$ as follows. For $w \in \mathbb{R}^n$, let $u_{l,w}$ be the unique function in $C(\Gamma)$ which satisfies

$$u_{l,w}(x_i) = w_i, \quad \partial_s^2 u_{l,w} = l^2 u_{l,w}. \tag{5.12}$$

Explicitly, on segment Γ_j, we set

$$u_{l,w}(x) = w_{i_0(j)} \cosh ls_j(x) + \left[w_{i_1(j)} - w_{i_0(j)} \cosh la_j\right] \frac{\sinh ls_j(x)}{\sinh la_j}. \tag{5.13}$$

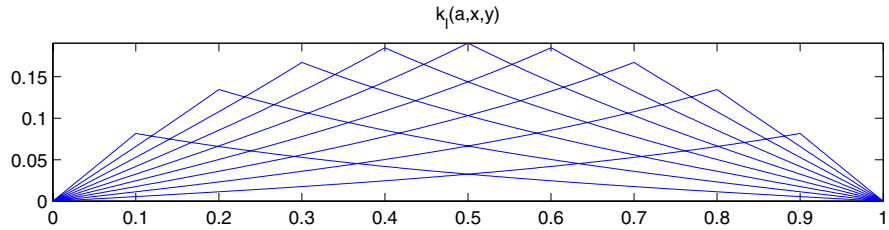


FIG. 5.2. *The function $k_l(a,x,y)$ satisfies $-\partial_x^2 k_l = l^2 k_l$ on $(0,y)$ and (y,a), is zero at $x=0$ and $x=a$, and has a unit (negative) jump in slope at $x=y$. In this plot, $a=1$, $l=2$, and $y=0.1,\dots,0.9$. Note that $\lim_{y\to 0}\frac{\partial}{\partial x}\big|_{x=0} k_l(a,x,y)=1$ with a similar result as $y\to 1$.*

Now we define the ith component of $T_l w$ to be the flux of $u_{l,w}$ into junction i:

$$(T_l w)_i = \mathcal{F}_i u_{l,w}. \tag{5.14}$$

Thus T_l converts values at junctions into fluxes at junctions of the solution to (5.12).

We next show that T_l is invertible. Suppose that $T_l w = 0$, i.e., that $\mathcal{F}_i u_{l,w} = 0$ at each junction. We wish to conclude that $u_{l,w} \equiv 0$, so we proceed by contradiction. Multiplying by (-1) if necessary, we assume the maximum value of $u_{l,w}$ is positive. The maximum cannot occur in the interior of a segment or at a gb-wall junction because the slope would be zero at such a maximum, while (5.12) would require that the second derivative be positive. It also cannot occur at a triple junction because the sum of the outward slopes is zero at such a junction: if any is positive it is not a maximum, and if each is zero we use (5.12) again. Thus we reach a contradiction and conclude that T_l is invertible.

To correct the flux boundary conditions of K_l to obtain G_l, we define $w_l(y) \in \mathbb{R}^n$ to give the values at the junctions of the solution to $\partial_x^2 u = l^2 u$ with the same junction fluxes as $K(\cdot,y)$:

$$w_l(y) = T_l^{-1}\left(\{\mathcal{F}_i K(\cdot,y)\}_{i=1}^n\right) \qquad (y \text{ not an endpoint}). \tag{5.15}$$

Note that $\mathcal{F}_i K(\cdot,y)$ is nonzero only when y is on a segment incident to junction i, and by (5.11), as y approaches junction i we have

$$\lim_{y\to x_i}\left(\{\mathcal{F}_k K(\cdot,y)\}_{k=1}^n\right) = -e_i \in \mathbb{R}^n. \tag{5.16}$$

We define $G_l(x,y)$ by

$$G_l(x,y) = \begin{cases} K(x,y) - u_{l,w_l(y)}(x), & y \text{ not an endpoint}, \\ u_{l,T_l^{-1}(e_i)}(x), & y = x_i. \end{cases} \tag{5.17}$$

By (5.16), $G_l(x,y)$ is continuous on $\Gamma\times\Gamma$, and by construction, for fixed y in the interior of some segment, $G_l(x,y)$ as a function of x satisfies the boundary conditions (5.1). It is readily verified using the corresponding property of $k_l(a,x,y)$ that the operator

$$\mathcal{G}_l f(x) = \int_\Gamma G_l(x,y) f(y)\, dy \tag{5.18}$$

is the inverse of $L + l^2$ in the sense that for all $f \in C(\Gamma)$ and all $u \in \widetilde{C}^2(\Gamma) \cap C(\Gamma)$ satisfying (5.1),

$$(L + l^2)\mathcal{G}_l f = f, \qquad \mathcal{G}_l(L + l^2)u = u. \tag{5.19}$$

$\mathcal{G}_l$ is self-adjoint since L is self-adjoint, and as a result, $G_l(x, y) = G_l(y, x)$ for all $x, y \in \Gamma$.

THEOREM 5.1. *There exists an orthonormal basis $\{\varphi_n\}_{n=1}^{\infty}$ for $L^2(\Gamma)$ and an increasing sequence of nonnegative numbers λ_n growing without bound such that*

$$\varphi_n \in \widetilde{C}^{\infty}(\Gamma) \cap C(\Gamma), \tag{5.20}$$

$$\mathcal{F}_i \varphi_n = 0 \qquad (i \text{ any junction label}), \tag{5.21}$$

$$L\varphi_n = \lambda_n^2 \varphi_n. \tag{5.22}$$

The domain of L satisfies

$$\{f \in \widetilde{C}^2(\Gamma) : f \text{ satisfies (5.1)}\} \subset \mathcal{D}(L) \subset \{f \in \widetilde{C}^1(\Gamma) : f \text{ satisfies (5.1)}\}. \tag{5.23}$$

Proof. In the standard way [7, 8], we can show that $\mathcal{G}_l$ is a self-adjoint, compact operator on $L^2(\Gamma)$; thus $\mathcal{G}_l$ has a complete orthonormal set $\{\varphi_n\}_{n=1}^{\infty}$ of eigenfunctions with eigenvalues converging to zero. It is also readily shown that for all $f \in L^1(\Gamma) \supset L^2(\Gamma)$,

$$\mathcal{G}_l f \in \{\eta \in \widetilde{C}^1(\Gamma) \,:\, \eta \text{ satisfies the boundary conditions (5.1)}\}; \tag{5.24}$$

hence the eigenfunctions are continuous and satisfy flux boundary conditions. Differentiating (5.18), it follows that if $f \in \widetilde{C}^r(\Gamma)$, then $\mathcal{G}_l f \in \widetilde{C}^{r+2}(\Gamma)$; thus the eigenfunctions belong to $\widetilde{C}^{\infty}(\Gamma)$ by a bootstrap argument. By (5.19), they are eigenfunctions of $L + l^2$ with reciprocal eigenvalues. Since $L + l^2$ is invertible for $l > 0$, we conclude that the eigenvalues of L form an unbounded sequence of nonnegative numbers $\{\lambda_n^2\}$. Finally, (5.23) holds when we *redefine* L to be $\mathcal{G}_l^{-1} - l^2$. Then $\mathcal{D}(L) = \operatorname{ran}(\mathcal{G}_l)$, so the first inclusion follows from (5.19) and the second from (5.24). ∎

5.4. The kernel of *L*. The segments Γ_j of the grain boundary network can be grouped together into connected components as sets in $\mathbb{R}^2$. We decompose the numbers $1, \dots, N$ into a collection $\mathcal{J}$ of disjoint sets such that each $J \in \mathcal{J}$ is the set of indices of the segments Γ_j that belong to component $\Gamma_J = \cup_{j \in J} \Gamma_j$. We number these subsets arbitrarily $\mathcal{J} = \{J_1, \dots, J_d\}$ and define the functions $e_k \in L^2(\Gamma)$ for $1 \le k \le d$ by

$$e_k(x) = \begin{cases} |\Gamma_{J_k}|^{-\frac{1}{2}}, & x \in \Gamma_{J_k}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.25}$$

Here $|\Gamma_{J_k}| = \sum_{j \in J_k} |\Gamma_j|$ is the sum of the lengths of the segments making up component k. Note that each e_k is continuous at all junctions since all segments that meet at a junction belong to the same connected component.

PROPOSITION 5.2. *The functions $\{e_k\}_{k=1}^{d}$ form a basis for* $\ker(L)$.

Proof. Theorem 5.1 ensures that $\ker(L)$ is finite dimensional and is spanned by functions $\{\varphi_1, \dots, \varphi_{d'}\}$ satisfying (5.20)–(5.22) with $\lambda_n = 0$. Since each e_k also satisfies these conditions, we have $d' \ge d$, and it remains to show that any $\varphi \in \ker(L)$ is constant on each connected component. Suppose not. Then there is a segment Γ_{j*}

on which φ is not constant. Since $L\varphi = 0$, φ is linear on each segment. Starting with Γ_{j^*}, there is a path to an outer wall along which φ strictly increases. This is because φ satisfies (5.21), so if φ increases along a segment as we approach a triple junction, it must also increase along one of the other segments as we leave the junction. Since φ strictly increases along the path, no interior node can be revisited, and eventually the path reaches a wall with a positive slope, which contradicts (5.21). ∎

5.5. An isomorphism. In this section we show that $L^{\frac{1}{2}}$ becomes an isomorphism from $H^1(\Gamma)$ onto $L^2(\Gamma)$ if we modify it slightly to eliminate its kernel. It will be useful to define the operators

$$P = I - \sum_{n=1}^{d} (\cdot, \varphi_n)\varphi_n, \qquad A = L + \sum_{n=1}^{d} (\cdot, \varphi_n)\varphi_n. \tag{5.26}$$

Note that P is the orthogonal projection onto the subspace

$$\operatorname{ran}(L) = \ker(L)^{\perp} = \left\{ f \in L^2(\Gamma) \,:\, \int_{\Gamma_{J_k}} f\, ds = 0,\ 1 \le k \le d \right\} \tag{5.27}$$

and $L = AP = PA$. A and P are self-adjoint since the φ_n are orthogonal.

THEOREM 5.3. *For any absolutely continuous f such that $f' \in L^2(\Gamma)$, if we write $f = \sum a_n \varphi_n$, then*

$$\|f\|^2 = \sum_{n=1}^{\infty} |a_n|^2, \qquad \|f'\|^2 = \sum_{n=1}^{\infty} |a_n \lambda_n|^2. \tag{5.28}$$

There is a constant C such that

$$\|f\|_{L^2} \le \|A^{-\frac{1}{2}} f\|_{H^1} \le C \|f\|_{L^2} \qquad \big(f \in L^2(\Gamma)\big); \tag{5.29}$$

i.e., $A^{-\frac{1}{2}}$ is an isomorphism from $L^2(\Gamma)$ onto $H^1(\Gamma)$ and is therefore compact as an operator on $L^2(\Gamma)$. The domain of $L^{\frac{1}{2}}$ is $H^1(\Gamma)$, which requires continuity but imposes no constraints on the derivatives at junctions.

Proof. On each segment, we have

$$\varphi_n(x) = c_{n,j} \cos(\lambda_n s_j(x) - \theta_{n,j}) \qquad (x \in \Gamma_j). \tag{5.30}$$

We define

$$\psi_n = \lambda_n^{-1} \varphi_n' \qquad (n > d). \tag{5.31}$$

Note that ψ_n is *not* continuous on Γ but is zero at gb-wall junctions and satisfies appropriate jump conditions at triple junctions so that

$$\big[\psi_n g\big]_{\Gamma} = 0 \qquad (g \in C(\Gamma)). \tag{5.32}$$

We claim that $\{\psi_n\}_{n>d}$ is an orthonormal basis for the subset of functions $f \in L^2$ such that f is the derivative of an absolutely continuous function in $C(\Gamma)$. This subset will not be all of L^2 as soon as there are loops in the grain boundary network, since the integral around a loop must be zero for a continuous antiderivative to exist. By (5.8), we have the orthogonality condition

$$\int_{\Gamma} \psi_n \psi_m = \lambda_n^{-1} \left(\big[\varphi_n \psi_m\big]_{\Gamma} - \int_{\Gamma} \psi_m' \varphi_n \right) = \int_{\Gamma} \varphi_m \varphi_n = \delta_{mn}. \tag{5.33}$$

To prove completeness, suppose $W \in C(\Gamma)$ is absolutely continuous and $w = W'$ belongs to L^2. Suppose further that for all $n > d$, we have

$$\int_\Gamma w\psi_n = 0. \tag{5.34}$$

We must show that $w \equiv 0$. Integrating by parts, we obtain

$$\int_\Gamma w\psi_n = \big[W\psi_n\big]_\Gamma - \int_\Gamma \psi_n' W = \lambda_n \int_\Gamma W\varphi_n = 0 \qquad (n > d). \tag{5.35}$$

Since $\{\varphi_n\}_{n=1}^\infty$ is a basis, we conclude that W is a linear combination of $\varphi_1, \dots, \varphi_d$. Thus W is a constant on each connected component of Γ and $w = W' = 0$, as desired.

For any $f \in L^2$, we may expand $f = \sum_{n=1}^\infty a_n\varphi_n$ and apply the Parseval identity to conclude that $\|f\|^2 = \sum |a_n|^2$. If f is absolutely continuous and its derivative is in L^2, then we have $f' = \sum_{n>d} b_n\psi_n$ with

$$b_n = \int_\Gamma f'\psi_n = \big[f\psi_n\big]_\Gamma - \int_\Gamma f\psi_n' = \lambda_n \int_\Gamma f\varphi_n = \lambda_n a_n. \tag{5.36}$$

Since $\lambda_n = 0$ for $n \le d$, the Parseval identity gives the result $\|f'\|^2 = \sum_1^\infty |\lambda_n a_n|^2$. Therefore we have

$$\begin{aligned} \|f\|_{H^1}^2 &= \|f\|_{L^2}^2 + \|f'\|_{L^2}^2 = \sum_{n=1}^\infty (1+\lambda_n^2)|a_n|^2, \\ \|A^{\frac12} f\|_{L^2}^2 &= \sum_{n=1}^d |a_n|^2 + \sum_{n=d+1}^\infty \lambda_n^2 |a_n|^2. \end{aligned} \tag{5.37}$$

As a result, we obtain

$$\|A^{\frac12} f\|_{L^2} \le \|f\|_{H^1} \le C\|A^{\frac12} f\|_{L^2} \qquad \big(f \in H^1(\Gamma)\big), \tag{5.38}$$

with $C = \lambda_*^{-1}\sqrt{1+\lambda_*^2}$, where $\lambda_* = \lambda_{d+1}$ is the smallest nonzero eigenvalue of $L^{\frac12}$. ∎

DEFINITION 5.4. *The operator G is defined via*

$$G = A^{-1} - \sum_{n=1}^d (\cdot, \varphi_n)\varphi_n. \tag{5.39}$$

G is the pseudoinverse of L in the sense that they have the same kernel and eigenfunctions with reciprocal (or zero) eigenvalues. The properties of A imply that G is self-adjoint and compact on $L^2(\Gamma)$ and satisfies $G = PA^{-1} = A^{-1}P$ and $LG = P$.

6. Dynamics. In this section we show that if the grain boundary network is nondegenerate, then the equation

$$\eta_t = SL\eta, \qquad \eta(0) = \eta_0, \tag{6.1}$$

generates an analytic semigroup $\{E_t : t \ge 0\}$ of bounded linear operators on $H^1(\Gamma)$. See [31] for the degenerate case and the books [13, 32, 16, 1] for background information on semigroup theory. As mentioned in section 3.3, the solution $\eta(t)$ when the electromigration force is present is given by

$$\eta(t) = E_t(\eta_0 + \psi) - \psi. \tag{6.2}$$

The boundary conditions on η enforcing chemical potential continuity and flux balance at junctions hold as a consequence of the analyticity of E_t and the properties of the domain $\mathcal{D}(SL)$. The role of singularities in the stress field near corners and junctions is also discussed.

6.1. The semigroup generated by SL. In this section we show that there is a Riesz basis (a basis equivalent to an orthonormal basis [11]) for $H^1(\Gamma)$ consisting of eigenfunctions of SL. This allows us to exhibit the semigroup operator E_t explicitly, study its properties, and approximate it numerically [30, 26]. Throughout this section, we assume the grain boundary network is nondegenerate so that B is injective and $SB = I$ on $L^2(\Gamma)$.

LEMMA 6.1. *Let us denote* $\ker(L) = \operatorname{span}\{e_k\}_{k=1}^d$ *by* $\{e\}$. *Then*

$$\ker(SL) = \{e\}, \qquad \operatorname{ran}(SL) = \{Be\}^\perp. \tag{6.3}$$

Proof. SL is densely defined in $L^2(\Gamma)$ since $\mathcal{D}(S)$ is dense, A^{-1} is bounded with dense range, and $\mathcal{D}(SL) = A^{-1}\mathcal{D}(S)$. Clearly $\{e\} \subset \ker(SL)$. Since S is injective on its domain, if $Lx \in \mathcal{D}(S)$ is nonzero, so is SLx. Thus $\ker(SL) = \{e\}$ as claimed. Suppose $y = SLx$. Then $By = Lx$ belongs to $\operatorname{ran}(L) = \{e\}^\perp$, so $(y, Be) = 0$ for all $e \in \ker(L)$, as claimed. ∎

LEMMA 6.2. *The (nonorthogonal) projection on* $L^2(\Gamma)$ *given by*

$$Q \text{ projects along } \{e\} \text{ onto } \{Be\}^\perp \tag{6.4}$$

is well defined.

Proof. Since B is injective, $\{e\}$ and $\{Be\}$ have the same dimension. We must show that $\{e\} \cap \{Be\}^\perp = \{0\}$. Suppose $x \in \{e\} \cap \{Be\}^\perp$. Then $(x, Bx) = 0$, which implies $x = 0$ since B is self-adjoint and negative definite. ∎

Remark 6.3. Q may be written explicitly as $I - (\cdot, w_k)e_k$ (summation implied), where $w_k = (Be_j)\alpha_{jk}$ and $(e_i, Be_j)\alpha_{jk} = \delta_{ik}$. Since the L^2 inner product $(\cdot, w_k)$ is a bounded linear functional on $H^1(\Gamma)$ and $e_k \in H^1(\Gamma)$, Q is also a well-defined projection on $H^1(\Gamma)$. By contrast, the L^2 adjoint $Q^* = I - (\cdot, e_k)w_k$ is generally not defined on $H^1(\Gamma)$ due to the possibility of singularities in the arc length derivative of w_k near junctions.

LEMMA 6.4. *The following diagram is commutative in the sense that for each block* $X \underset{g}{\overset{f}{\rightleftarrows}} Y$ *we have* $f \circ g = id_Y$ *and* $g \circ f = id_{\mathcal{D}(f)}$:

$$\{Be\}^\perp \underset{Q}{\overset{P}{\rightleftarrows}} \{e\}^\perp \underset{G}{\overset{L}{\rightleftarrows}} \{e\}^\perp \underset{B}{\overset{S}{\rightleftarrows}} \{Be\}^\perp . \tag{6.5}$$

Proof. P and Q both project along $\{e\}$, so $PQ = P$ and $QP = Q$. Since $\operatorname{ran}(P) = \{e\}^\perp$ and $\operatorname{ran}(Q) = \{Be\}^\perp$, the block involving P and Q is commutative. Since LG is the identity on $\{e\}^\perp$ and GL is the identity on $\mathcal{D}(L) \cap \{e\}^\perp = \operatorname{ran}(G)$, the block involving L and G is commutative. If $(x, Be) = 0$, then $(Bx, e) = 0$, so B maps $\{Be\}^\perp$ to $\{e\}^\perp$. Since SB is the identity on $L^2(\Gamma)$ and BS is the identity on $\mathcal{D}(S) = \operatorname{ran}(B)$, the block involving S and B is commutative. ∎

THEOREM 6.5. *There is a Riesz basis* $\{\phi_k\}$ *for* $H^1(\Gamma)$ *and a nonincreasing, unbounded sequence of numbers* $\lambda_k \le 0$ *such that* $SL\phi_k = \lambda_k\phi_k$.

Proof. The operator $K = G^{\frac{1}{2}}Q^*BQG^{\frac{1}{2}}$ is compact and self-adjoint, so there is an orthonormal basis $\{\varphi_k\}$ for $L^2(\Gamma)$ such that $K\varphi_k = \mu_k\varphi_k$, ($\mu_k \in \mathbb{R}$, $\mu_k \to 0$). Note

that $\mu_k = (\varphi_k, K\varphi_k) = (QG^{\frac{1}{2}}\varphi_k, BQG^{\frac{1}{2}}\varphi_k) \le 0$, with equality iff $QG^{\frac{1}{2}}\varphi_k = 0$. Note that $\ker(QG^{\frac{1}{2}}) = \ker(G^{\frac{1}{2}}) = \{e\}$ since $\ker(Q) = \{e\}$ and $\operatorname{ran}(G^{\frac{1}{2}}) \subset \{e\}^\perp$. We may therefore assume $\mu_1, \dots, \mu_d = 0$, and the remaining μ_k form an increasing sequence of negative numbers converging to zero. Note that

$$QGQ^*B(QG^{\frac{1}{2}}\varphi_k) = \mu_k(QG^{\frac{1}{2}}\varphi_k). \tag{6.6}$$

From the above remarks, we know $QG^{\frac{1}{2}}\varphi_k$ is nonzero for $k > d$. We define

$$\phi_k = \left\{ \begin{array}{ll} \varphi_k, & k = 1, \dots, d \\ QG^{\frac{1}{2}}\varphi_k, & k > d \end{array} \right\}, \quad \lambda_k = \left\{ \begin{array}{ll} 0, & k = 1, \dots, d \\ \mu_k^{-1}, & k > d \end{array} \right\}. \tag{6.7}$$

From the definition of Q, it follows that $Q^*B = BQ$, so $QGBQ\phi_k = \mu_k\phi_k$. If $k \le d$, then $SL\phi_k = 0$ since $\phi_k \in \{e\}$. Otherwise, $Q\phi_k = \phi_k$ and Lemma 6.4 gives $SL\phi_k = SL(\mu_k^{-1}QGB\phi_k) = \lambda_k\phi_k$. It is readily verified that for $k \ge 1$,

$$\phi_k = \left[(I - P) + QG^{\frac{1}{2}}\right]\varphi_k, \qquad \varphi_k = \left[(I - Q) + L^{\frac{1}{2}}\right]\phi_k. \tag{6.8}$$

Since $[I - P + QG^{\frac{1}{2}}]$ is bounded from $L^2(\Gamma)$ to $H^1(\Gamma)$ and its inverse $[I - Q + L^{\frac{1}{2}}]$ is bounded in the other direction, they are isomorphisms. Thus the ϕ_k form a Riesz basis for $H^1(\Gamma)$ as claimed. ∎

THEOREM 6.6. *The initial value problem $\eta_t = SL\eta$, $\eta(0) = \eta_0$ generates an analytic semigroup $\{E_t : t \ge 0\}$ of bounded linear operators on $H^1(\Gamma)$.*

Proof. Since the ϕ_k form a Riesz basis, the mapping from $H^1(\Gamma)$ to l^2 giving the coefficients of the expansion $\eta_0 = \sum_k a_k\phi_k$ is an isomorphism, and there is a constant C independent of η_0 such that $C^{-2}\|\eta_0\|_{H^1}^2 \le \sum_{k=1}^\infty |a_k|^2 \le C^2\|\eta_0\|_{H^1}^2$. These coefficients may be determined via

$$a_k = ([I - Q + L^{\frac{1}{2}}]\eta_0, \varphi_k) = (\eta_0, \phi_k^*), \qquad \phi_k^* = [I - Q^* + L^{\frac{1}{2}}]\varphi_k, \tag{6.9}$$

where $(\cdot, \cdot)$ is the $L^2(\Gamma)$ inner product. For $\eta_0 \in H^1(\Gamma)$ we define

$$E_t\eta_0 = \textstyle\sum_k a_k e^{\lambda_k t}\phi_k, \qquad a_k = (\eta_0, \phi_k^*). \tag{6.10}$$

E_t is bounded for any $t \ge 0$ since $\lambda_k \le 0$ for all k, and hence $\|E_t\eta_0\|_{H^1} \le C^2\|\eta_0\|_{H^1}$. E_0 is clearly the identity on $H^1(\Gamma)$, and $E_{t+s} = E_tE_s$ since $(\phi_j, \phi_k^*) = \delta_{jk}$. For fixed η_0, the mapping $t \mapsto E_t\eta_0$ is continuous for $t \ge 0$ since

$$\|E_t\eta_0 - E_s\eta_0\|_{H^1} \le C\left(\sum_{k=1}^N |a_k(e^{\lambda_k t} - e^{\lambda_k s})|^2 + \sum_{k=N+1}^\infty |a_k|^2\right)^{\frac{1}{2}} \tag{6.11}$$

may be made arbitrarily small by choosing N large enough to make the second term small and then s close enough to t to make the first term small. Note that for any fixed $t > 0$ and $\eta_0 \in H^1(\Gamma)$ we have

$$\left\|\frac{E_{t+h}\eta_0 - E_t\eta_0}{h} - SLE_t\eta_0\right\|_{H^1} \le C\left(\sup_{k\ge1}\left|t^{-1}f_1(\lambda_k t)f_2(\lambda_k h)\right|\right)\left(\sum_{k=1}^\infty |a_k|^2\right)^{\frac{1}{2}}, \tag{6.12}$$

where $f_1(z) = ze^z$ and $f_2(z) = \frac{1}{z}(e^z - 1) - 1$. The supremum may be made arbitrarily small by taking h sufficiently close to zero since f_1 and f_2 are bounded on

the negative real axis, $\lim_{z\to-\infty} f_1(z) = 0$, and $\lim_{z\to 0} f_2(z) = 0$. Thus we see that SL is the generator of $\{E_t : t \geq 0\}$. To show that this semigroup is analytic, we need only check that $\limsup_{t\downarrow 0} t\|SLE_t\| < \infty$, which follows from $t\|SLE_t\eta_0\|_{H^1} \leq C\left(\sup_k |f_1(\lambda_k t)|\right)\left(\sum_1^\infty |a_k|^2\right)^{1/2} \leq C^2 e^{-1}\|\eta_0\|_{H^1}$. ∎

Remark 6.7. Equation (6.10) leads to a useful numerical method in which ϕ_k, ϕ_k^*, and λ_k are computed by approximating $QGBQ$ (which has the same eigenfunctions as SL with reciprocal or zero eigenvalues) using a singularity capturing least squares finite element method; see [30, 26].

Remark 6.8. Since E_t is an operator on $H^1(\Gamma)$ and the formula for the evolution of normal stress is given by

$$\eta(t) = E_t(\eta_0 + \psi) - \psi, \tag{6.13}$$

we should verify that ψ belongs to $H^1(\Gamma)$. Since ψ is the solution to the Laplace equation on a domain with corners, it is smooth in the interior of Ω with singularities of the form

$$r^\lambda \phi(\theta) \qquad (\lambda = \pi/\omega \geq 1/2) \tag{6.14}$$

near reentrant corners of opening angle $\omega \leq 2\pi$ [12, 19]. As a result, the restriction of ψ to Γ is continuous on Γ, differentiable in the interior of each Γ_j, and its derivative with respect to arc length cannot diverge at the endpoints faster than $s^{(\lambda_{\min}-1)}$. (It will diverge at all only if the grain boundary terminates at a reentrant corner of the domain.) If we assume that Ω has no cracks with $\omega = 2\pi$ or, if it has cracks, that the crack tips do not lie on grain boundaries, then $\lambda_{\min} > \frac{1}{2}$ and $\psi \in H^1(\Gamma)$ as claimed.

6.2. The steady state stress distribution. Since $\lambda_1 = \cdots = \lambda_d = 0$, E_t leaves $\{e\} = \ker(SL)$ invariant. Since $\lambda_k \leq \lambda_{d+1} < 0$ for $k > d$, E_t takes any vector in $\{Be\}^\perp = \mathrm{ran}(SL)$ to zero as $t \to \infty$. More precisely, one may show [31] that $\lim_{t\to\infty} E_t = I - Q$ in norm. Thus

$$\eta_{\text{steady}} = \lim_{t\to\infty} \eta(t) = (I - Q)(\eta_0 + \psi) - \psi. \tag{6.15}$$

As a result, in the nondegenerate case there are constants c_j such that

$$\eta_{\text{steady}} = -\psi + c_j e_j \qquad \text{(summation implied)}. \tag{6.16}$$

We also observe that $(\eta_{\text{steady}} - \eta_0) = -Q(\eta_0 + \psi) \in \{Be\}^\perp$, so

$$(\eta_{\text{steady}} - \eta_0, Be_k) = 0 \qquad (1 \leq k \leq d). \tag{6.17}$$

Since B is self-adjoint, this implies $(g_{\text{steady}} - g_0, e_k) = 0$, which is a statement of mass conservation on each connected component of Γ. Using (6.16) and (6.17), we have

$$c_j(e_j, Be_k) = (\eta_0 + \psi, Be_k) \qquad (1 \leq k \leq d), \tag{6.18}$$

which determines the c_j uniquely due to the fact that the $d\times d$ matrix with components (e_j, Be_k) is invertible. Note that the steady state flux $\partial_s(\eta_{\text{steady}}+\psi)$ is zero; this ceases to be true in the degenerate case [31], where (6.16) has additional nonconstant terms.

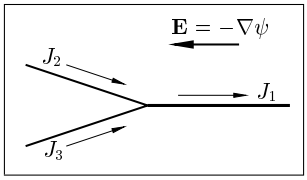


FIG. 6.1. *The electric field can lead to a flux imbalance at a triple (or wall) junction which must be immediately compensated by stress gradients to satisfy mass conservation. As a result, ψ generally will not lie in $\mathcal{D}(L)$ although $\eta(t) + \psi \in \mathcal{D}(L)$ when $t > 0$.*

6.3. Boundary conditions. ψ does not necessarily satisfy zero flux boundary conditions at junctions, and therefore, although $\psi \in \mathcal{D}(L^{1/2})$, it is not necessarily in $\mathcal{D}(L)$; see Figure 6.1. On the other hand, because E_t is analytic, we have range$(E_t) \subset \mathcal{D}(SL)$ for all $t > 0$. Therefore $\eta(t)$ in (6.13) has the property that

$$\eta(t) + \psi \in \mathcal{D}(L) \qquad (t > 0). \tag{6.19}$$

This implies chemical potential continuity and flux balance at all junctions for $t > 0$ (conditions 4b, 5b, and 5c in Figure 2.3). The grain growth function g may be obtained from η via $g(t) = B\eta(t)$, which automatically satisfies the compatibility conditions 4a and 5a of Figure 2.3 by virtue of the definition of B in (4.6). We have therefore proved that the grain boundary diffusion problem is well posed.

6.4. Stress singularities at junctions. It is well known that solutions to elliptic systems (such as the Lamé equations) on domains with corners and interface junctions exhibit singularities at these junctions. In the current case, as the normal stress η evolves on the grain boundary network, the stress and displacement fields in the bulk grains evolve as the solution to the grain boundary normal stress problem with η specified on Γ; see Definition 4.1. The general theory [30, 19, 24] states that the singular part of the solution may be written as a sum of power solutions (each component of the form $r^\lambda \phi(\theta)$ in local polar coordinates) to the *homogeneous* boundary value problem. As a result, the singular part of the solution near a given junction satisfies boundary conditions (4.1)–(4.4) with $\eta \equiv 0$ along the grain boundaries entering the junction. Although a different linear combination of stress components will generally diverge as the junction is approached along Γ, $\mathbf{n} \cdot \sigma\mathbf{n}$ will remain finite and well behaved, and all boundary conditions in Figure 2.3 describe quantities that remain finite despite the singularities. The corresponding displacement jump g will also remain finite, although it will generally exhibit infinite slopes and discontinuities (compatible with the boundary conditions) at junctions.

We have therefore demonstrated a mechanism through which stress components directly involved in the mass transport process remain bounded and well behaved while other "hidden" stress components grow very large and develop singularities; these components may be responsible for void nucleation and stress-induced damage but are omitted from commonly used scalar stress-generation models. In [30, 26], the first several terms in the asymptotic expansion for $\mathbf{u}$ and σ are computed a priori and added to the finite element basis to improve accuracy without mesh refinement. These singular functions often are very complicated, with singularity exponents clustered together in the complex plane.

Appendix. Infinite interconnect line.

In this section, we work out an exact solution to the stress-driven grain boundary diffusion problem $g_t = -\eta_{xx}$ for an infinite interconnect line with a single grain

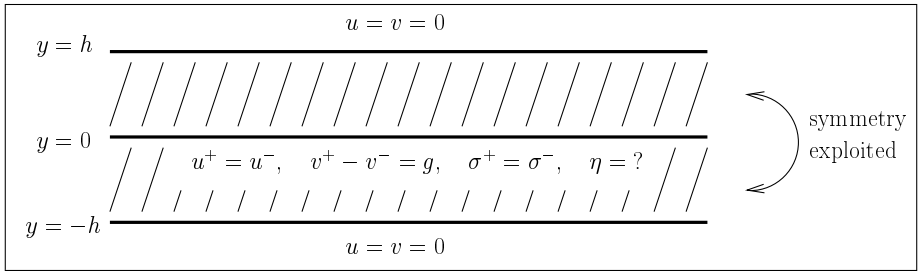


FIG. A.1. *The geometry and boundary conditions for the infinite strip.*

boundary running through the center. This provides useful insight about the nature of the diffusion process without the complication of boundary conditions at junctions or singularities in the stress field. The approach is to solve the elastic equations for a sinusoidal grain growth g and then use separation of variables and the Fourier transform to determine the evolution for an arbitrary initial condition.

A.1. Elastic equations for sinusoidal grain growth. Suppose the grain boundary coincides with the x-axis, and let h denote the width of each grain, as shown in Figure A.1. If we let $u = u_1$, $v = u_2$ and define κ, α, β, γ, and τ by

$$\kappa = \frac{\lambda + 3\mu}{\lambda + \mu}, \qquad \sigma = \mu \begin{pmatrix} \alpha - \gamma & \tau \\ \tau & \alpha + \gamma \end{pmatrix}, \qquad \beta = \frac{2}{\kappa + 1}(v_x - u_y), \tag{A.1}$$

then complex variable methods in plane elasticity [22, 30] can be used to guarantee the existence of holomorphic ϕ and ψ (known as Muskhelishvili functions) such that

$$\begin{aligned} \alpha + i\beta &= 2\phi', \\ \gamma + i\tau &= \bar{z}\phi'' + \psi', \\ u + iv &= \frac{1}{2}\left(\kappa\phi - z\bar{\phi}' - \bar{\psi}\right). \end{aligned} \tag{A.2}$$

By symmetry, for any displacement jump $g(x)$ the variables in the top grain will be related to the variables in the bottom grain via

$$u^+(x,y) = u^-(x,-y), \qquad v^+(x,y) = -v^-(x,-y), \tag{A.3}$$

$$\alpha^+(x,y) = \alpha^-(x,-y), \qquad \beta^+(x,y) = -\beta^-(x,-y), \tag{A.4}$$

$$\gamma^+(x,y) = \gamma^-(x,-y), \qquad \tau^+(x,y) = -\tau^-(x,-y). \tag{A.5}$$

Thus it is sufficient to restrict attention to the top grain. At $y = h$, we impose Dirichlet boundary conditions $u = v = 0$. Along the grain boundary, the four conditions $u^+ = u^-$, $v^+ - v^- = g$, $\alpha^+ + \gamma^+ = \alpha^- + \gamma^-$, $\tau^+ = \tau^-$ reduce to

$$\tau = 0, \qquad v = \frac{g}{2} \qquad \text{(boundary conditions along grain boundary)}. \tag{A.6}$$

We observe that in the limit as $h \to \infty$, these boundary conditions coincide with the problem of a rigid stamp without friction on a half-space and can be solved using singular integrals [23]. We omit details since the result for finite h covers this case in the limit.

For finite h, the singular integral approach does not work (at least not easily), so instead we take g of the form

$$g(x) = c_1 \cos \omega x + c_2 \sin \omega x \tag{A.7}$$

and make an ansatz for the form of the Muskhelishvili functions:

$$\phi = (a_1 + ia_2)\cos \omega z + (a_3 + ia_4)\sin \omega z, \tag{A.8}$$

$$\psi = (a_5 + ia_6)\cos \omega z + (a_7 + ia_8)\sin \omega z + (a_9 + ia_{10})z\cos \omega z + (a_{11} + ia_{12})z \sin \omega z. \tag{A.9}$$

We wish to determine if there are real coefficients a_i for which the boundary conditions are satisfied. We begin by constructing the 4×12 real matrix $A_0(\omega, \kappa, h, x)$ whose ith column contains the boundary conditions $u(y = h)$, $v(y = h)$, $v(y = 0)$, $\tau(y = 0)$ for the ϕ and ψ corresponding to a_i. For example, the second column corresponds to $\phi = i \cos \omega z$, $\psi = 0$:

$$\operatorname{col}_2(A_0) = \frac{1}{2}\begin{pmatrix} \omega h \sin \omega x \cosh \omega h - \omega x \cos \omega x \sinh \omega h + \kappa \sin \omega x \sinh \omega h \\ -\omega h \cos \omega x \sinh \omega h - \omega x \sin \omega x \cosh \omega h + \kappa \cos \omega x \cosh \omega h \\ -\omega x \sin \omega x + \kappa \cos \omega x \\ -2\omega^2 x \cos \omega x \end{pmatrix}. \tag{A.10}$$

Next we define the 16×12 real matrix $A(\omega, \kappa, h)$ by expanding each row of A_0 into four rows containing the coefficients of $\cos(\omega x)$, $\sin(\omega x)$, $x\cos(\omega x)$, $x\sin(\omega x)$. To satisfy the boundary conditions (A.6), we need to find $a \in R^{12}$ such that $Aa = b$, where b contains the desired coefficients of the terms $\cos(\omega x)$, $\sin(\omega x)$, $x\cos(\omega x)$, $x\sin(\omega x)$ in the boundary conditions. Explicitly, b and the second column of A are given by

$$b = \frac{1}{2}\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_1 \\ c_2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \operatorname{col}_2(A) = \frac{1}{2}\begin{pmatrix} 0 \\ \omega h \cosh \omega h + \kappa \sinh \omega h \\ -\omega \sinh \omega h \\ 0 \\ -\omega h \sinh \omega h + \kappa \cosh \omega h \\ 0 \\ 0 \\ -\omega \cosh \omega h \\ \kappa \\ 0 \\ 0 \\ -\omega \\ 0 \\ 0 \\ -2\omega^2 \\ 0 \end{pmatrix}. \tag{A.11}$$

We verify that a solution exists by computing the nullspace of A^T symbolically and checking that $b \in (\ker A^T)^\perp = \text{image } A$. We then select 12 linearly independent rows of A (and the corresponding rows of b) and solve $Aa = b$ symbolically. The resulting a determines ϕ and ψ, which we use to compute $\eta = \sigma_{22} = \alpha + \gamma$ along the grain boundary. This has to be done only once since the parameters such as κ and h appear

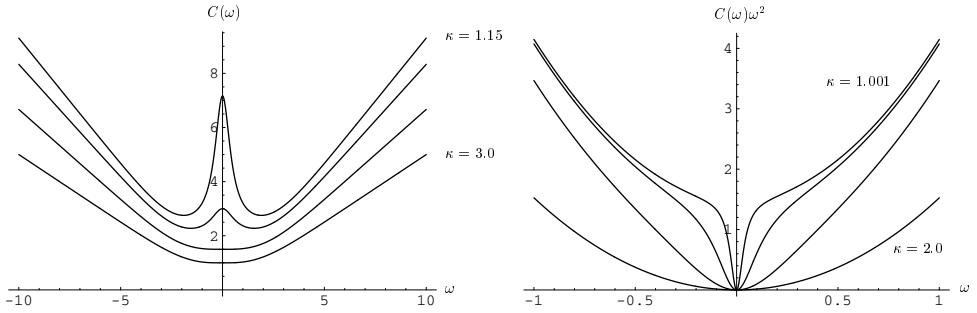


FIG. A.2. *Left: plot of $C(\omega)$ for $h = 1$ and $\kappa = 1.15, 1.4, 2.0, 3.0$. Right: plot of the dissipation rate $C(\omega)\omega^2$ for $h = 1$ and $\kappa = 1.001, 1.01, 1.1, 2.0$. Note that $C(\omega)$ diverges in the incompressible ($\kappa \to 1$), long wavelength ($\omega \to 0$) limit and that although $C(\omega)$ is not monotonic for $\kappa < 2$, $C(\omega)\omega^2$ is monotonic for $\omega \geq 0$. The envelope of the graphs of $C(\omega)\omega^2$ is $\frac{3}{2} + \frac{27}{10}\omega^2$ near the origin as $\kappa \to 1$.*

symbolically. All of this, including the construction of A_0 and A via (A.2), (A.8), and (A.9), can be done without difficulty using Mathematica or Maple.

The result of this computation is that along the grain boundary, η is a constant multiple of g for any c_1, c_2, ω:

$$\eta(x) = -C(\omega)g(x), \qquad C(\omega) = \frac{\omega[1 + \kappa^2 + 4h^2\omega^2 + 2\kappa\cosh 2h\omega]}{(1+\kappa)[\kappa \sinh 2h\omega - 2h\omega]}. \tag{A.12}$$

A plot of $C(\omega)$ for a few values of κ is given in Figure A.2. For large and small ω, we see that $C(\omega)$ has the asymptotic form

$$C(\omega) = \frac{2\omega}{1+\kappa} \qquad (|h\omega| \gg 1), \tag{A.13}$$

$$C(\omega) = \frac{1}{h}\left\{\frac{\kappa+1}{2(\kappa-1)} - \frac{(\kappa-2)(\kappa-3)}{3(\kappa-1)^2}(h\omega)^2 + \cdots\right\} \qquad (|h\omega| \ll 1). \tag{A.14}$$

A.2. Evolution for an arbitrary initial condition. The fact that $\eta(x) = -C(\omega)g(x)$ when g varies harmonically allows us to use the Fourier transform to solve the grain boundary diffusion problem for an arbitrary initial condition $g(x, t = 0)$. Note that the solution to $g_t = -\eta_{xx}$ with $g(x, t = 0) = \cos\omega x$ is given by

$$g(x,t) = e^{-C(\omega)\omega^2 t}\cos\omega x. \tag{A.15}$$

This gives the time evolution of each Fourier mode. If we write g at $t = 0$ as

$$g(x,0) = \int_{-\infty}^{\infty} e^{i\omega x}\hat{g}(\omega,0)\,d\omega, \tag{A.16}$$

then at any later time g will be

$$g(x,t) = \int_{-\infty}^{\infty} e^{i\omega x}\hat{g}(\omega,t)\,d\omega = \int_{-\infty}^{\infty} e^{i\omega x}e^{-C(\omega)\omega^2 t}\hat{g}(\omega,0)\,d\omega. \tag{A.17}$$

It is instructive to compare the dissipation rate $C(\omega)\omega^2$ to that for the heat equation and the linearized surface diffusion equation:

Dissipation Rate	Equation
$b\omega^2$	$u_t = bu_{xx} \quad (b > 0)$
$C(\omega)\omega^2$	$g_t = -\eta_{xx}$
$b\omega^4$	$u_t = -bu_{xxxx} \quad (b > 0)$

From Figure A.2 and (A.12)–(A.14), we see that low-frequency modes decay like the heat equation with $b = (\kappa+1)[2h(\kappa-1)]^{-1}$, whereas high-frequency modes decay as $\exp(-b|\omega|^3)$ with $b = 2(1+\kappa)^{-1}$, which is halfway between the heat equation and the linearized surface diffusion equation.

REFERENCES

[1] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, The Netherlands, 1976.

[2] I. A. Blech, *Electromigration in thin aluminum films on titanium nitride*, J. Appl. Phys., 47 (1976), pp. 1203–1208.

[3] I. A. Blech and C. Herring, *Stress generation by electromigration*, Appl. Phys. Lett., 29 (1976), pp. 131–133.

[4] A. F. Bower and D. Craft, *Analysis of failure mechanisms in the interconnect lines of microelectronic circuits*, Fatigue Fract. Engrg. Mater. Struct., 21 (1998), pp. 611–630.

[5] P. G. Ciarlet, *Mathematical Elasticity*, Vol. 1, North–Holland, Amsterdam, 1993.

[6] A. C. F. Cocks and S. P. A. Gill, *A variational approach to two dimensional grain growth I. Theory*, Acta Mater., 44 (1996), pp. 4765–4775.

[7] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, Krieger, Malabar, FL, 1984.

[8] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I, Wiley Interscience, New York, 1989.

[9] V. B. Fiks, *Mechanism of ion mobility in metals*, Soviet Physics—Solid State, 1 (1959), pp. 14–28.

[10] D. Fridline, *Finite Element Modeling of Electromigration and Stress Voiding in Microelectronic Interconnects*, Ph.D. thesis, Brown University, Providence, RI, 2001.

[11] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr. 18, AMS, Providence, RI, 1969.

[12] P. Grisvard, *Singularities in Boundary Value Problems*, Res. Notes Appl. Math. 22, Masson, Paris, 1992.

[13] E. Hille and R. S. Phillips, *Functional Analysis and Semi-groups*, AMS, Providence, RI, 1957.

[14] P. S. Ho and T. Kwok, *Electromigration in metals*, Rep. Prog. Phys., 52 (1989), pp. 301–348.

[15] H. Huntington and A. Grone, *Current-induced marker motion in gold wires*, J. Phys. Chemistry Solids, 20 (1961), pp. 76–87.

[16] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.

[17] R. Kirchheim, *Stress and electromigration in Al-lines of integrated circuits*, Acta Metal. Mater., 40 (1992), pp. 309–323.

[18] M. A. Korhonen, P. Borgesen, K. N. Tu, and C.-Y. Li, *Stress evolution due to electromigration in confined metal lines*, J. Appl. Phys., 73 (1993), pp. 3790–3799.

[19] V. A. Kozlov, V. G. Maz'ya, and J. Rossmann, *Elliptic Boundary Value Problems in Domains with Point Singularities*, AMS, Providence, RI, 1997.

[20] L. D. Landau and E. M. Lifshitz, *Theory of Elasticity*, 3rd ed., Butterworth–Heinemann, Oxford, UK, 1986.

[21] W. W. Mullins, *Mass transport at interfaces in single component systems*, Metal. Mater. Trans. A, 26 (1995), pp. 1917–1929.

[22] N. I. Muskhelishvili, *Some Basic Problems of the Mathematical Theory of Elasticity*, 2nd English ed., P. Noordhoff, Groningen, The Netherlands, 1963.

[23] N. I. Muskhelishvili, *Singular Integral Equations*, 2nd ed., Dover, New York, 1992.

[24] B. A. Plamenevskij, *Elliptic boundary value problems in domains with piecewise smooth boundary*, in Partial Differential Equations 9, Encyclopaedia Math. Sci. 79, M. S. Agranovich, Y. V. Egorov, and M. A. Shubin, eds., Springer-Verlag, Berlin, 1997.

[25] M. E. Sarychev, Y. V. Zhitnikov, L. Borucki, C. L. Liu, and T. M. Makhviladze, *General model for mechanical stress evolution during electromigration*, J. Appl. Phys., 86 (1999), pp. 3068–3075.

[26] J. A. Sethian and J. Wilkening, *A numerical model of stress driven grain boundary diffusion*, J. Comput. Phys., 193 (2003), pp. 275–305.
[27] R. S. Sorbello, *Theory of electromigration*, in Solid State Physics, Vol. 51, H. Ehrenreich and F. Spaepen, eds., Academic Press, New York, 1997, pp. 159–231.
[28] B. G. Streetman and S. Banerjee, *Solid State Electronic Devices*, Prentice–Hall, Englewood Cliffs, NJ, 2000.
[29] K.-N. Tu, J. W. Mayer, and L. C. Feldman, *Electronic Thin Film Science for Electrical Engineers and Materials Scientists*, Macmillan, New York, 1992.
[30] J. Wilkening, *Mathematical Analysis and Numerical Simulation of Electromigration*, Ph.D. thesis, University of California, Berkeley, 2002.
[31] J. Wilkening, L. Borucki, and J. A. Sethian, *Analysis of stress-driven grain boundary diffusion. Part* II*: Degeneracy*, SIAM J. Appl. Math., 64 (2004), pp. 1864–1886.
[32] K. Yosida, *Functional Analysis*, Springer-Verlag, Heidelberg, 1980.
[33] P. V. Zant, *Microchip Fabrication: A Practical Guide to Semiconductor Processing*, McGraw–Hill, New York, 2000.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 1864–1886

ANALYSIS OF STRESS-DRIVEN GRAIN BOUNDARY DIFFUSION. PART II: DEGENERACY*

JON WILKENING†, LEN BORUCKI‡, AND J. A. SETHIAN§

Abstract. The stress-driven grain boundary diffusion problem is a continuum model of mass transport phenomena in microelectronic circuits due to high current densities (electromigration) and gradients in normal stress along grain boundaries. The model involves coupling many different equations and phenomena, and difficulties such as nonlocality, complex geometry, and singularities in the stress tensor have left open such mathematical questions as existence of solutions and compatibility of boundary conditions. In this paper and its companion, we address these issues and establish a firm mathematical foundation for this problem.

We study the properties of a type of Dirichlet-to-Neumann map that involves solving the Lamé equations with interesting interface boundary conditions. We identify a new class of degenerate grain boundary networks that lead to unsuppressed linear growth modes that are suggestive of continental drift in plate tectonics. We use techniques from semigroup theory to prove that the problem is well posed and that the stress field relaxes to a steady state distribution which may or may not completely balance the electromigration force. In the latter (degenerate) case, the displacements continue to grow without bound along stress-free modes.

Key words. grain boundary, diffusion, electromigration, elasticity, semigroups

AMS subject classifications. 35Q72, 47D03, 74F99

DOI. 10.1137/S0036139903438247

1. Introduction.

Electromigration is a diffusion process in which high current densities act as a driving force to transport ions in a metallic lattice in the direction of electron flow by transferring momentum through scattering [10]. As microelectronic circuits become smaller and current densities become higher, failure due to electromigration damage in interconnect lines becomes an everincreasing problem in the design of circuits. Grain boundaries, void surfaces, and passivation interfaces are fast diffusion paths along which the diffusion constant typically is seven to eight orders of magnitude higher than in the grains; therefore, most of the mass transport occurs at these locations. The inhomogeneous redistribution of atoms leads to the development of stresses in the line. Stress gradients along grain boundaries and surface tension at void surfaces both contribute to the flux of atoms, usually suppressing electromigration and increasing the lifetime of the line.

Many experimental, theoretical, and numerical studies have been done to investigate the role of various combinations of electromigration, stress gradients, surface

*Received by the editors December 6, 2002; accepted for publication (in revised form) November 26, 2003; published electronically August 4, 2004.
http://www.siam.org/journals/siap/64-6/43824.html

†Courant Institute of Mathematical Sciences, New York, NY 10012 (wilken@cims.nyu.edu). The research of this author was supported in part by a Department of Energy Computational Science Graduate Student Fellowship while the author was at U.C. Berkeley; by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract DE-AC03-76SF00098; and by the National Science Foundation through grant DMS-0101439.

‡Motorola, Inc., Tempe, AZ 85284 (LenBorucki@intelligentplanar.com). The research of this author was supported in part by the Division of Mathematical Sciences of the National Science Foundation, University-Industry Program.

§Department of Mathematics and Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94721 (sethian@math.berkeley.edu). The research of this author was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract DE-AC03-76SF00098 and by the Division of Mathematical Sciences of the National Science Foundation.

diffusion, temperature, and anisotropy on the transport of atoms in the bulk grains, along void surfaces, along grain boundaries, and at passivation interfaces. We refer the reader to the companion paper [13] for a discussion of this literature.

The goal of this paper and its companion [13] is to provide a rigorous treatment of a modest subset of the mass transport phenomena that occur in microelectronic circuits. In particular, our model is two dimensional and neglects void evolution, curvature-driven grain boundary motion, plastic deformation, and thermal effects. Instead, we focus on the coupling of electromigration to stress generation, which is difficult due to nonlocality, stiffness, complex geometry, and stress singularities at junctions where boundary conditions involving normal stress are imposed.

In [13], we describe the problem physically, state the equations and boundary conditions, find an exact solution for an infinite interconnect line, recast the problem for a finite geometry as an ordinary differential equation on a Hilbert space involving two unbounded operators L and S, analyze the operator L, and prove that the problem is well posed (using techniques from semigroup theory) under the simplifying assumption that the grain boundary network is nondegenerate. We summarize many of these results in section 2.

In section 3, we prove that S (a type of Dirichlet-to-Neumann map) is self-adjoint, negative, closed, and densely defined. These properties are stated (omitting proofs) in [13] and play an essential role in our analysis of the nondegenerate and general cases. To define S, we study weak solutions to the grain boundary normal stress problem. This leads us to identify a new class of degenerate grain boundary networks for which S has a nontrivial (but always finite dimensional) kernel. We use an energy argument to prove self-adjointness and negativity, and we present a counting argument that is useful for characterizing degeneracy.

In section 4, we prove that the equation governing the evolution of normal stress on the grain boundary network Γ generates an analytic semigroup of bounded linear operators on $H^1(\Gamma)$. We also show how to use this semigroup to determine the evolution of displacement and stress inside each grain, which is a nontrivial task since the grain boundary normal stress problem is not uniquely solvable in the degenerate case without additional information about the jump in displacement across Γ. We show that the stress field relaxes to a steady state distribution which may or may not completely balance the electromigration force along grain boundaries. In the latter (degenerate) case, the displacement field describing the motion of the grains continues to grow without bound along stress-free modes, leading to behavior that resembles continental drift in plate tectonics.

We remark that such growth modes are quite harmless to an interconnect line. They correspond to a gradual transport of material from one side of each participating grain to the other, causing it to continually drift to avoid misfit with its neighbors, but not leading to stress generation or voiding.

2. Preliminaries. We model the interconnect line as a union $\Omega = \bigcup_{k=1}^{M} \Omega_k$ of disjoint polygonal grains, as shown in Figure 2.1. We denote the outer boundary (the "walls") of the domain by Γ_0, and we denote the grain boundary network by $\Gamma = \bigcup_{j=1}^{N} \Gamma_j$. Each line segment Γ_j is given an arbitrary orientation (a unit tangent vector $\mathbf{t}_j$) and an arc length parameter s which increases in the $\mathbf{t}_j$ direction. The unit normal $\mathbf{n}_j$ points from right ($-$) to left ($+$) facing along $\mathbf{t}_j$. The net grain growth g is defined on Γ as the jump in normal component of displacement across the grain

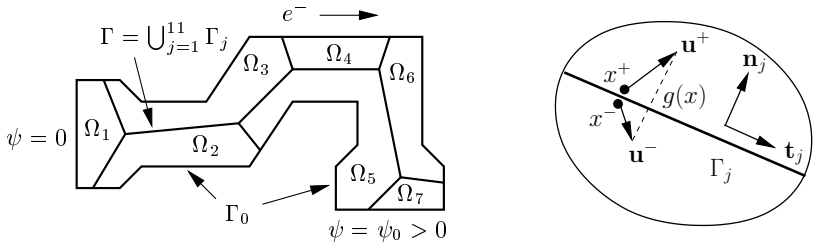


FIG. 2.1. *Left: geometry of an interconnect line. Right: $g(x)$ is the jump in normal component of displacement across Γ at x.*

boundary:

$$g(x) := [\mathbf{u}(x^+) - \mathbf{u}(x^-)] \cdot \mathbf{n}_j \qquad (x \in \Gamma_j). \tag{2.1}$$

It represents the distance the original grains have separated to accommodate the new material that occupies that space; see Figure 2.1. Note that g evolves as a function defined on Γ as $\mathbf{u}$ evolves on Ω; both Γ and Ω remain *fixed* in the reference configuration. The sign of g is independent of the orientation chosen for the segment.

The electric potential ψ is found by solving the Laplace equation subject to the boundary conditions $\psi = 0$ and $\psi = \psi_0$ at the two ends of the interconnect line and $\partial_n \psi = 0$ on all other walls. We assume the grain boundaries do not significantly affect the flow of electrons in the line, so boundary conditions are specified along Γ_0 only; Γ is invisible to ψ.

Each grain is assumed to deform elastically (assuming plane strain) and to satisfy the Lamé equations of linearized elasticity, $\mu \Delta \mathbf{u} + (\lambda + \mu) \nabla (\nabla \cdot \mathbf{u}) = \mathbf{0}$. The outer walls are assumed to be perfectly rigid, giving the two boundary conditions

$$\mathbf{u}(x) = \mathbf{0} \qquad (x \in \Gamma_0). \tag{2.2}$$

Along grain boundaries, four interface boundary conditions are specified:

$$\mathbf{u}(x^+) - \mathbf{u}(x^-) = g(x)\mathbf{n_j} \qquad (x \in \Gamma_j), \tag{2.3}$$

$$\sigma(x^+)\mathbf{n_j} = \sigma(x^-)\mathbf{n_j} \qquad (x \in \Gamma_j). \tag{2.4}$$

In other words, grains are not allowed to slide tangentially, the jump in normal component of displacement is specified to be $g(x)$, and both components of traction balance across the grain boundary. The traction condition is justified because we have adopted an Eulerian viewpoint for the meaning of displacement; see [13]. For a given deformation φ, $\mathbf{u}(x)$ is defined to be $x - \varphi^{-1}(x)$ rather than $\varphi(x) - x$. As a result, $g(x)\mathbf{n}_j = \varphi^{-1}(x^-) - \varphi^{-1}(x^+)$ rather than $\varphi(x^+) - \varphi(x^-)$ (the latter is shown in Figure 2.1). The material and Eulerian viewpoints have the same linearization.

After nondimensionalizing [13], the flux J of atoms along the grain boundary is given by $J = \partial_s(\eta + \psi)$. Here $\eta = \mathbf{n} \cdot \sigma \mathbf{n}$ is the normal stress along the grain boundary, $\psi = \psi\big|_\Gamma$ is the restriction of the electric potential to the grain boundary, and ∂_s is the derivative with respect to arc length along the grain boundary. The continuity equation expressing mass conservation is $\partial_t g + \partial_s J = 0$; hence the evolution of net grain growth is governed by

$$\partial_t g = -\partial_s^2(\eta + \psi) = L(Sg + \psi). \tag{2.5}$$

Here $L = -\frac{\partial^2}{\partial s^2}$ is the negative of the second derivative operator with respect to arc length on each grain boundary segment, and S maps a displacement jump g defined on Γ to the corresponding normal stress η on Γ by solving the Lamé equations, as discussed above. If we apply the operator S to both sides of (2.5), we obtain a differential equation for η:

$$\partial_t \eta = SL(\eta + \psi). \tag{2.6}$$

The solution to this equation is given by $\eta(t) = E_t(\eta + \psi) - \psi$, where $\{E_t : t \geq 0\}$ is the strongly continuous semigroup of linear operators generated by SL; see section 4. The solution to (2.5) is more complicated; it will be discussed in section 4.3.

Boundary conditions for chemical potential continuity and flux balance at junctions are enforced by requiring that $\eta + \psi$ belongs to the domain $\mathcal{D}(L)$ for $t > 0$; see [13]. Similarly, to ensure that g is actually a displacement jump, i.e., that there exists a displacement field $\mathbf{u}$ on Ω satisfying (2.3), we require that $g \in \mathcal{D}(S)$ for $t > 0$.

The domain $\mathcal{D}(S)$ is difficult to characterize; see section 3.3. To describe the domain $\mathcal{D}(L)$, it is useful to establish further notation. Let $C(\Gamma)$ denote the space of continuous functions on Γ, and let $\widetilde{C}(\Gamma)$ denote the space of functions f continuous on the interiors of the Γ_j with well-defined limits $f(x_i^j)$ at the endpoints x_i of Γ_j but with possibly different limiting values at x_i when approached from different segments. Differentiation is defined segment by segment, where we recall that each segment is given an arbitrary orientation along which the arc length parameter increases. We define

$$\widetilde{C}^r(\Gamma) = \{f : f^{(k)} \in \widetilde{C}(\Gamma),\ 0 \leq k \leq r\}. \tag{2.7}$$

Then the domain $\mathcal{D}(L)$ satisfies

$$\{f \in \widetilde{C}^2(\Gamma) : f \text{ satisfies } (*)\} \ \subset\ \mathcal{D}(L)\ \subset\ \{f \in \widetilde{C}^1(\Gamma) : f \text{ satisfies } (*)\}, \tag{2.8}$$

where $(*)$ refers to continuity and flux boundary conditions at all junctions. In other words, $f \in \widetilde{C}^r(\Gamma)$ satisfies $(*)$ if $f \in C(\Gamma)$ and at each junction x_i, $\sum_j \pm f'(x_i^j) = 0$, where the sum is over segments incident to x_i and the sign depends on whether the segment is parameterized toward or away from the junction.

Other key properties of L (all proved in [13]) are as follows. L is self-adjoint and positive. Its kernel consists of the functions

$$e_k(x) = \begin{cases} |\Gamma_{J_k}|^{-\frac{1}{2}}, & x \in \Gamma_{J_k}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

Here J_k is the set of indices such that $\Gamma_{J_k} := \cup_{j \in J_k} \Gamma_j$ is the kth connected component of Γ (treated as a point set in $\mathbb{R}^2$), and $|\Gamma_{J_k}| = \sum_{j \in J_k} |\Gamma_j|$ is the sum of the lengths of the segments making up component k. Let $d = \dim \ker(L)$, and define

$$P = I - \sum_{k=1}^{d} (\cdot, e_k) e_k, \quad A = L + \sum_{k=1}^{d} (\cdot, e_k) e_k, \quad G = A^{-1} - \sum_{n=1}^{d} (\cdot, e_k) e_k, \tag{2.10}$$

where $(\cdot, \cdot)$ is the L^2 inner product on Γ. Then P is the orthogonal projection onto the subspace

$$\operatorname{ran}(L) = \ker(L)^{\perp} = \left\{ f \in L^2(\Gamma)\ :\ \int_{\Gamma_{J_k}} f\, ds = 0,\ 1 \leq k \leq d \right\}, \tag{2.11}$$

and we have $L = AP = PA$, $LG = P$, and $GL = P\big|_{\mathcal{D}(L)}$. $A^{\frac{1}{2}}$ is an isomorphism from $H^1(\Gamma)$ to $L^2(\Gamma)$, where $H^1(\Gamma)$ consists of all $f \in C(\Gamma)$ which are absolutely continuous with (weak) derivative $f' \in L^2(\Gamma)$. Finally, $\mathcal{D}(L^{\frac{1}{2}}) = H^1(\Gamma)$.

3. Elasticity with interface boundary conditions. In this section we give rigorous definitions of the operators S and B, define weak solutions to the Lamé equations with appropriate interface boundary conditions along grain boundaries, and introduce the notion of degeneracy of a grain boundary network. We prove that S and B are self-adjoint and negative on $L^2(\Gamma)$, that the former is closed and densely defined, and that the latter is compact. We also provide a precise characterization of grain boundary degeneracy that is easy to check numerically.

In the previous section, we described S as a type of Dirichlet-to-Neumann operator that takes a displacement jump g on the grain boundary, solves the Lamé equations subject to the boundary conditions (2.2)–(2.4), and returns the normal stress η on Γ. For technical reasons, it is preferable to define S as the pseudoinverse of B, where B takes a specified normal stress η on the grain boundary, solves the Lamé equations subject to the boundary conditions

$$\mathbf{u}(x) = \mathbf{0} \qquad (x \in \Gamma_0), \tag{3.1}$$

$$[\mathbf{u}(x^+) - \mathbf{u}(x^-)] \cdot \mathbf{t}_j = 0 \qquad (x \in \Gamma_j), \tag{3.2}$$

$$[\sigma(x^+) - \sigma(x^-)]\mathbf{n}_j = \mathbf{0} \qquad (x \in \Gamma_j), \tag{3.3}$$

$$\mathbf{n}_j \cdot \sigma(x)\mathbf{n}_j = \eta(x) \qquad (x \in \Gamma_j), \tag{3.4}$$

and returns the jump in the normal component of displacement

$$(B\eta)(x) = [\mathbf{u}(x^+) - \mathbf{u}(x^-)] \cdot \mathbf{n}_j \qquad (x \in \Gamma_j). \tag{3.5}$$

The primary obstacle to this approach is that in the case of degeneracy, η must satisfy further conditions for a solution $\mathbf{u}$ to exist, and these solutions are not unique. In this case, we define B using appropriate projections so that its pseudoinverse S has the physical meaning described previously.

3.1. Boundary conditions. To impose Dirichlet boundary conditions at walls and no-slip boundary conditions across grain boundaries, we employ a Hilbert subspace H of $H^1(\Omega)^2$ defined as the kernel of appropriate trace operators. Recall that the inner product of the Sobolev space $H^1(\Omega)^2$ is given by

$$(\mathbf{u}, \mathbf{v}) = \int_\Omega (\mathbf{u} \cdot \mathbf{v} + \nabla \mathbf{u} : \nabla \mathbf{v}) dx, \tag{3.6}$$

where $(\nabla \mathbf{u})_{ij} = \partial_j u_i$ and $A : B = \sum_{ij} A_{ij} B_{ij}$. Note that the values of $\mathbf{u}$ in this space do not communicate across grain boundaries—the restriction of $\mathbf{u}$ to each Ω_k can be any function in $H^1(\Omega_k)^2$. The trace operators defined below map, respectively, a vector field $\mathbf{u} \in H^1(\Omega)^2$ to its value at the walls, to its jump in tangential component across grain boundaries, and to its jump in normal component across grain boundaries. We use γ_0 and γ_t to define the Hilbert space H given by

$$H = \left\{ \mathbf{u} \in H^1(\Omega)^2 \;\middle|\; \gamma_0 \mathbf{u} = \mathbf{0},\ \gamma_t \mathbf{u} = 0 \right\}. \tag{3.7}$$

We will need γ_n in section 3.3 to define weak solutions and also to define the operator B. Recall that N and M are the number of grain boundary segments and the

number of regions, respectively:

$$\Omega = \bigcup_{k=1}^{M} \Omega_k, \qquad \Gamma = \bigcup_{j=1}^{N} \Gamma_j, \qquad \Gamma_0 = \text{outer walls.} \tag{3.8}$$

THEOREM 3.1. *The following trace operators are compact:*

$$\gamma_0 : H^1(\Omega)^2 \to L^2(\Gamma_0)^2 : \mathbf{u} \mapsto \mathbf{u}\big|_{\Gamma_0}, \tag{3.9}$$

$$\gamma_t : H^1(\Omega)^2 \to L^2(\Gamma) : \mathbf{u} \mapsto \left([\mathbf{u}\big|_{\Gamma_1^+} - \mathbf{u}\big|_{\Gamma_1^-}] \cdot \mathbf{t}_1 \, , \, \ldots \, , \, [\mathbf{u}\big|_{\Gamma_N^+} - \mathbf{u}\big|_{\Gamma_N^-}] \cdot \mathbf{t}_N \right), \tag{3.10}$$

$$\gamma_n : H^1(\Omega)^2 \to L^2(\Gamma) : \mathbf{u} \mapsto \left([\mathbf{u}\big|_{\Gamma_1^+} - \mathbf{u}\big|_{\Gamma_1^-}] \cdot \mathbf{n}_1 \, , \, \ldots \, , \, [\mathbf{u}\big|_{\Gamma_N^+} - \mathbf{u}\big|_{\Gamma_N^-}] \cdot \mathbf{n}_N \right). \tag{3.11}$$

Here $\mathbf{u}\big|_{\Gamma_j^+}$ *is the trace of* $\mathbf{u}$ *on* Γ_j *from the left (i.e., the trace of* $\mathbf{u}\big|_{\Omega_k}$ *on* Γ_j*, where* Ω_k *lies to the left of* Γ_j*),* $\mathbf{u}\big|_{\Gamma_j^-}$ *is the trace of* $\mathbf{u}$ *on* Γ_j *from the right, and we have identified* $L^2(\Gamma)$ *with* $L^2(\Gamma_1) \times \cdots \times L^2(\Gamma_N)$.

Proof. Since Γ_0 is also a union of line segments, it suffices to show that for any region Ω_k and boundary segment $X \subset \partial\Omega_k$, the composite map

$$\mathbf{u} \mapsto \mathbf{u}\big|_{\Omega_k} \mapsto \left(\mathbf{u}\big|_{\Omega_k}\right)\big|_{\partial\Omega_k} \mapsto \left(\mathbf{u}\big|_{\Omega_k}\right)\big|_X \tag{3.12}$$

is compact. The first and last maps are just restriction operators from $H^1(\Omega)^2$ to $H^1(\Omega_k)^2$ and $L^2(\partial\Omega_k)^2$ to $L^2(X)^2$, so they are bounded. Since Ω_k is a polygon, it has a Lipschitz boundary, and hence [1, 4, 5] the trace operator

$$\gamma_k : H^1(\Omega_k)^2 \to H^{\frac{1}{2}}(\partial\Omega_k)^2 \tag{3.13}$$

is bounded. But $H^{\frac{1}{2}}(\partial\Omega_k)^2$ is compactly embedded in $L^2(\partial\Omega_k)^2$, so the middle map in (3.12) is compact, as required. ∎

3.2. Degenerate grain boundary networks. In this section we define the notion of grain boundary degeneracy, which characterizes the existence of unsuppressed growth modes consisting of stress-free infinitesimal rigid body motions in each grain. We also describe an algorithm for determining whether a given grain boundary network is degenerate and, if it is, for finding these modes. We present a counting argument that strongly suggests that grain geometries with convex grains and very few quadruple or higher order junctions will be nondegenerate. We verify numerically that randomly generated grain boundary networks with convex grains are indeed nondegenerate.

DEFINITION 3.2. *A grain boundary network* Γ *is said to be* degenerate *if* H *contains a nonzero function* $\mathbf{u}$ *consisting of infinitesimal rigid body motions (defined below) on each grain.*

To gain geometric insight, we construct a procedure for testing the degeneracy of a given grain boundary network. An infinitesimal rigid body motion is a displacement field of the form

$$u_1(x, y) = a - cy, \qquad u_2(x, y) = b + cx, \tag{3.14}$$

where a, b, c are arbitrary real numbers. Let (x_k, y_k) be some fixed point in Ω_k, and let r_k be a characteristic length scale for the kth grain. For any $\mathbf{u} \in H(\Omega)^2$ consisting

of infinitesimal rigid body motions, there are parameters a_k, b_k, and c_k for each region such that

$$\mathbf{u}\big|_{\Omega_k} = \left(a_k - c_k \frac{y - y_k}{r_k},\, b_k + c_k \frac{x - x_k}{r_k}\right). \tag{3.15}$$

We wish to determine if there is a nontrivial choice of these parameters such that the corresponding $\mathbf{u}$ belongs to H. We define the vector w by

$$w = (a_1, b_1, c_1, \ldots, a_M, b_M, c_M)^T \tag{3.16}$$

and construct a matrix A with $3M$ columns such that

$$Aw = 0 \;\Leftrightarrow\; \mathbf{u}_w \in H. \tag{3.17}$$

Clearly, if Ω_k touches an outer wall, then the condition that $\gamma_0 \mathbf{u} = 0$ requires that $a_k = b_k = c_k = 0$. We assume that the M_0 outer grains appear first in the list, and we let

$$A_{ij} = \delta_{ij} \qquad (1 \le i \le 3M_0,\, 1 \le j \le 3M) \tag{3.18}$$

so that A has the block structure

$$A = \begin{pmatrix} I & 0 \\ * & A' \end{pmatrix}. \tag{3.19}$$

Note that A is injective iff A' is injective. For each edge that borders a grain with index $k > M_0$, we add a row to A to impose the condition that $\gamma_t \mathbf{u} = 0$. Explicitly, letting (x, y) be any point on edge j and denoting the left and right grains by l and k, the row added to A enforces the equation

$$\left(a_l - c_l \frac{y - y_l}{r_l} - a_k + c_k \frac{y - y_k}{r_k},\, b_l + c_l \frac{x - x_l}{r_l} - b_k - c_k \frac{x - x_k}{r_k}\right) \cdot \mathbf{t}_j \;=\; 0, \tag{3.20}$$

which is clearly linear in the components of w. Note that adding $\alpha \mathbf{t}_j$ to (x, y) does not affect the validity of this equation, so if it holds for one point on the edge, it holds at all points on the edge.

The number of edges that contribute an equation to A can be computed as follows. Let $\Omega' = \cup_{k=M_0+1}^{M} \Omega_k$, and consider the planar graph $\Gamma' = \partial\Omega'$. For example, in Figure 3.1(a), Γ' consists of the eight segments bordering unshaded regions. The Euler relation

$$n + f - e = 1 \qquad (\Gamma' \text{ connected}) \tag{3.21}$$

gives the relationship between the number of vertices, regions, and edges of this graph, where we use 1 instead of 2 since we do not count the unbounded region. If $\overline{\Omega'}$ is multiply connected, this formula holds for each of the c connected subgraphs of Γ', so (3.21) should be modified to read

$$n + f - e = c \qquad (\Gamma' \text{ has } c \text{ connected components}). \tag{3.22}$$

Let n_p be the number of vertices with p incident edges. Then since each edge has two endpoints, we have

$$\sum_{p=2}^{\infty} p n_p = 2e. \tag{3.23}$$

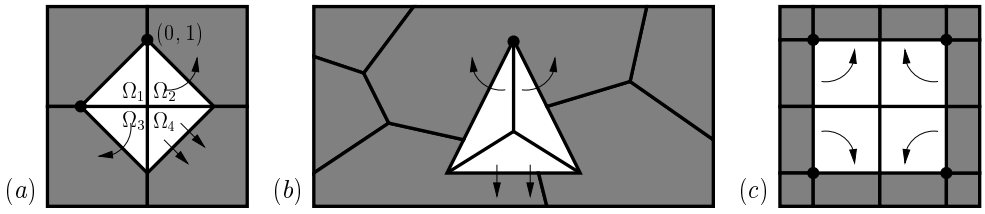


FIG. 3.1. *Examples of degenerate grain boundary networks.* (a) Ω' *has four regions and eight edges, so* A' *has* $12-8=4$ *more columns than rows. This also could have been obtained using* (3.24) *with* $n_3=4, n_4=1$. *The arrows represent one of the functions* $\mathbf{u}$ *in the four dimensional space* H_d, *namely,* $\mathbf{u}\big|_{shaded}=\mathbf{u}\big|_{\Omega_1}=\mathbf{0}$, $\mathbf{u}\big|_{\Omega_2}=(1-y,x)$, $\mathbf{u}\big|_{\Omega_3}=(y,-1-x)$, $\mathbf{u}\big|_{\Omega_4}=(1,-1)$. (b) A' *has six rows and nine columns. Note that nonconvexity in the outer grains allows all nodes of* Ω' *to have* $p=3$, *whereas normally* $n_2\geq 3$. (c) *This time* A' *has the same number of rows and columns, yet it still has a one dimensional kernel since the sign pattern* $+-+-$ *is periodic at a quadruple junction.*

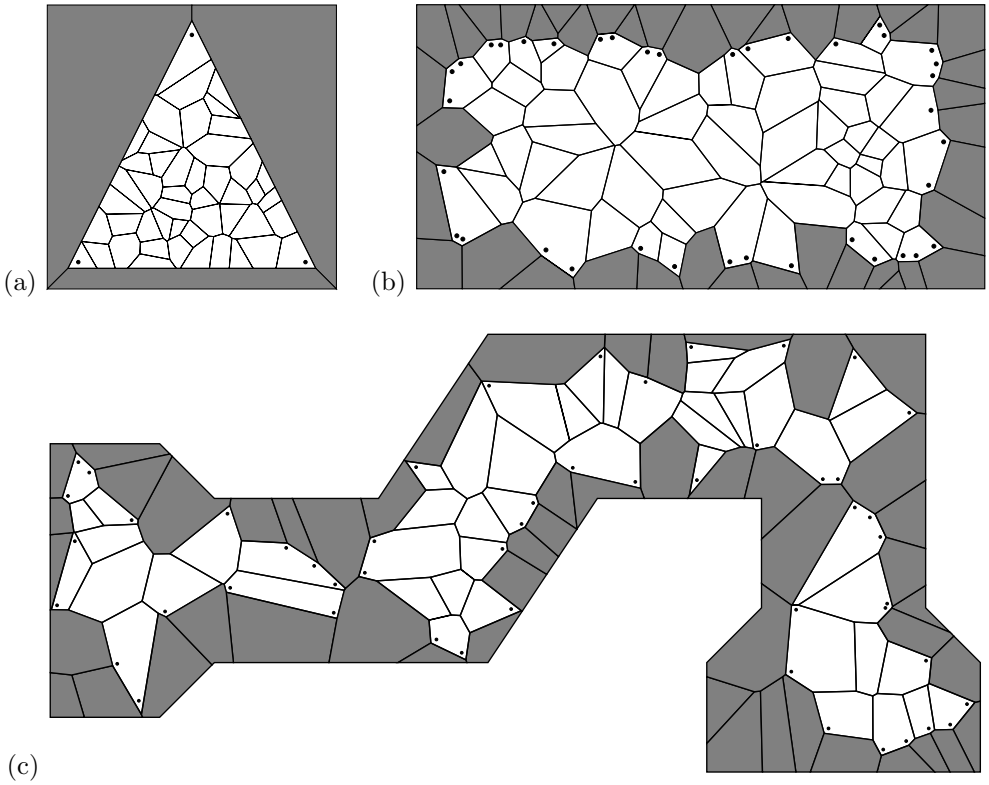


FIG. 3.2. *Typical examples of the geometries generated while computing the condition numbers in Table* 3.1. *Each* • *marks a corner with* $p=2$ *incident edges. All other junctions of* Γ' *have* $p=3$. *Triple junctions often occur clustered together in random Voronoi diagrams, giving the appearance of higher order junctions; this does not give rise to poorly conditioned matrices* A'. *Large shear forces may develop across short edges in such cases if there is not enough redundancy in the other equations (e.g., in case* (a) *here), but that is not a relevant issue when deciding whether a grain boundary network is degenerate. Since at least three corners with* $p=2$ *are needed to traverse the outer boundary of each connected component of* Γ', $e-3f$ *in* (3.24) *is guaranteed to be nonnegative.* (a) *Same number of equations as unknowns* ($e=3f$). (b) $n_2=38$, *so there are* $n_2-3=35$ *more equations than unknowns.* (c) $n_2=51$ *and* $c=3$, *so* $e-3f=42$. *Nonconvexity of grains at outer walls due to reentrant corners clearly will not lead to the difficulties that arose in Figure* 3.1(b).

TABLE 3.1

Condition number of A' for randomly generated Voronoi diagrams. The geometry labels correspond to Figure 3.2*, which shows examples of* (a)-50, (b)-100, *and* (c)-100. *In case* (a)*, we add three outer regions to the Voronoi diagram so that the number of equations is equal to the number of unknowns (a worst-case scenario). Geometries* (b) *and* (c) *are more realistic, although they tend to have more pathologies (such as clustered triple points and short edges) than would likely be found in a real grain boundary network. All the matrices tested were extremely well conditioned, which supports our conjecture that if each grain is convex and no quadruple or higher order junctions occur in the grain boundary network, then Γ is nondegenerate.*

Geometry	Regions	Trials	max	min	mean	std dev
(a)	200	10000	172.6	31.7	37.8	3.2
(a)	100	10000	48.4	21.6	26.2	2.1
(a)	50	10000	35.0	14.4	18.2	1.6
(b)	200	10000	31.8	13.0	16.8	1.3
(b)	100	10000	31.5	8.2	11.1	1.1
(c)	200	10000	26.8	7.4	9.9	1.1
(c)	100	10000	21.4	4.5	6.9	1.1

Multiplying (3.22) by 3 and subtracting (3.23), we obtain

$$e - 3f = \sum (3-p)n_p - 3c \qquad (A' \text{ is an } e \times 3f \text{ matrix}). \tag{3.24}$$

Since each region of Ω' contributes three unknowns and each edge contributes one equation, we see that a necessary condition for the grain boundary network to be nondegenerate is that the right-hand side be nonnegative—otherwise A' will have more columns than rows, and hence a nontrivial kernel. This necessary condition is automatically satisfied if each grain is convex and $n_p = 0$ for $p > 3$, i.e., if we require that all grain boundary junctions be gb-wall or triple junctions: traversing the outer boundary of each of the c components of $\overline{\Omega'}$, we will encounter at least three changes in direction of more than 180 degrees; each of these angles contributes to n_2 since the third segment must prevent nonconvex outer grains (the grains touching walls), and thus $n_2 \geq 3c$. Some examples of degenerate grain boundary networks are shown in Figure 3.1.

CONJECTURE 3.3. *If each Ω_k is convex and no quadruple or higher order junctions occur in the grain boundary network, then Γ is nondegenerate.*

To test this conjecture, we wrote a PERL program to choose M points at random in a polygonal domain U, compute the Voronoi diagram of these points, chop Voronoi regions that cross ∂U, set up the matrix A', and call Matlab to compute its condition number as the ratio of largest to smallest singular value. The points (x_k, y_k) are taken to be the average of the vertices of grain k, and r_k is taken to be $\sqrt{\text{area}_k/\pi}$. The purpose of x_k, y_k, and r_k is to improve the condition number of A' by scaling the effect of c_k to be commensurate with a_k and b_k. The PERL program repeats the above procedure many times (opening a pipe to Matlab at the beginning) and computes the minimum, maximum, mean, and standard deviation of the condition numbers. The results are summarized in Table 3.1. Typical examples of the resulting grain boundary structures are shown in Figure 3.2. Although convex grains do not necessarily arise from Voronoi diagrams, we see no reason that these would not be a good representative sample, especially in light of the fact that *all* the matrices A' that we generated in this way were extremely well conditioned even for grain boundary networks where several triple points had almost coalesced into higher order junctions.

Even if the conjecture is false, this numerical experiment shows that "typical" grain boundary networks are nondegenerate, and we have provided a method for

finding all degeneracies of any grain boundary network (possibly containing nonconvex grains and higher order junctions):

PROCEDURE 3.4 (finding all degeneracies). *Construct the matrix A, find a basis $w_1, \dots, w_q$ for its kernel, and record the corresponding displacements $\mathbf{u}_1, \dots, \mathbf{u}_q$ using* (3.15) *and* (3.16). *These are a basis for the subspace H_d of stress-free (grain by grain) infinitesimal rigid body motions in H.*

DEFINITION 3.5. *A* degenerate grain growth mode *is a function $h \in L^2(\Gamma)$ of the form $h = \gamma_n(\mathbf{u})$ for some $\mathbf{u} \in H_d$. We denote the space of such functions by $\gamma_n(H_d)$.*

Remark 3.6. We will see later that $\ker(S) = \ker(B) = \gamma_n(H_d)$.

LEMMA 3.7. *γ_n is injective on H_d. Thus if $\{\mathbf{u}_k\}_{k=1}^q$ is a basis for H_d, then the functions $h_k = \gamma_n(\mathbf{u}_k)$ form a basis for $\gamma_n(H_d)$.*

Proof. Suppose $\mathbf{u} \in H_d$ and $\gamma_n(\mathbf{u}) = 0$. Then, since $H_d \subset H$, we also have $\gamma_t(\mathbf{u}) = 0$ and $\gamma_0(\mathbf{u}) = 0$. Thus $\mathbf{u}$ is continuous across grain boundaries, is zero on the outer walls, and consists of infinitesimal rigid body motions on each grain. Continuity across grain boundaries implies that the rigid body parameters are the same in each grain, for if l and r index the parameters on either side of a grain boundary segment, then

$$\begin{aligned} a_l - a_r - (c_l - c_r)y &= 0, \\ b_l - b_r + (c_l - c_r)x &= 0 \end{aligned} \tag{3.25}$$

for each (x, y) on the segment. Using two points on the grain boundary, we find $c_l = c_r$, so (3.25) implies $a_l = a_r$ and $b_l = b_r$. Dirichlet conditions at the walls then give that $a = b = c = 0$ in all grains, as required. ∎

Remark 3.8. We may assume the h_k are orthonormal in $L^2(\Gamma)$ (using a Gram–Schmidt procedure, if necessary).

THEOREM 3.9. *Each $h \in \gamma_n(H_d)$ has zero mass on every connected component of the grain boundary network:*

$$\int_{\Gamma_{J_i}} h \, ds = 0 \qquad (i = 1, \dots, d). \tag{3.26}$$

Proof. Let $\mathbf{u} \in H_d$, $h = \gamma_n \mathbf{u}$, and $i \in \{1, \dots, d\}$. Define

$$K_i = \{k : \Omega_k \text{ does not touch } \Gamma_0 \text{ but does touch } \Gamma_{J_i}\}, \tag{3.27}$$

$$J_i' = \{j \in J_i : \Gamma_j \text{ borders an } \Omega_k \text{ which doesn't touch } \Gamma_0\}. \tag{3.28}$$

In Figure 3.3, J_1' contains 12 segment indices, K_1 contains 4 region indices, J_2' contains 6 indices, K_2 contains 3 indices, and J_3' and K_3 are empty. Since $\mathbf{u}$ is an infinitesimal rigid body motion on each grain, it is divergence free, so for $1 \le k \le M$ we have

$$\int_{\partial\Omega_k} \mathbf{u} \cdot \mathbf{n} \, ds = \iint_{\Omega_k} \nabla \cdot \mathbf{u} \, dx = 0 \qquad (\mathbf{n} = \text{outward unit normal}). \tag{3.29}$$

Summing (3.29) over $k \in K_i$ (with an empty sum meaning zero), we obtain

$$0 = \sum_{k \in K_i} \int_{\partial\Omega_k} \mathbf{u} \cdot \mathbf{n} \, ds = \sum_{j \in J_i'} \int_{\Gamma_j} [\mathbf{u}(x^-) - \mathbf{u}(x^+)] \cdot \mathbf{n}_j \, ds = \int_{\Gamma_{J_i}} -h \, ds. \tag{3.30}$$

Here we have used the following facts: $\mathbf{n}_j$ is the unit inward normal on the left $(+)$ grain and the unit outward normal on the right $(-)$ grain; the condition $\gamma_0 \mathbf{u} = \mathbf{0}$

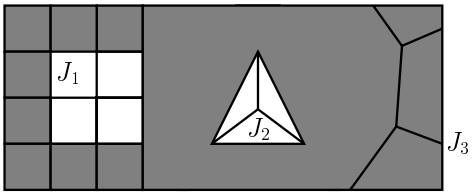


FIG. 3.3. *Shown here, Γ consists of three connected components Γ_{J_i}, so $\dim\ker(L) = 3$. The space H_d (and hence $\ker(S)$) is four dimensional (one degree of freedom in the unshaded structure on the left, three in the middle; see Figure* 3.1*). Each $\mathbf{u} \in H_d$ is zero on the shaded grains.*

implies that $\mathbf{u}$ is zero on any Ω_k touching an outer wall since $\mathbf{u}$ consists of infinitesimal rigid body motions; if $j \in J_i'$, then Γ_j borders either one or two Ω_k with $k \in K_i$—in the former case, $\mathbf{u}(x^{\pm})$ is zero on the other side because that region touches a wall; if $k \in K_i$, all boundary segments Γ_j of Ω_k have $j \in J_i'$; finally, $h = \gamma_n \mathbf{u}$ is zero on the segments Γ_j with $j \in J_i \setminus J_i'$ since both adjacent regions touch a wall. Figure 3.4 illustrates a similar argument in the next section. ∎

3.3. Weak solutions. In this section, we define weak solutions of the grain boundary normal stress and displacement jump problems, and we give rigorous definitions of the operators B and S. Many complications arise in the case of degenerate grain boundaries that make the analysis difficult. We prove that B is compact, self-adjoint, and negative, and we show that $\ker(B) = \gamma_n(H_d)$. The operator S is defined as the pseudoinverse of B, inheriting self-adjointness and negativity.

We will need to make use of the bilinear form

$$a(\mathbf{u}, \mathbf{v}) := \int_\Omega \sigma(\mathbf{u}) : \varepsilon(\mathbf{v})\, dx = \int_\Omega [\lambda\, (\nabla \cdot \mathbf{u})(\nabla \cdot \mathbf{v}) + 2\mu\, \varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v})]\, dx, \tag{3.31}$$

which induces the seminorm $\|\mathbf{u}\|_a = \sqrt{a(\mathbf{u}, \mathbf{u})}$ on $H^1(\Omega)^2$. Here

$$\varepsilon(\mathbf{u})_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i), \qquad \sigma(\mathbf{u}) = \lambda \operatorname{tr} \varepsilon(\mathbf{u}) I + 2\mu\varepsilon(\mathbf{u}), \tag{3.32}$$

and there is clearly a constant C such that

$$\|\mathbf{u}\|_a \le C\|\mathbf{u}\|_{H^1(\Omega)^2} \qquad \left(\mathbf{u} \in H^1(\Omega)^2\right). \tag{3.33}$$

For any $\eta \in L^2(\Gamma)$, we define the linear functional $l_\eta \in H'$ by

$$l_\eta(\mathbf{v}) = -(\eta, \gamma_n \mathbf{v})_{L^2(\Gamma)} = -\int_\Gamma (\eta)(\gamma_n \mathbf{v})\, ds \qquad (\mathbf{v} \in H). \tag{3.34}$$

Note that

$$\|l_\eta\|_{H'} \le \|\gamma_n\|_{\mathcal{L}(H^1(\Omega)^2, L^2(\Gamma))} \|\eta\|_{L^2(\Gamma)}. \tag{3.35}$$

DEFINITION 3.10. *A weak solution to the grain boundary normal stress problem for a given normal stress $\eta \in L^2(\Gamma)$ is a function $\mathbf{u} \in H$ which satisfies*

$$a(\mathbf{u}, \mathbf{v}) = -\int_\Gamma (\eta)(\gamma_n \mathbf{v})\, ds \qquad (\mathbf{v} \in H). \tag{3.36}$$

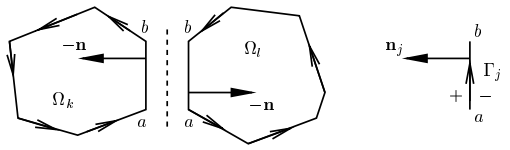


FIG. 3.4. *The contribution of a particular grain boundary segment to* $\sum_k \int_{\partial\Omega_k} \mathbf{v}\big|_{\Omega_k} \cdot \sigma\mathbf{n}$ *contains precisely one term from the left grain and one term from the right grain. Segments on the outer walls do not contribute since* $\gamma_0\mathbf{v} = 0$.

PROPOSITION 3.11. *Any classical solution is a weak solution.*

Proof. Suppose $\mathbf{u}$ is a classical solution with corresponding stress tensor σ and $\mathbf{v} \in H$. Then on Ω_k we have

$$\begin{aligned}(3.37)\qquad \int_{\partial\Omega_k} \mathbf{v}\big|_{\Omega_k} \cdot (\sigma\mathbf{n})\, ds &= \int_{\Omega_k} \partial_j\,(v_i\sigma_{ij})\, dx = \int_{\Omega_k} \overbrace{(\partial_j\sigma_{ij})}^{0}\, v_i + \sigma_{ij}\,(\partial_j v_i)\, dx \\ &= \int_{\Omega_k} \sigma_{ij}\left(\frac{\partial_j v_i + \partial_i v_j}{2}\right) dx = \int_{\Omega_k} \sigma : \varepsilon(\mathbf{v})\, dx = a_k(\mathbf{u}\big|_{\Omega_k},\, \mathbf{v}\big|_{\Omega_k}).\end{aligned}$$

When we sum over all regions, the right-hand side becomes $a(\mathbf{u}, \mathbf{v})$, and the left-hand side becomes a sum over all segments of Γ with one term coming from the left grain and one term coming from the right grain; see Figure 3.4. Since $\gamma_t\mathbf{v} = 0$ and σ is continuous across each segment, the sum of these two terms for the jth segment is

$$(3.38)\qquad \int_{\Gamma_j} (\mathbf{v}^+ \cdot \sigma(-\mathbf{n}_j))\, ds + \int_{\Gamma_j} (\mathbf{v}^- \cdot \sigma\mathbf{n}_j)\, ds = -\int_{\Gamma_j} (\gamma_n\mathbf{v})\mathbf{n}_j \cdot \sigma\mathbf{n}_j\, ds.$$

Summing over all segments and using the boundary condition $\sigma_{nn} = \eta$, we obtain (3.36) as desired. ∎

PROPOSITION 3.12. *If the grain boundary network is degenerate, then a necessary condition for a weak solution to exist is that* $\eta \perp \gamma_n(H_d)$. *If a solution does exist, it is only defined modulo* H_d.

Proof. Fix η, and suppose a solution $\mathbf{u}$ exists. For any $\mathbf{w} \in H_d$, we use (3.36) to conclude that

$$(3.39)\qquad -\int_\Gamma (\eta)(\gamma_n\mathbf{w})\, ds = a(\mathbf{u}, \mathbf{w}) = \int_\Omega \sigma(\mathbf{u}) : \overbrace{\varepsilon(\mathbf{w})}^{0}\, dx = 0.$$

Thus η is orthogonal to $\gamma_n\mathbf{w}$. For any $\mathbf{v} \in H$ we have $a(\mathbf{w}, \mathbf{v}) = 0$, so

$$(3.40)\qquad a(\mathbf{u} + \mathbf{w}, \mathbf{v}) = a(\mathbf{u}, \mathbf{v}) = -\int_\Gamma (\eta)(\gamma_n\mathbf{v})\, ds,$$

and $\mathbf{u} + \mathbf{w}$ is also a weak solution. ∎

DEFINITION 3.13. *We define the space* $\tilde{H}$ *by the relation* $\gamma_n(\tilde{H}) \subset \gamma_n(H_d)^\perp$:

$$(3.41)\qquad \tilde{H} = \{\mathbf{u} \in H \;\big|\; l_h(\mathbf{u}) = 0 \text{ whenever } h \in \gamma_n(H_d)\}.$$

Remark 3.14. Since γ_n is injective on H_d, $\mathbf{u} \equiv 0$ is the only (grain by grain) infinitesimal rigid body motion in $\tilde{H}$, i.e., $\tilde{H} \cap H_d = \{0\}$. Since the codimension of $\tilde{H}$ is at most $q := \dim H_d$ by (3.41), the decomposition $H = \tilde{H} \oplus H_d$ holds.

THEOREM 3.15. *The bilinear form $a(\cdot,\cdot)$ is coercive on $\tilde{H}$, and therefore $\|\cdot\|_a$ is a norm on $\tilde{H}$ equivalent to the one inherited from $H^1(\Omega)^2$.*

Proof. Since $\lambda \geq 0$ and $\mu > 0$, it suffices to show that there is a $c > 0$ such that

$$\|\mathbf{u}\|_\varepsilon^2 := \int_\Omega \varepsilon(\mathbf{u}) : \varepsilon(\mathbf{u})\, dx \geq c \int_\Omega [\mathbf{u}\cdot\mathbf{u} + \nabla\mathbf{u} : \nabla\mathbf{u}]\, dx = c\|\mathbf{u}\|_{\tilde{H}}^2 \qquad \left(\mathbf{u} \in \tilde{H}\right). \tag{3.42}$$

Suppose not. Then there is a sequence of unit vectors $\mathbf{u}_n \in \tilde{H}$ such that

$$\|\mathbf{u}_n\|_\varepsilon^2 \to 0. \tag{3.43}$$

Since the unit ball of $H^1(\Omega)^2$ is compact in $L^2(\Omega)^2$, there is a subsequence which converges in $L^2(\Omega)^2$. Replacing the original sequence with the subsequence, we may assume

$$\|\mathbf{u}_m - \mathbf{u}_n\|_{L^2(\Omega)^2}^2 \to 0 \qquad (m, n \to \infty). \tag{3.44}$$

Since each Ω_k is polygonal, Korn's inequality [4, 3] guarantees the existence of positive constants c_k such that

$$\int_{\Omega_k} \varepsilon(\mathbf{u}) : \varepsilon(\mathbf{u})\, dx + \|\mathbf{u}\|_{L^2(\Omega_k)^2}^2 \geq c_k \|\mathbf{u}\|_{H^1(\Omega_k)^2}^2 \qquad \left(\mathbf{u} \in H^1(\Omega_k)^2\right). \tag{3.45}$$

Letting $c = \min_k c_k$ and summing over all regions, we obtain

$$\|\mathbf{u}\|_\varepsilon^2 + \|\mathbf{u}\|_{L^2(\Omega)^2}^2 \geq c\|\mathbf{u}\|_{H^1(\Omega)^2}^2 \qquad \left(\mathbf{u} \in H^1(\Omega)^2\right). \tag{3.46}$$

Replacing $\mathbf{u}$ by $\mathbf{u}_m - \mathbf{u}_n$ in (3.46) and using (3.43) and (3.44), we find that $\{\mathbf{u}_n\}$ is a Cauchy sequence in $\tilde{H}$. We let $\mathbf{u}^* = \lim_n \mathbf{u}_n$ and note that because $\|\cdot\|_\varepsilon$ is continuous in $H^1(\Omega)^2$, $\|\mathbf{u}^*\|_\varepsilon = \lim_n \|\mathbf{u}_n\|_\varepsilon = 0$. Thus $\varepsilon(\mathbf{u}^*) \equiv 0$. It is straightforward to show [3] that the only solutions to $\varepsilon(\mathbf{u}) \equiv 0$ on a domain U are infinitesimal rigid body motions. Thus $\mathbf{u}^*\big|_{\Omega_k}$ is an infinitesimal rigid body motion for each k, and by Remark 3.14, $\mathbf{u}^* \equiv \mathbf{0}$. But this is impossible since we must also have $\|\mathbf{u}^*\|_{\tilde{H}} = \lim_n \|\mathbf{u}_n\|_{\tilde{H}} = 1$. Therefore $a(\cdot,\cdot)$ is coercive on $\tilde{H}$ as claimed. ∎

PROPOSITION 3.16. *If $\eta \in L^2(\Gamma)$ is nonzero, then $l_\eta \in H'$ is nonzero.*

Proof. We define

$$\mathcal{C} = \{\eta \in \widetilde{C}^1(\Gamma) \cap C(\Gamma) \;\big|\; \eta = 0 \text{ at all junctions}\}, \tag{3.47}$$

where $\widetilde{C}^r(\Gamma)$ was defined in (2.7), and we claim that

$$\mathcal{C} \subset \{\eta \in L^2(\Gamma) \;\big|\; \exists\, \mathbf{v} \in H \text{ s.t. } \eta = \gamma_n \mathbf{v}\}. \tag{3.48}$$

To see this, let $\eta \in \mathcal{C}$, and extend η to $C(\Gamma \cup \Gamma_0)$ by setting $\eta = 0$ on the outer walls. Decompose η into a sum

$$\eta = \sum_k \eta_k, \qquad \eta_k = \frac{1}{2}\eta\big|_{\partial\Omega_k}. \tag{3.49}$$

Since each restriction η_k is C^1 on the (closed) segments of $\partial\Omega_k$ and zero at the corners, the x- and y-components ξ_k^1, ξ_k^2 of $-\eta_k\mathbf{n}$ also have this property ($\mathbf{n}$ is the outward unit normal). This is sufficient to ensure that each ξ_k^i belongs to $H^1(\partial\Omega_k) \subset H^{\frac{1}{2}}(\partial\Omega_k)$, which implies [5] that there are functions $v_k^i \in H^1(\Omega_k)$ whose trace is equal to ξ_k^i

on the boundary. Defining $\mathbf{v} \in H$ grain by grain to have components v_k^1, v_k^2 proves (3.48). Since $\mathcal{C}$ is dense in $L^2(\Gamma)$ and (3.48) holds, we have

$$l_\eta = 0 \quad \Rightarrow \quad \eta \perp \mathcal{C} \quad \Rightarrow \quad \eta = 0 \quad \text{a.e.} \tag{3.50}$$

as desired. ∎

COROLLARY 3.17. *If $\eta \in \gamma_n(H_d)^\perp$ is nonzero, then l_η is nonzero in $\tilde{H}'$.*

Proof. Such an l_η is nonzero when acting on H and is zero on H_d, so it must be nonzero on $\tilde{H}$ due to $H = \tilde{H} \oplus H_d$. ∎

DEFINITION 3.18. *The projection R is defined as the orthogonal projection in $L^2(\Gamma)$ onto $\gamma_n(H_d)^\perp$. Explicitly, we have*

$$R = I - \sum_{k=1}^{q} (\cdot, h_k) h_k, \tag{3.51}$$

where the h_k form an L^2-orthonormal basis for $\gamma_n(H_d)$, as discussed in Remark 3.8.

THEOREM 3.19 (weak solutions). *For any $\eta \in L^2(\Gamma)$, there exists a unique weak solution $\mathbf{u}[\eta] \in \tilde{H}$ to the grain boundary normal stress problem with normal stress $R\eta$ on Γ. There is a constant C independent of η such that*

$$\|\mathbf{u}[\eta]\|_{H^1(\Omega)^2} \leq C \|\eta\|_{L^2(\Gamma)}. \tag{3.52}$$

Moreover, if $\mathbf{u} \equiv 0$, then $\eta \in \gamma_n(H_d)$.

Proof. We produce a candidate solution $\mathbf{u} \in \tilde{H}$ using the Lax–Milgram theorem and the fact that $a(\cdot, \cdot)$ is bounded and coercive on $\tilde{H}$ while $l_{R\eta}$ is a bounded linear functional on $\tilde{H}$. The solution $\mathbf{u}$ is the unique function in $\tilde{H}$ satisfying

$$a(\mathbf{u}, \mathbf{v}) = l_{R\eta}(\mathbf{v}) \qquad (\mathbf{v} \in \tilde{H}), \tag{3.53}$$

which we must show holds for all $\mathbf{v} \in H$. Since $H = \tilde{H} \oplus H_d$, it suffices to check the result for $\mathbf{v} \in H_d$: we have $a(\mathbf{u}, \mathbf{v}) = 0$ since $\mathbf{v}$ is a rigid body motion on each grain, and $l_{R\eta}(\mathbf{v}) = -(R\eta, \gamma_n \mathbf{v}) = 0$ since R projects onto $\gamma_n(H_d)^\perp$. Equation (3.52) follows from coercivity, (3.53), and (3.35):

$$c\|\mathbf{u}\|^2 \leq a(\mathbf{u}, \mathbf{u}) \leq \|l_{R\eta}\| \, \|\mathbf{u}\| \leq \|\gamma_n\| \, \|R\| \, \|\mathbf{u}\| \, \|\eta\|. \tag{3.54}$$

If $\eta \notin \gamma_n(H_d)$, then $R\eta$ satisfies the hypothesis of Corollary 3.17, so $l_{R\eta}$ is nonzero in $\tilde{H}'$. By (3.53), the solution $\mathbf{u}$ cannot be identically zero. ∎

DEFINITION 3.20. *The operator $B : L^2(\Gamma) \to L^2(\Gamma)$ is defined via*

$$B\eta := \gamma_n \mathbf{u}[\eta]. \tag{3.55}$$

Note that in the case of grain boundary degeneracy, $\mathbf{u}[\eta]$ involves a projection of η and a selection criterion for choosing among the nonunique solutions in H.

THEOREM 3.21. *B is compact, self-adjoint, and negative and satisfies*

$$\ker(B) = \gamma_n(H_d). \tag{3.56}$$

Proof. B is compact because $\eta \mapsto \mathbf{u}[\eta]$ is bounded and γ_n is compact. Using (3.36) and the fact that $(\eta, w)_{L^2} = (R\eta, w)_{L^2}$ for $w \in \gamma_n(\tilde{H})$, we have

$$\int_\Gamma \eta_0 B\eta_1 \, ds = \int_\Gamma \eta_0 \gamma_n \mathbf{u}[\eta_1] \, ds = \int_\Gamma (R\eta_0)(\gamma_n \mathbf{u}[\eta_1]) \, ds = -a(\mathbf{u}[\eta_0], \mathbf{u}[\eta_1]). \tag{3.57}$$

Since $a(\cdot,\cdot)$ is symmetric and coercive on $\tilde{H}$, B is self-adjoint and negative:

$$(\eta_0, B\eta_1)_{L^2} = (\eta_1, B\eta_0)_{L^2}, \qquad (\eta_0, B\eta_0)_{L^2} \le 0 \qquad \left(\eta_0, \eta_1 \in L^2(\Gamma)\right). \tag{3.58}$$

Note that $(\eta, B\eta)_{L^2}$ is related to the elastic energy stored in the grains:

$$E = \frac{1}{2} a\left(\mathbf{u}[\eta], \mathbf{u}[\eta]\right) = -\frac{1}{2}\int_\Gamma \eta B\eta \, ds \qquad \left(\eta \in L^2(\Gamma)\right). \tag{3.59}$$

Since $E = 0$ iff $\mathbf{u} \equiv 0$, and $\mathbf{u}[\eta] \equiv 0$ precisely when $\eta \in \gamma_n(H_d)$, we find that $\ker(B) = \gamma_n(H_d)$ as claimed. ∎

DEFINITION 3.22. *The operator S is defined to be the pseudoinverse of B.*

Remark 3.23. Since B is self-adjoint and compact, it has an orthonormal eigendecomposition $B = \sum_1^\infty \beta_k(\cdot, \chi_k)\chi_k$ with $(\beta_1 = \cdots = \beta_q = 0)$ and the remaining β_k forming an increasing sequence of negative numbers converging to zero. S is defined as $S = \sum_1^\infty \alpha_k(\cdot, \chi_k)\chi_k$, where $\alpha_k = 0$ for $k \le q$ and $\alpha_k = 1/\beta_k$ for $k > q$. Since S is defined with respect to an orthonormal basis, we know it is self-adjoint, densely defined, and negative, and its range is $\gamma_n(H_d)^\perp$. Note that the operators S and B satisfy $SB = R$, $BS = R\big|_{\mathcal{D}(S)}$.

DEFINITION 3.24. *A solution of the grain boundary displacement jump problem for a given $g \in \mathcal{D}(S)$ is a solution $\mathbf{u}$ of the normal stress problem with $\eta = Sg$, subject to the additional requirement that $\gamma_n \mathbf{u} = g$.*

THEOREM 3.25. *For any $g \in \mathcal{D}(S)$ there is a unique solution $\mathbf{u}(g)$ of the grain boundary displacement jump problem.*

Proof. Suppose $g \in \mathcal{D}(S)$. Since γ_n is injective on H_d and $\text{range}(I - R) = \gamma_n(H_d)$, there is a unique $\mathbf{u}_d[g] \in H_d$ such that $\gamma_n(\mathbf{u}_d[g]) = (I - R)g$. Clearly

$$\mathbf{u}(g) = \mathbf{u}_d[g] + \mathbf{u}[Sg] \tag{3.60}$$

is the desired solution, where $\mathbf{u}[Sg]$ is the unique solution in $\tilde{H}$ with normal stress Sg specified on Γ; see Proposition 3.12 and note that $\gamma_n(\mathbf{u}(g)) = (I - R)g + BSg = g$. ∎

Remark 3.26. In the degenerate case, the condition $\gamma_n \mathbf{u} = g$ removes the indeterminacy of the solution to the normal stress problem. As a result, the operator S truly maps g to the corresponding normal stress η, whereas B has nonphysical projections built into it for the convenience of being defined on all of $L^2(\Gamma)$.

Remark 3.27. The domain $\mathcal{D}(S)$ is quite complicated due to the variety of ways self-similar solutions of the Lamé equations can behave near grain boundary junctions; see [9, 12]. In particular, even for smooth functions η that are continuous at junctions, $g = B\eta$ generally will be discontinuous at junctions and exhibit infinite slopes. As a result, it would be very difficult to define weak solutions to the grain boundary displacement jump problem directly (without using the grain boundary normal stress problem) and to characterize those g for which the resulting normal stress η is meaningful in the trace sense. The above approach allows us to define S and derive its properties via the compact operator B, which avoids these complications.

4. Dynamics. In this section we show that the equation

$$\eta_t = SL\eta, \qquad \eta(0) = \eta_0, \tag{4.1}$$

generates an analytic semigroup $\{E_t : t \ge 0\}$ of bounded linear operators on $H^1(\Gamma)$. As mentioned previously, the solution $\eta(t)$ when the electromigration force is present

is then given by

$$\eta(t) = E_t(\eta_0 + \psi) - \psi. \tag{4.2}$$

The boundary conditions on $\eta(t) + \psi$ at junctions hold for $t > 0$ as a consequence of the analyticity of E_t and the properties of $\mathcal{D}(SL)$. We will also show that the evolution of grain growth is given by

$$g(t) = R_1 B\eta(t) + (I - R_1)g_0 + [(I - R_1)L\psi]t, \tag{4.3}$$

where R_1 is a projection with kernel of dimension $q = \dim\ker S$ (the degree of degeneracy of the grain boundary network). The term that grows linearly in time corresponds to a continual transport of material around the grains, leading to stress-free rigid body motions in each grain suggestive of continental drift in plate tectonics.

4.1. Semigroup theory. We briefly review the elements of semigroup theory we will need in what follows. A family $\{E_t : t \geq 0\}$ of bounded linear operators on a Banach space X is called a *strongly continuous semigroup* if

$$\begin{aligned} &\text{(i)} && E_{t+s} = E_t E_s \quad (t, s \geq 0), \\ &\text{(ii)} && E_0 = id_X, \\ &\text{(iii)} && t \mapsto E_t x \text{ is continuous on } [0, \infty) \text{ for each fixed } x \in X. \end{aligned} \tag{4.4}$$

If $\|E_t\| \leq 1$ for all $t \geq 0$, $\{E_t\}$ is called a *contraction semigroup*. The *infinitesimal generator* A of a strongly continuous semigroup $\{E_t\}$ is given by

$$Ax = \lim_{h \to 0^+} [E_h x - x]/h \qquad (x \in \mathcal{D}(A)), \tag{4.5}$$

where $\mathcal{D}(A)$ is the set of all $x \in X$ for which the limit exists. It can be proved [2] that $\mathcal{D}(A)$ is dense in X, that A on $\mathcal{D}(A)$ is a closed operator, and that for $x \in \mathcal{D}(A)$, $t \mapsto E_t x$ is continuously differentiable and satisfies

$$\frac{d}{dt} E_t x = AE_t x = E_t Ax \qquad (0 \leq t < \infty). \tag{4.6}$$

The semigroup $\{E_t\}$ is said to be differentiable if $E_t X \subset \mathcal{D}(A)$ for $t > 0$, in which case [2] it is infinitely differentiable and for each $t > 0$ the operators $E_t^{(n)}$ given by

$$E_t^{(n)} x := \frac{d^n}{dt^n} E_t x = A^n E_t x \tag{4.7}$$

are bounded and satisfy

$$E_t^{(n)} x = (E'_{t/n})^n x \qquad (t > 0). \tag{4.8}$$

A differentiable semigroup is said to be analytic if

$$\limsup_{t \to 0} t\|E'_t\| = \alpha < \infty, \tag{4.9}$$

which is equivalent [11] to having a holomorphic extension E_λ given locally by

$$E_\lambda x = \sum_{n=0}^{\infty} \frac{(\lambda - t)^n}{n!} E_t^{(n)} x \qquad \left(t > 0, \quad |\lambda - t| < \frac{t}{\alpha e}, \quad x \in X\right). \tag{4.10}$$

THEOREM 4.1. *If X is a Hilbert space and A is a closed, densely defined, negative, self-adjoint operator on X, then A is the infinitesimal generator of a contraction semigroup $\{E_t\}$ with holomorphic extension $\{E_\lambda : \operatorname{Re}\lambda > 0\}$, and $\alpha \leq e^{-1}$ in* (4.9).

Proof. See [7, 14]. □

4.2. The semigroup generated by SL. The main obstacle to solving (4.1) is that although S and L are each self-adjoint, they do not commute, and hence SL is not self-adjoint. If L were invertible, the obvious thing to do in this situation (see, e.g., [8]) would be to define a new variable $y = L^{\frac{1}{2}}\eta$, use Theorem 4.1 to obtain the solution $y(t)$ of the equation

$$y_t = L^{\frac{1}{2}} S L^{\frac{1}{2}} y, \qquad y(0) = y_0, \tag{4.11}$$

with $y_0 = L^{\frac{1}{2}}\eta_0$, and check that $\eta = L^{-\frac{1}{2}} y$ satisfies (4.1). Since L has a d dimensional kernel, we cannot directly obtain η from y in this way, and we will instead rely on the knowledge that $y(t) - y_0 \in \operatorname{ran}(L^{\frac{1}{2}} S L^{\frac{1}{2}})$ for all time while $\eta(t) - \eta_0 \in \operatorname{ran}(SL)$.

CONVENTION 4.2. *Generic elements of* $\ker(L)$ *and* $\ker(S)$ *will be denoted* e *and* h *so that the notation* $\{e, Gh\}$*, for example, represents the space* $\ker(L) \oplus G\ker(S)$.

Recall from (2.10) and (3.51) that we have defined $d = \dim\{e\}$, $q = \dim\{h\}$, and P and R as the orthogonal projections onto $\{e\}^\perp$ and $\{h\}^\perp$, respectively:

$$P = I - \sum_{k=1}^{d} (\cdot, e_k) e_k, \qquad R = I - \sum_{k=1}^{q} (\cdot, h_k) h_k. \tag{4.12}$$

By Theorem 3.9, we know $\{e\} \perp \{h\}$; hence P and R commute. Moreover, B is injective on $\{e\}$ (and G on $\{h\}$) since $\{e\} \cap \ker(B) = \{e\} \cap \{h\} = \{0\}$.

LEMMA 4.3. *The following identities hold:*

$$\begin{aligned} &\ker(SL) = \{e, Gh\}, \quad \ker(LS) = \{Be, h\}, \quad \ker(L^{\frac{1}{2}} S L^{\frac{1}{2}}) = \{e, G^{\frac{1}{2}} h\}, \\ &\operatorname{ran}(SL) = \{Be, h\}^\perp, \quad \operatorname{ran}(LS) = \{e, Gh\}^\perp, \quad \operatorname{ran}(L^{\frac{1}{2}} S L^{\frac{1}{2}}) = \{e, G^{\frac{1}{2}} h\}^\perp. \end{aligned} \tag{4.13}$$

Proof. SL is densely defined in $L^2(\Gamma)$ since $\mathcal{D}(S)$ is dense, G is bounded with range dense in $\{e\}^\perp$, and $\mathcal{D}(SL) = \{e\} \oplus G\mathcal{D}(S)$. Likewise $\mathcal{D}(LS) = \{h\} \oplus B\mathcal{D}(L)$ and $\mathcal{D}(L^{\frac{1}{2}} S L^{\frac{1}{2}}) = \{e\} \oplus G^{\frac{1}{2}}[\{h\} \oplus B\mathcal{D}(L^{\frac{1}{2}})]$ are dense in $L^2(\Gamma)$. Clearly, $\ker(SL) \supset \{e, Gh\}$. Since $\{h\} \subset \{e\}^\perp$ and LG is the identity on $\{e\}^\perp$, the only vectors mapped to $\{h\}$ by L belong to $\{e, Gh\}$, so the reverse inclusion also holds. A similar argument establishes $\ker(LS) = \{Be, h\}$. For $\ker(L^{\frac{1}{2}} S L^{\frac{1}{2}})$, we use

$$(x, L^{\frac{1}{2}} S L^{\frac{1}{2}} x) = 0 \quad \Leftrightarrow \quad -(|S|^{\frac{1}{2}} L^{\frac{1}{2}} x, |S|^{\frac{1}{2}} L^{\frac{1}{2}} x) = 0 \quad \Leftrightarrow \quad |S|^{\frac{1}{2}} L^{\frac{1}{2}} x = 0 \tag{4.14}$$

and argue as in the other two cases. The result $\operatorname{ran}(SL) \subset \ker(LS)^\perp$ follows from the fact that $(SL)^* = LS$, and $\operatorname{ran}(SL) \supset \{Be, h\}^\perp$ is a consequence of Lemma 4.9 below. Similar arguments give $\operatorname{ran}(LS)$ and $\operatorname{ran}(L^{\frac{1}{2}} S L^{\frac{1}{2}})$. ∎

Remark 4.4. Since $\{e\} \perp \{Gh\}$, $\{Be\} \perp \{h\}$ and $\{e\} \perp \{G^{\frac{1}{2}} h\}$, the kernels in (4.13) all have dimension $d + q = \dim\{e\} + \dim\{h\}$.

DEFINITION 4.5. *We define the (nonorthogonal) projections* P_1*,* R_1*, and* Q *on* $L^2(\Gamma)$ *via*

$$P_1 \text{ projects along } \{e\} \text{ onto } \{Be\}^\perp, \tag{4.15}$$

$$R_1 \text{ projects along } \{h\} \text{ onto } \{Gh\}^\perp, \tag{4.16}$$

$$Q \text{ projects along } \{e, Gh\} = \ker(SL) \text{ onto } \{Be, h\}^\perp = \operatorname{ran}(SL). \tag{4.17}$$

Remark 4.6. In general, if X and Y are finite dimensional subspaces of the same dimension such that $X \cap Y^\perp = \{0\}$, the projection along X onto $Y^\perp$ exists and is

given by $I-(\cdot,w_k)x_k$ (summation is implied). Here $\{x_k\}$ is a basis for X, $\{y_k\}$ is a basis for Y, $w_k=y_j\alpha_{jk}$, and $(x_i,y_j)\alpha_{jk}=\delta_{ik}$. To verify that P_1, R_1, and Q are well defined, we must check the condition $X\cap Y^\perp=\{0\}$.

Suppose $x\in\{e\}\cap\{Be\}^\perp$. Then $(x,Bx)=-(|B|^{\frac12}x,|B|^{\frac12}x)=0$, which implies $x\in\{h\}$. Since $\{e\}\perp\{h\}$, $x=0$ as required. An identical argument works for R_1.

Suppose $x\in\{e,Gh\}\cap\{Be,h\}^\perp$. Then there is $e_0\in\{e\}$ and $h_0\in\{h\}$ such that $x=e_0+Gh_0$. Since $x\perp\{h\}$ and $e_0\perp h_0$, we have $(x,h_0)=(Gh_0,h_0)=0$. Since G is self-adjoint and positive, this implies $h_0\in\{e\}$ so that $x=e_0+0$. But now we have $x\in\{e\}\cap\{Be\}^\perp$, which implies $x=0$ from the above argument.

Remark 4.7. Note that there are $w_k\in\{Be\}$ and $z_k\in\{Gh\}$ such that

$$P_1=I-\sum_{k=1}^{d}(\cdot,w_k)e_k,\qquad R_1=I-\sum_{k=1}^{q}(\cdot,z_k)h_k. \tag{4.18}$$

As a result, in addition to being bounded in $L^2(\Gamma)$, P_1 is also bounded as an operator on $H^1(\Gamma)$ since the e_k belong to this space. On the other hand, the L^2 adjoint $P_1^*=I-(\cdot,e_k)w_k$ is not necessarily defined on $H^1(\Gamma)$ due to the possibility of singularities in the derivative of w_k near junctions. Similarly, R_1^* is a projection in $H^1(\Gamma)$ while R_1 generally is not due to discontinuities in the h_k at junctions.

Remark 4.8. Q may be written $Q=P_1R_1^*$ since $\{e\}\perp\{Gh\}$ and $\{e\}\perp\{h\}$.

LEMMA 4.9. *The following diagrams are commutative in the sense that for each block* $X\underset{g}{\overset{f}{\rightleftarrows}}Y$ *we have* $f\circ g=id_Y$ *and* $g\circ f=id_{\mathcal{D}(f)}$*:*

(4.19)

$$\{e,G^{\frac12}h\}^\perp\underset{G^{\frac12}}{\overset{L^{\frac12}}{\rightleftarrows}}\{e,Gh\}^\perp\underset{R_1}{\overset{R}{\rightleftarrows}}\{e,h\}^\perp\underset{B}{\overset{S}{\rightleftarrows}}\{Be,h\}^\perp\underset{P_1}{\overset{P}{\rightleftarrows}}\{e,h\}^\perp\underset{G^{\frac12}}{\overset{L^{\frac12}}{\rightleftarrows}}\{e,G^{\frac12}h\}^\perp,$$

$$\{Be,h\}^\perp\underset{P_1}{\overset{P}{\rightleftarrows}}\{e,h\}^\perp\underset{G}{\overset{L}{\rightleftarrows}}\{e,Gh\}^\perp\underset{R_1}{\overset{R}{\rightleftarrows}}\{e,h\}^\perp\underset{B}{\overset{S}{\rightleftarrows}}\{Be,h\}^\perp. \tag{4.20}$$

Proof. P and P_1 both project along $\{e\}$, so $PP_1=P$ and $P_1P=P_1$. Since $\{e\}\perp\{h\}$, both projections leave $\{h\}^\perp$ invariant. Since $\mathrm{ran}(P)=\{e\}^\perp$ and $\mathrm{ran}(P_1)=\{Be\}^\perp$, the blocks involving P and P_1 are commutative. Identical arguments may be used for the blocks involving R and R_1.

Note that if $(x,Gh)=0$, then $(Gx,h)=0$; i.e., G maps $\{Gh\}^\perp$ into $\{h\}^\perp$. Since LG is the identity on $\{e\}^\perp$ (recall $\mathcal{D}(L)=\mathrm{ran}(G)\oplus\{e\}$) and GL is the identity on $\{e\}^\perp\cap\mathcal{D}(L)$, the blocks involving L and G are commutative. Identical arguments may be used for the remaining blocks. ∎

DEFINITION 4.10. *We say that* T *is the* pseudoinverse *of the bounded operator* K *on the Hilbert space* H *if there are closed subspaces* X *and* Y *(not necessarily orthogonal) such that* $H=X\oplus Y$, $\ker(T)=X=\ker(K)$, *and*

$$TKy=y\quad(y\in Y),\qquad KTy=y\quad(y\in Y\cap\mathcal{D}(T)). \tag{4.21}$$

In particular, we require $\mathrm{ran}(K)\subset\mathcal{D}(T)$.

LEMMA 4.11. *Such a* T *is closed.*

Proof. First we claim that $\mathrm{ran}(T)=Y$. Clearly, (4.21) implies $\mathrm{ran}(T)\supset Y$. To prove the reverse inclusion, suppose $x_1+y_1=T(x_2+y_2)$ with $x_i\in X$, $y_i\in Y$. Then $Ky_1=KTy_2=y_2$, so $y_1=TKy_1=Ty_2=x_1+y_1$, which implies $x_1=0$ as required.

Now suppose $a_k \to a$, $Ta_k \to b$. We must show $Ta = b$. Note that $b \in Y$ since each $Ta_k \in Y$ and Y is closed. Decompose $a_k = x_k + y_k$ and $a = x + y$ using $H = X \oplus Y$. Then $y_k \to y$ since the projection along X onto Y is continuous. We also know $y_k = KTy_k = KTa_k \to Kb$ since K is continuous. Thus $y = Kb$ and $a = x + y \in \mathcal{D}(T)$. Finally, $Ta = Ty = TKb = b$ since $b \in Y$. ∎

Remark 4.12. If there is an eigenbasis for K, then it is an eigenbasis for both operators, and the eigenvalues are reciprocal or zero. When K is not self-adjoint, this definition differs from the usual definition in linear algebra that T and K should have the same SVD bases (exchanging left and right singular vectors) with reciprocal (or zero) singular values. The current definition is more useful for eigenvalue problems while the usual one is more useful for least squares problems. The definitions coincide when T and K are self-adjoint.

THEOREM 4.13. *The following pseudoinverse relationships hold:*

$$L^{\frac{1}{2}} S L^{\frac{1}{2}} = \operatorname{pinv}(G^{\frac{1}{2}} Q^* B Q G^{\frac{1}{2}}), \tag{4.22}$$

$$SL = \operatorname{pinv}(QGBQ), \tag{4.23}$$

$$LS = \operatorname{pinv}(Q^* B G Q^*). \tag{4.24}$$

Proof. Since $SR = S$ and $LP = L$, the left-to-right compositions in (4.19) and (4.20) are $L^{\frac{1}{2}} S L^{\frac{1}{2}}$ and SL, respectively. Because P_1 leaves $\{h\}^\perp$ invariant, Q and P_1 agree on $\{h\}^\perp$. Likewise Q^* and R_1 agree on $\{e\}^\perp$, so the right-to-left compositions are $G^{\frac{1}{2}} Q^* B Q G^{\frac{1}{2}}$ and QGQ^*B, respectively.

Clearly, $K := G^{\frac{1}{2}} Q^* B Q G^{\frac{1}{2}}$ annihilates $X := \{e, G^{\frac{1}{2}} h\} = \ker(L^{\frac{1}{2}} S L^{\frac{1}{2}})$, and (4.19) ensures that (4.21) holds with $T := L^{\frac{1}{2}} S L^{\frac{1}{2}}$, $Y := \{e, G^{\frac{1}{2}} h\}^\perp$ as required.

The operator QGQ^*B does not have the same kernel as $T := SL$; however, this is easily corrected using $K := QGQ^*BQ$ instead. We then have $\ker(K) = \ker(T) = \{e, Gh\} =: X$ by (4.13) and (4.17). Equation (4.20) implies that (4.21) holds with $Y := \{Be, h\}^\perp$, which complements X in $L^2(\Gamma)$ since Q is a well-defined projection. Finally, by Remark 4.8 and the identities $R_1^* G R_1 = R_1^* G = G R_1$, $P_1^* B P_1 = P_1^* B = B P_1$, it follows that $QGQ^*BQ = QGBQ$. The proof for LS is similar. ∎

LEMMA 4.14. *Equation* (4.11) *generates an analytic contraction semigroup* $\{\tilde{E}_t : t \geq 0\}$ *of bounded linear operators on* $L^2(\Gamma)$. *For each* $t \geq 0$, $P\tilde{E}_t P = \tilde{E}_t P$.

Proof. We showed that $L^{\frac{1}{2}} S L^{\frac{1}{2}}$ has dense domain in the proof of Lemma 4.3. It is closed by Lemma 4.11 and Theorem 4.13, self-adjoint since L and S are self-adjoint, and negative since S is negative:

$$(x, L^{\frac{1}{2}} S L^{\frac{1}{2}} x) = (L^{\frac{1}{2}} x, S L^{\frac{1}{2}} x) \leq 0. \tag{4.25}$$

Theorem 4.1 may therefore be applied to conclude that $L^{\frac{1}{2}} S L^{\frac{1}{2}}$ is the generator of an analytic contraction semigroup $\{\tilde{E}_t : t \geq 0\}$ of bounded linear operators. Since $\operatorname{ran}(L^{\frac{1}{2}} S L^{\frac{1}{2}}) \subset \{e\}^\perp$, for $y_0 \in L^2(\Gamma)$ we have

$$(\tilde{E}_t y_0, e_k) = (\tilde{E}_0 y_0, e_k) + \int_0^t (L^{\frac{1}{2}} S L^{\frac{1}{2}} \tilde{E}_s y_0, e_k)\, ds = (y_0, e_k). \tag{4.26}$$

Hence $\tilde{E}_t$ leaves $\{e\}^\perp$ invariant, and $P\tilde{E}_t P = \tilde{E}_t P$ as claimed. ∎

THEOREM 4.15. *The family* $\{E_t : t \geq 0\}$ *given by*

$$E_t = (I - P_1) + P_1 G^{\frac{1}{2}} \tilde{E}_t L^{\frac{1}{2}} \tag{4.27}$$

is an analytic semigroup in $H^1(\Gamma)$. Its infinitesimal generator is SL.

Proof. Since $G^{\frac12}$ is bounded from L^2 to H^1, $L^{\frac12}$ is bounded from H^1 to L^2, and P_1 is bounded in H^1, there is a $C>0$ such that $\|P_1G^{\frac12}\tilde E_tL^{\frac12}\|_{H^1}\le C\|\tilde E_t\|_{L^2}$; therefore, each E_t is bounded in $H^1(\Gamma)$. Property (4.4)(iii) follows similarly: pick $x\in H^1$, and let $y=L^{\frac12}x$; then we have

$$\|(E_t-E_s)x\|_{H^1}=\|P_1G^{\frac12}(\tilde E_t-\tilde E_s)y\|_{H^1}\le C\|(\tilde E_t-\tilde E_s)y\|_{L^2}\underset{t\to s}{\longrightarrow}0. \tag{4.28}$$

Properties (4.4)(i) and (4.4)(ii) follow immediately from (4.27) and the corresponding properties of $\tilde E$, using the relations

$$P_1P=P_1,\quad PP_1=P,\quad L^{\frac12}=L^{\frac12}P,\quad G^{\frac12}L^{\frac12}=P=L^{\frac12}G^{\frac12},\quad P\tilde E_sP=\tilde E_sP. \tag{4.29}$$

The analyticity may be seen by computing

$$\limsup_{t\to0}t\|E_t'\|_{H^1}\le C\limsup_{t\to0}t\|\tilde E_t'\|_{L^2}<\infty. \tag{4.30}$$

To prove that the generator of E_t is SL, we first note that

$$L^{\frac12}E_t=L^{\frac12}P_1G^{\frac12}\tilde E_tL^{\frac12}=P\tilde E_tL^{\frac12}=\tilde E_tL^{\frac12} \tag{4.31}$$

and compute

$$E_t'=P_1G^{\frac12}\big(L^{\frac12}SL^{\frac12}\tilde E_t\big)L^{\frac12}=P_1PSL^{\frac12}\big(L^{\frac12}E_t\big)=P_1SLE_t=SLE_t. \tag{4.32}$$

∎

Proposition 4.16. *There is a Riesz basis $\{\phi_k\}$ for $H^1(\Gamma)$ and a nonincreasing, unbounded sequence of numbers $\lambda_k\le0$ such that $SL\phi_k=\lambda_k\phi_k$.*

Proof. Since $G^{\frac12}Q^*BQG^{\frac12}$ is self-adjoint and compact, the spectral theorem gives an L^2 orthonormal basis of eigenfunctions $\{\varphi_k\}_{k=1}^\infty$, which by Theorem 4.13 is also an eigenbasis of $L^{\frac12}SL^{\frac12}$: φ_k is either in the kernel of both operators, or it is an eigenfunction of each with reciprocal eigenvalues. Since S is negative, the eigenvalues λ_k of $L^{\frac12}SL^{\frac12}$ satisfy $\lambda_k\le0$. Since $L^{\frac12}SL^{\frac12}$ commutes with P, we may assume the φ_k are also eigenfunctions of P (with eigenvalue 0 or 1). Define

$$\phi_k=\begin{cases}\varphi_k, & P\varphi_k=0,\\ P_1G^{\frac12}\varphi_k & \text{otherwise.}\end{cases} \tag{4.33}$$

In the first case we have $SL\phi_k=0$. In the second, we obtain

$$SLP_1G^{\frac12}\varphi_k=SL^{\frac12}\varphi_k=P_1SL^{\frac12}\varphi_k=P_1G^{\frac12}L^{\frac12}SL^{\frac12}\varphi_k=\lambda_kP_1G^{\frac12}\varphi_k. \tag{4.34}$$

The ϕ_k are related to the φ_k via

$$\phi_k=\big[(I-P)+P_1G^{\frac12}\big]\varphi_k,\qquad \varphi_k=\big[(I-P_1)+L^{\frac12}\big]\phi_k. \tag{4.35}$$

Since $[I-P+P_1G^{\frac12}]$ is bounded from $L^2(\Gamma)$ to $H^1(\Gamma)$ and its inverse $[I-P_1+L^{\frac12}]$ is bounded in the other direction, they are isomorphisms. Thus the ϕ_k form a Riesz basis (a basis equivalent to an orthonormal basis [6]) for $H^1(\Gamma)$ as claimed. ∎

Remark 4.17. Equation (4.27) could also have been written

$$E_t=[I-P+P_1G^{\frac12}]\,\tilde E_t\,[I-P_1+L^{\frac12}]. \tag{4.36}$$

Remark 4.18. For $\eta_0 \in H^1(\Gamma)$, the coefficients in the expansion $\eta_0 = \sum_k a_k \phi_k$ can be determined via

$$a_k = ([I - P_1 + L^{\frac{1}{2}}]\eta_0, \varphi_k)_{L^2} = (\eta_0, \phi_k^*)_{L^2}, \qquad \phi_k^* = [I - P_1^* + L^{\frac{1}{2}}]\varphi_k. \tag{4.37}$$

The ϕ_k^* are eigenfunctions of LS with eigenvalues λ_k since $LS\phi_k^* = [(SL - P_1 SL)^* + LSL^{\frac{1}{2}}]\varphi_k = \lambda_k L^{\frac{1}{2}}\varphi_k = \lambda_k \phi_k^*$. They belong to $L^2(\Gamma)$ but need not belong to $H^1(\Gamma)$ due to possible singularities in $\partial_s \phi_k^*$ at junctions. For $\eta_0 \in H^1(\Gamma)$ the expansions

$$\eta_0 = \sum_{k=1}^{\infty} a_k \phi_k, \qquad E_t \eta_0 = \sum_{k=1}^{\infty} a_k e^{\lambda_k t}\phi_k, \qquad a_k = (\eta_0, \phi_k^*)_{L^2(\Gamma)}, \tag{4.38}$$

hold in $H^1(\Gamma)$. Note that the L^2 norm of ϕ_k^* diverges as $k \to \infty$, but when $\eta_0 \in H^1(\Gamma)$, the inner products a_k in (4.38) do not; they are square summable.

Remark 4.19. The expansions (4.38) lead to a useful numerical method in which the ϕ_k, ϕ_k^*, and λ_k are computed by approximating the pseudoinverse $\mathrm{pinv}(SL) = QGBQ$ using a singularity-capturing least squares finite element method; see [12, 9].

PROPOSITION 4.20. $\lim_{t\to\infty} E_t = I - Q$ *in norm.*

Proof. Recall that $\dim\ker(SL) = d + q$ so that $\lambda_1 = \cdots = \lambda_{d+q} = 0$. Since $\{\phi_k\}_{k=1}^{d+q}$ is a basis for $\mathrm{range}(I - Q)$, we have

$$[E_t - (I - Q)]\eta_0 = \sum_{k=d+q+1}^{\infty} a_k e^{\lambda_k t}\phi_k. \tag{4.39}$$

Since the mapping $\eta_0 \mapsto \langle a_k\rangle_{k=1}^{\infty}$ with $a_k = (\eta_0, \phi_k^*)_{L^2}$ is an isomorphism from $H^1(\Gamma)$ to l^2, there is a constant C such that

$$\|(E_t - I + Q)\eta_0\|_{H^1} \le Ce^{\lambda^* t}\|\eta\|_{H^1} \qquad (t \ge 0,\ \eta_0 \in H^1(\Gamma)) \tag{4.40}$$

with $\lambda^* = \lambda_{d+q+1} < 0$. Thus $\|E_t - (I - Q)\|_{H^1} \le Ce^{\lambda^* t} \to 0$ as $t \to \infty$ as claimed. ∎

Remark 4.21. Since E_t is an operator on $H^1(\Gamma)$ and the formula for the evolution of normal stress is given by

$$\eta(t) = E_t(\eta_0 + \psi) - \psi, \tag{4.41}$$

we should verify that ψ belongs to $H^1(\Gamma)$. This is done in the companion paper [13].

Remark 4.22. Since E_t is analytic, we have $\mathrm{ran}(E_t) \subset \mathcal{D}(SL)$ for all $t > 0$. Therefore $\eta(t)$ in (4.41) has the property that

$$\eta(t) + \psi \in \mathcal{D}(L) \qquad (t > 0). \tag{4.42}$$

Thus although ψ does not necessarily satisfy zero flux boundary conditions at junctions, the normal stress η immediately compensates so that for all $t > 0$, flux balance holds. As long as there is a displacement jump $g(t)$ compatible with the evolution of $\eta(t)$, we have proved that the grain boundary diffusion problem is well posed.

4.3. The evolution of *g*. In the nondegenerate case, the evolution of g is easily determined from the evolution of η in (4.41) via

$$g(t) = B\eta(t) \qquad \text{(nondegenerate case)}. \tag{4.43}$$

The situation becomes much more complicated (and rather interesting) in the degenerate situation where the subspace $\{h\} = \ker(S) = \ker(B)$ is nontrivial. In that case, each function in $\{h\}$ is a growth mode which is not suppressed by the grain boundary diffusion process. If such a mode h is activated by ψ, it will grow linearly in time without bound, as will the corresponding $\mathbf{u} \in H_d$ such that $\gamma_n \mathbf{u} = h$; see Definition 3.5. The nonlinear picture in which $\mathbf{u}$ is replaced by a collection of genuine (as opposed to infinitesimal) rigid body motions on each grain resembles continental drift in plate tectonics, at least superficially. The steady state stress distribution does not fully cancel the flux due to electromigration, and material is continually transported around the participating grains, causing them to drift in order to avoid misfit with their neighbors as material is removed from one side and deposited on the other.

THEOREM 4.23. *The evolution of g is given by*

$$g(t) = R_1 B\eta(t) + (I - R_1)g_0 + [(I - R_1)L\psi]t. \tag{4.44}$$

Proof. Recall that the projections R and R_1 may be written

$$R = I - \sum_{k=1}^{q} (\cdot, h_k)h_k, \quad R_1 = I - \sum_{k=1}^{q} (\cdot, z_k)h_k \quad \big(h_k \in \{h\}, z_k \in \{Gh\}\big). \tag{4.45}$$

Note that $(I - R_1)L = \sum_{k=1}^{q} (\cdot, Lz_k)h_k$ is actually a bounded operator on $L^2(\Gamma)$, so its domain may be extended from $\mathcal{D}(L)$ to $L^2(\Gamma)$. Since $\ker(S) = \{h\}$, we see that $S(I - R_1) = 0$. Therefore

$$Sg = SB\eta = R\eta = \eta, \tag{4.46}$$

where the last step follows from the fact that $\eta_0 := Sg_0 \in \operatorname{ran}(R)$ and

$$\eta_t = SL(\eta + \psi) \quad \Rightarrow \quad \eta - \eta_0 \in \operatorname{ran}(SL) \subset \operatorname{ran}(R). \tag{4.47}$$

We next use $R_1 R = R_1$ and $BS = R\big|_{\mathcal{D}(S)}$ to conclude that

$$g(0) = R_1 BSg_0 + (I - R_1)g_0 = g_0. \tag{4.48}$$

Finally, we check that g solves the evolution equation $g_t = L(Sg + \psi)$:

$$\begin{aligned} g_t &= R_1 BSL(\eta + \psi) + (I - R_1)L\psi \\ &= R_1 L(Sg + \psi) + (I - R_1)L\psi \\ &= L(Sg + \psi) - [(I - R_1)L](Sg + \psi - \psi) \\ &= L(Sg + \psi). \end{aligned} \tag{4.49}$$

In the last step, we used the fact that $(I - R_1)LSg = 0$ since $\operatorname{ran}(LS) \subset \{Gh\}^{\perp} = \ker(I - R_1)$ by (4.13) and (4.16). In the second-to-last step we were careful not to break up $(I - R_1)L$ when acting on ψ since the latter may not belong to the domain of L. In contrast, the function $(Sg + \psi)$ belongs to $\mathcal{D}(L)$ for $t > 0$, as discussed in Remark 4.22. ∎

Remark 4.24. Once $g(t)$ is known, the stress and displacement fields inside the grains are uniquely determined as the solution to the grain boundary displacement jump problem; see Definition 3.24 and Theorem 3.25.

5. Conclusion. The stress-driven grain boundary diffusion problem involves coupling many different equations and phenomena that lead to interesting behavior due to the interplay between nonlocality, singular behavior, and complex geometry. By posing the problem as an evolution of functions defined on the grain boundary, we were able to use methods of semigroup theory to answer fundamental questions of existence, uniqueness, and appropriateness of boundary conditions. In the process, we discovered a class of degenerate grain boundaries that exhibit interesting behavior.

Placing this problem back into the larger model, which includes void and vacancy evolution, grain boundary sliding, etc., it would be interesting to study the behavior of the solution in the vicinity of a junction where a void meets a grain boundary. Here again, questions of appropriate boundary conditions arise, thermodynamic arguments are murky, and singularities in the stress tensor and electric field together with the stiffness inherent in grain boundary diffusion and curvature-driven surface diffusion make the problem difficult to attack theoretically and numerically.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] V. Barbu, *Nonlinear semigroups and differential equations in Banach spaces*, Noordhoff, Leyden, The Netherlands, 1976.
[3] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 1997.
[4] P. G. Ciarlet, *Mathematical Elasticity, Vol.* 1, North–Holland, Amsterdam, 1993.
[5] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer–Verlag, Berlin, 1986.
[6] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr. 18, AMS, Providence, RI, 1969.
[7] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.
[8] S. G. Krein, *Linear Differential Equations in Banach Space*, Transl. Math. Monogr. 29, AMS, Providence, RI, 1971.
[9] J. A. Sethian and J. Wilkening, *A numerical model of stress driven grain boundary diffusion*, J. Comput. Phys., 193 (2003), pp. 275–305.
[10] R. S. Sorbello, *Theory of electromigration*, in Solid State Physics, Vol. 51, H. Ehrenreich and F. Spaepen, eds., Academic Press, New York, 1997, pp. 159–231.
[11] J. A. van Casteren, *Generators of Strongly Continuous Semigroups*, Pitman, Boston, 1985.
[12] J. Wilkening, *Mathematical Analysis and Numerical Simulation of Electromigration*, Ph.D. thesis, University of California, Berkeley, 2002.
[13] J. Wilkening, L. Borucki, and J. A. Sethian, *Analysis of stress-driven grain boundary diffusion. Part* I, SIAM J. Appl. Math. 64 (2004), pp. 1839–1863.
[14] K. Yosida, *Functional Analysis*, Springer-Verlag, Heidelberg, 1980.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 1887–1906

ON THE EVOLUTION OF DOMAIN WALLS IN HARD FERROMAGNETS*

PAOLO PODIO-GUIDUGLI† AND GIUSEPPE TOMASSETTI†

Abstract. We propose a sharp-interface theory for the dynamics of domain walls in highly anisotropic ("hard") ferromagnetic bodies. Starting from the Gilbert equation, we consider the asymptotic regime when the hardness parameter goes to infinity, and we use the technique of matched expansions to derive a system of two evolution equations for the domain wall, regarded as a smooth surface. The first equation, apart for a nonlocal forcing term, has the standard form for a surface set in motion according to its mean curvature. The second relates the normal velocity to the internal structure of the domain wall.

Key words. micromagnetics, domain walls, matched asymptotic expansions, motion by curvature

AMS subject classifications. 35K57, 74N20

DOI. 10.1137/S003613990343402X

1. Introduction.

The formation and the evolution of magnetic domains are the core subject in the mathematical modeling of saturated ferromagnetic bodies. A *magnetic domain* is a region in space where, at a given time, the *magnetization*, a unit vector, has constant direction. Adjacent domains are separated by thin zones where the magnetization direction undergoes large spatial changes; for all practical purposes, these zones can be modeled as sharp interfaces, the *domain walls*. Magnetic domains and their walls are collectively referred to as *domain-wall structures*.

Domain structures are usually observed in static circumstances. To predict their evolution is the goal of *dynamic micromagnetics*, a discipline that was first constituted as a chapter of continuum mechanics in 1963 by Brown [4]. Brown, building on a path-breaking paper by Landau and Lifshitz [16] that appeared in 1935, coined the name "micromagnetics" for the study in a variational format of the static problem of domain formation; it seems that the term "dynamic micromagnetics" was first used in [6], in 1996.

The variational theory of micromagnetics interprets domain structures as the result of minimizing a suitable energy functional [4, 18, 2, 13]. In particular, the occurrence of patchwise-constant energy minimizers accounts for the formation of magnetic domains and domain walls. Dynamic micromagnetics is a much less developed theory, centered about the *(Landau–Lifshitz–) Gilbert equation* [16, 9, 22], a nonlinear parabolic PDE that rules the evolution of the magnetization in a rigid ferromagnet (or, more generally, in a ferromagnet being at mechanical rest, in a sense made precise in [6, 3]). The existence of global-in-time weak solutions to this equation has been established [29, 3], as has their characteristic nonuniqueness [1]; their form has been studied numerically [20].

Here we concentrate on the evolution problem of domain walls, however they were

*Received by the editors September 2, 2003; accepted for publication (in revised form) January 20, 2004; published electronically August 4, 2004. This work was supported by Progetto Cofinanziato 2000, "Modelli Matematici per la Scienza dei Materiali," and by TMR contract FMRX-CT98-0229, "Phase Transitions in Crystalline Solids."

http://www.siam.org/journals/siap/64-6/43402.html

†Dipartimento di Ingegneria Civile, Università di Roma Tor Vergata, Viale del Politecnico, 1-00133 Roma, Italy (ppg@uniroma2.it, tomassetti@ing.uniroma2.it).

formed. We propose to view a domain wall as a smooth surface $\mathcal{S}$ whose dynamics, in the absence of deformations, is a direct consequence of the Gilbert equation alone. Implementing a way to obtain an evolution equation for $\mathcal{S}$ first proposed in [27] (see also [28]), we take the anisotropy modulus as the parameter inducing the asymptotic regime we consider, and we assume that a solution to the Gilbert equation can be constructed by matching two regular expansions, the one holding in a tubular neighborhood of $\mathcal{S}$, the other away from $\mathcal{S}$; roughly speaking, the matching conditions yield the desired evolution equation. This equation has the form, apart for the typically nonlocal forcing term, of the classic equation of motion of a surface by its own curvature.

Hubert and Schäfer [13, p. 215] describe well the role and significance to be ascribed to micromagnetics: "The calculation of domain wall structures is by far the most important contribution of micromagnetics to the analysis of magnetic domains. This is true for two reasons: experimentally, domain walls are difficult to access because they change their properties at surfaces where they can be primarily observed. Also, it is in most cases difficult to isolate a single wall from its neighbours to measure its properties. Usually, domain walls interact in a complicated network." In connection with this last remark, we believe that our finding an evolution equation for a single wall opens the way to a study of wall junctions.

Our paper is organized as follows. In section 2 we detail the form of the Gilbert equation on which we base our study. In section 3 we illustrate the static solution of the Gilbert equation due to Landau and Lifshitz [16], a suitable preliminary to a dynamic solution that we also illustrate, the traveling-wave solution, first discovered by Walker [30], which mimics the motion of a flat wall in an infinite body under the action of a uniform external magnetic field parallel to the wall plane. In section 4 we delineate the asymptotic regime under which our motion equations will emerge as a consequence of the Gilbert equation. In sections 5 and 6 we address the derivation of those equations using the method of matched asymptotic expansions. Both our approach and our methods, as well as our results, differ from those of authors who have looked at the same problem before us (and in greater generality, because they have not ignored mechanical deformation, as we do). We briefly discuss their findings, and compare them with ours, in section 7.

2. The Gilbert equation. For a rigid and saturated ferromagnetic body, here identified for short with a domain Ω of $\mathbb{R}^3$, the evolution of the magnetization $\mathbf{m}$, a unimodular vector field on $\Omega \times (0,T)$, is modeled by the Landau–Lifshitz equation, which we write in the Gilbert format:

$$\gamma^{-1}\dot{\mathbf{m}} = \mathbf{m} \times (\mathbf{h}_{\text{eff}} + \mathbf{d}) \quad \text{in } \Omega \times (0,T), \quad \gamma < 0, \tag{2.1}$$

where γ is the opposite of the *gyromagnetic ratio*, $\mathbf{h}_{\text{eff}}$ is the *effective field*, and $\mathbf{d}$ is the *dissipation field* [9, 21].

For the effective field, we take

$$\mathbf{h}_{\text{eff}} = \alpha\Delta\mathbf{m} + \beta(\mathbf{e}\cdot\mathbf{m})\mathbf{e} + \mathbf{h} + \mathbf{h}_{\text{ext}}. \tag{2.2}$$

Here, (i) $\alpha > 0$ is the *exchange constant*, $\beta > 0$ the *anisotropy constant*, and $\mathbf{e}$ a unit vector parallel to the *easy axis* of magnetization of the material; (ii) $\mathbf{h}$ is the *stray field*, defined to be the only $L^2(\mathbb{R}^3)$ solution of the *Maxwell equations* in the quasi-static approximation, namely,

$$\operatorname{curl}\mathbf{h} = \mathbf{0}, \quad \operatorname{div}\mathbf{h} = -\operatorname{div}(\chi_\Omega\mathbf{m}) \quad \text{in } \mathbb{R}^3 \times (0,T), \tag{2.3}$$

with χ_Ω the characteristic function of Ω; and (iii) $\mathbf{h}_{\text{ext}}$ is the *external field*, a forcing term that can be chosen at will. Our choice for the dissipation field is standard:

$$\mathbf{d} = -\mu\dot{\mathbf{m}}, \quad \mu > 0, \tag{2.4}$$

with μ the *dissipation constant*. (See [21] for a discussion of other thermodynamically admissible dissipation mechanisms in ferromagnets.) As explained in the appendix to [27],

$$\dim[\gamma] = (\text{time})^{-1}, \quad \dim[\mu] = \text{time}, \quad \dim[\alpha] = (\text{length})^2,$$

and β is dimensionless; consequently, both $\mathbf{h}$ and $\mathbf{h}_{\text{ext}}$ are dimensionless as well.

With these constitutive prescriptions, the Gilbert equation takes the form we study here, namely,

$$\gamma^{-1}\dot{\mathbf{m}} + \mu\mathbf{m}\times\dot{\mathbf{m}} = \mathbf{m}\times(\alpha\Delta\mathbf{m} + \beta(\mathbf{m}\cdot\mathbf{e})\mathbf{e} + \mathbf{h} + \mathbf{h}_{\text{ext}}) \quad \text{in } \Omega\times(0,T). \tag{2.5}$$

In the next section we review some known explicit and exact solutions to this equation.

3. Flat walls. Consider an infinite body ($\Omega \equiv \mathbb{R}^3$) composed of a uniaxial material so oriented as to have $\mathbf{e} = \mathbf{c}_3$ (with $\mathbf{c}_3$ the third vector of a given orthonormal basis), and look for solutions

$$\mathbf{m} = \mathbf{m}(x_1, t) \tag{3.1}$$

of the Gilbert equation (2.5) whose spatial dependence is only through the first coordinate x_1 of the typical point $\mathbf{x} \in \Omega$ and such that

$$\lim_{x_1\to\pm\infty} \mathbf{m}(x_1, t) = \pm\mathbf{c}_3$$

at all times t. For such solutions, the Maxwell equations (2.3) yield

$$\mathbf{h}(x_1, t) = -(\mathbf{m}(x_1, t)\cdot\mathbf{c}_1)\mathbf{c}_1 + \mathbf{h}_\infty(t). \tag{3.2}$$

We dispose of the space constant $\mathbf{h}_\infty$ by asking that

$$\lim_{x_1\to\pm\infty} \mathbf{h}(x_1, t) = \mathbf{0} \tag{3.3}$$

at all times. The Gilbert equation then becomes

$$\gamma^{-1}\dot{\mathbf{m}} + \mu\mathbf{m}\times\dot{\mathbf{m}} = \mathbf{m}\times(\alpha\mathbf{m}'' + \mathbf{T}\,\mathbf{m}) \quad \text{in } \mathbb{R}\times(0,T), \tag{3.4}$$

where $(\cdot)'$ denotes differentiation with respect to x_1 and where

$$\mathbf{T} = \beta\mathbf{c}_3\otimes\mathbf{c}_3 - \mathbf{c}_1\otimes\mathbf{c}_1. \tag{3.5}$$

3.1. The static solution of Landau and Lifshitz. A time-independent solution of type (3.1) was found by Landau and Lifshitz [16] for the case when the external field is everywhere null. In this case (3.4) becomes

$$\mathbf{0} = \mathbf{m}\times(\alpha\mathbf{m}'' + \mathbf{T}\,\mathbf{m}) \quad \text{in } \mathbb{R}. \tag{3.6}$$

If we parametrize $\mathbf{m}$ by means of two scalar fields ϑ and φ (Figure 3.1) such that

$$\begin{aligned} \mathbf{m}(\vartheta,\varphi) &= \cos\vartheta\,\mathbf{c}_3 + \sin\vartheta\,\mathbf{t}(\varphi), \quad \vartheta\in[0,\pi], \\ \mathbf{t}(\varphi) &= \sin\varphi\,\mathbf{c}_1 + \cos\varphi\,\mathbf{c}_2, \quad \varphi\in[-\pi,\pi[, \end{aligned} \tag{3.7}$$

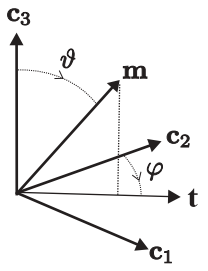


FIG. 3.1. *Parametrization of* $\mathbf{m}$.

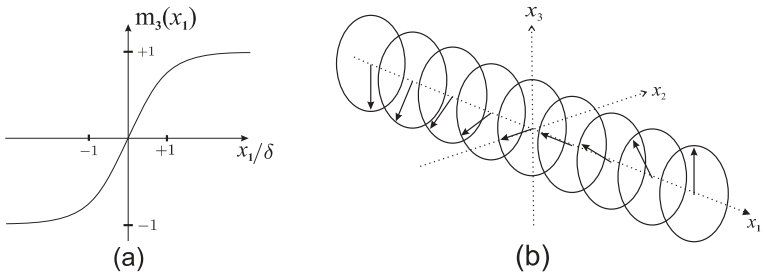


FIG. 3.2. *Wall profile.*

then the Landau–Lifshitz solution has the form

$$\begin{aligned} &\vartheta(x_1) = \arccos \tanh \frac{x_1}{\delta}, \\ &\varphi(x_1) = 0 \end{aligned} \tag{3.8}$$

with

$$\delta = \left(\frac{\alpha}{\beta}\right)^{\frac{1}{2}}. \tag{3.9}$$

Landau and Lifshitz found this solution by minimizing over the set of unimodular vector fields $\mathbf{m}$ on $\mathbb{R}$ the functional

$$\mathbf{m} \mapsto \int_{-\infty}^{+\infty} \left(\alpha|\mathbf{m}'|^2 - \mathbf{m}\cdot\mathbf{T}\mathbf{m}\right) dx_1,$$

whose Euler–Lagrange equation is (3.6). We see from (3.8) that, at a large distance from the origin (that is to say, for $|x_1/\delta| \gg 1$), the magnetization is nearly parallel to the easy axis ($\mathbf{m}(x_1) \times \mathbf{e} \simeq \mathbf{0}$), its direction being determined by the sign of x_1. The body can be thought of as partitioned into two infinite *magnetic domains*, $x_1 < -\delta$ (where $\mathbf{m} \simeq \mathbf{e}$) and $\delta < x_1$ (where $\mathbf{m} \simeq -\mathbf{e}$), separated by the *domain wall*, a transition layer of thickness 2δ within which the magnetization undergoes most of its overall 180° rotation in a plane perpendicular to $\mathbf{c}_1$ (Figure 3.2); the accompanying stray field is everywhere null. The situation depicted in Figure 3.2(b) is commonly referred to as a *Bloch wall*.

3.2. Walker's dynamic solution. The Landau–Lifshitz solution was generalized by Walker [30] to the case when the external field has the form

$$\mathbf{h}_{\text{ext}} = -H\,\mathbf{c}_3 \tag{3.10}$$

with H a real constant such that

$$|H| < \frac{1}{2}\gamma\mu. \tag{3.11}$$

The traveling-wave solution found by Walker is

$$\begin{aligned} \vartheta(x_1,t) &= \arccos\tanh\frac{x_1 - v_H t}{\delta_H}, \\ \varphi(x_1,t) &= \varphi_H. \end{aligned} \tag{3.12}$$

The Walker solution coincides with the Landau–Lifshitz solution for an observer moving with velocity $v_H\mathbf{c}_1$ because, at each time t, the dissipation associated with the evolution of the magnetization $\mathbf{m}$ is exactly compensated by the working of the external field $\mathbf{h}_{\text{ext}}$ [23]. For Walker, the vector $\mathbf{m}$ must lie in a plane through the easy axis forming an angle φ_H with the plane of the domain wall such that

$$\sin(2\varphi_H) = -\frac{2}{\gamma\mu}\,H, \tag{3.13}$$

whence the bounds (3.11) on the applied field H; $\mathbf{m}$ is nearly parallel to the easy axis, everywhere except in a layer whose thickness is of the order

$$\delta_H = \left(\frac{\alpha}{\beta + \sin^2\varphi_H}\right)^{\frac{1}{2}} \tag{3.14}$$

(cf. $(3.8)_2$ and (3.9)).

All in all, the Walker solution pictures a flat domain wall moving with velocity

$$v_H = \frac{1}{\mu}\left(\frac{\alpha}{\beta + \sin^2\varphi_H}\right)^{\frac{1}{2}} H \tag{3.15}$$

in the direction of the x_1-axis.[1]

What we learn from this explicit solution of the Gilbert equation is that two material parameters happen to be of special importance in the dynamics of domain walls: the *wall thickness* δ and the *wall mobility*

$$\nu := \lim_{H\to 0}\frac{v_H}{H} = \frac{\delta}{\mu}\,. \tag{3.16}$$

Indeed, these two parameters play a crucial role in tuning the asymptotics yielding our sharp-interface theory, what we do in the next section.

4. Basic scalings. Given a length scale L and a time scale T, we introduce the dimensionless independent variables

$$\tilde{\mathbf{x}} = L^{-1}\mathbf{x}, \quad \tilde{t} = T^{-1}t. \tag{4.1}$$

Then, the material parameters in the Gilbert equation (2.5) scale as

$$\tilde{\gamma} = T\gamma, \quad \tilde{\mu} = T^{-1}\mu, \quad \tilde{\alpha} = L^{-2}\alpha. \tag{4.2}$$

[1] Combining (3.15) with (3.11) yields a bound on $|v_H|$, referred to as the "breakdown velocity" [13].

As noted in section 2, the anisotropy constant β carries no dimension and hence is not affected by the scaling (4.1). A dimensionless version of the Gilbert equation (2.5) is

$$\beta\,\tilde{\delta}^{-1}\,\tilde{\nu}(\tau\,\dot{\mathbf{m}}+\mathbf{m}\times\dot{\mathbf{m}})=\mathbf{m}\times(\tilde{\delta}^2\Delta\mathbf{m}+(\mathbf{m}\cdot\mathbf{e})\mathbf{e}+\beta^{-1}(\mathbf{h}+\mathbf{h}_{\rm ext})), \tag{4.3}$$

where the dimensionless wall thickness and mobility are defined by

$$\tilde{\delta}=\left(\frac{\tilde{\alpha}}{\beta}\right)^{\frac{1}{2}},\quad \tilde{\nu}=\frac{\tilde{\delta}}{\tilde{\mu}}, \tag{4.4}$$

and where

$$\tau=(\gamma\mu)^{-1}. \tag{4.5}$$

The physical assumption that underlies our sharp-interface theory under construction is that *β is as large as necessary to make the wall thickness as small as desired*. One way to achieve this is to identify our smallness parameter with β^{-1}:

$$\beta=\varepsilon^{-1}. \tag{4.6}$$

Then, we stipulate that *the dimensionless wall thickness scales like ε and the dimensionless wall mobility scales like* 1:

$$\tilde{\delta}=\varepsilon,\quad \tilde{\nu}=1. \tag{4.7}$$

To achieve this, we choose L and T as follows:

$$L=\varepsilon^{-1/2}\,\alpha^{1/2},\quad T=\varepsilon^{-1}\mu. \tag{4.8}$$

With these choices, (4.3) becomes

$$\varepsilon^2\,(\tau\,\dot{\mathbf{m}}+\mathbf{m}\times\dot{\mathbf{m}})=\mathbf{m}\times\left(\varepsilon^2\Delta\mathbf{m}+(\mathbf{e}\cdot\mathbf{m})\mathbf{e}+\varepsilon(\mathbf{h}+\mathbf{h}_{\rm ext})\right). \tag{4.9}$$

When ε is small, exchange interactions become negligible with respect to anisotropy interactions, and we expect $\mathbf{m}$ to quickly converge toward a local minimum of the anisotropy energy. Hence, in a short time, the region Ω presents itself as partitioned into magnetic domains, in each of which $\mathbf{m}$ is nearly parallel to $\mathbf{e}$. Thin transition layers, the domain walls, separate neighboring domains with opposite magnetization. As anticipated in the introduction, we disregard the process of domain formation and focus on the evolution of domain boundaries, in the limit for $\varepsilon\downarrow 0$ when they are expected to become surfaces. To simplify matters, we take $\Omega=\mathbb{R}^3$ so that the Maxwell equations take the form

$$\operatorname{curl}\mathbf{h}=\mathbf{0},\quad \operatorname{div}(\mathbf{h}+\mathbf{m})=0\quad \text{in } \mathbb{R}^3\times(0,T) \tag{4.10}$$

and need not be scaled.

5. Time and space differentiation following a moving surface. Let $\mathcal{S}_t=\mathcal{S}(t)$ be a smooth oriented surface, smoothly evolving over the time interval $\mathcal{T}\subset(0,T)$. For each $t\in\mathcal{T}$, let $\mathcal{W}_t$ be a *tubular neighborhood* of $\mathcal{S}_t$ of thickness $2h_t$, that is to say, an open set of $\mathbb{R}^3$ in one-to-one correspondence with the set $\mathcal{S}_t\times(-h_t,+h_t)$ by way of the mapping

$$\mathcal{S}_t\times(-h_t,+h_t)\ni(\mathbf{s},d)\ \leftrightarrow\ \mathbf{s}+d\,\mathbf{n}_t(\mathbf{s})=\mathbf{x}\in\mathcal{W}_t, \tag{5.1}$$

where $\mathbf{n}_t(\mathbf{s})$ denotes the positively oriented unit normal to $\mathcal{S}_t$ at its point $\mathbf{s}$. We take $\mathcal{T}$ short enough for the set

$$\mathcal{W} := \bigcap_{t\in\mathcal{T}} \mathcal{W}_t$$

to be open and nonempty; needless to say, $\mathcal{W}$ contains a tubular neighborhood of $\mathcal{S}_t$ for all $t \in \mathcal{T}$. For each $\mathbf{x} \in \mathcal{W}$ and for each $t \in \mathcal{T}$, we denote by $\hat{\mathbf{s}}_t$ the *projection* of $\mathbf{x}$ on $\mathcal{S}_t$ and by $\hat{d}_t$ the *signed distance* of $\mathbf{x}$ from $\mathcal{S}_t$. Precisely, we set

$$\mathbf{s} = \hat{\mathbf{s}}_t(\mathbf{x}), \quad d = \hat{d}_t(\mathbf{x}); \tag{5.2}$$

we also set, consistently,

$$\mathbf{n} = \hat{\mathbf{n}}_t(\mathbf{x}) := \mathbf{n}_t(\hat{\mathbf{s}}_t(\mathbf{x})). \tag{5.3}$$

5.1. Some preparatory results. For each $\varepsilon > 0$ fixed, we introduce the *scaled* signed distance

$$r_\varepsilon = \hat{r}_\varepsilon(\mathbf{x}, t) := \varepsilon^{-1}\hat{d}_t(\mathbf{x}) \tag{5.4}$$

and consider the following identity over $\mathcal{W} \times \mathcal{T}$:

$$\hat{\mathbf{s}}_t(\mathbf{x}) + \varepsilon\, \hat{r}_\varepsilon(\mathbf{x}, t)\, \hat{\mathbf{n}}_t(\mathbf{x}) = \mathbf{x}. \tag{5.5}$$

Differentiating (5.5) with respect to t and taking the scalar product of both sides with $\mathbf{n}$ we get

$$r_\varepsilon^{\bullet} = -\varepsilon^{-1} v, \tag{5.6}$$

where

$$v = \hat{v}_t(\mathbf{x}) := \hat{\mathbf{s}}_t^{\bullet}(\mathbf{x}) \cdot \hat{\mathbf{n}}_t(\mathbf{x}) \tag{5.7}$$

is the *normal velocity* of the surface $\mathcal{S}_t$ at point $\hat{\mathbf{s}}_t(\mathbf{x})$.

Differentiating (5.5) with respect to $\mathbf{x}$ and using the fact that

$$\operatorname{grad} d = \mathbf{n}, \tag{5.8}$$

we obtain that

$$\operatorname{grad}\mathbf{s} = \mathbf{P} - \varepsilon\, r_\varepsilon \operatorname{grad}\mathbf{n}, \quad \mathbf{P} = \hat{\mathbf{P}}_t(\mathbf{x}) := \mathbf{1} - \hat{\mathbf{n}}_t(\mathbf{x}) \otimes \hat{\mathbf{n}}_t(\mathbf{x}). \tag{5.9}$$

Next, we differentiate with respect to $\mathbf{x}$ relation (5.3) and use the chain rule to get

$$\operatorname{grad}\mathbf{n} = (\partial_{\mathbf{s}}\mathbf{n}_t)\operatorname{grad}\mathbf{s} = -\mathbf{L}\operatorname{grad}\mathbf{s}, \tag{5.10}$$

where the *Weingarten tensor* $\mathbf{L}$ of the surface $\mathcal{S}_t$ can be thought of as a tensor field over $\mathcal{W} \times \mathcal{T}$ according to the following definition:

$$\mathbf{L} = \hat{\mathbf{L}}_t(\mathbf{x}) := -\partial_{\mathbf{s}}\mathbf{n}_t\,|_{\mathbf{s}=\hat{\mathbf{s}}_t(\mathbf{x})}. \tag{5.11}$$

As is well known,

$$\mathbf{LP} = \mathbf{L}. \tag{5.12}$$

Combining (5.9), (5.10), and (5.12), we arrive at

$$\operatorname{grad}\mathbf{n} = -\mathbf{L} + \varepsilon\, r_\varepsilon\, \mathbf{L}\operatorname{grad}\mathbf{n}. \tag{5.13}$$

We take the trace of both sides of (5.13) and obtain

$$\operatorname{div}\mathbf{n} = -k + \varepsilon\, r_\varepsilon\, \mathbf{L}\cdot\operatorname{grad}\mathbf{n}, \tag{5.14}$$

where

$$k = \hat{k}_t(\mathbf{x}) := \operatorname{tr}\hat{\mathbf{L}}_t(\mathbf{x}) \tag{5.15}$$

is twice the mean curvature of $\mathcal{S}_t$ at point $\mathbf{s} = \hat{\mathbf{s}}_t(\mathbf{x})$.

5.2. Time rate, gradient, Laplacian, and curl of a vector field. Let now

$$\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$$

be a smooth vector field over $\mathbb{R}^3 \times \mathcal{T}$. For each $\varepsilon > 0$ fixed, the *inner representation* of $\mathbf{v}$ is delivered by the map $\check{\mathbf{v}}_\varepsilon$ over $\mathbb{R}\times\mathcal{W}\times\mathcal{T}$ defined by

$$\check{\mathbf{v}}_\varepsilon(r, \mathbf{x}, t) := \mathbf{v}(\hat{\mathbf{s}}_t(\mathbf{x}) + \varepsilon\, r\, \hat{\mathbf{n}}_t(\mathbf{x}), t). \tag{5.16}$$

A consequence of this definition is that, for each $(r, \mathbf{x}, t) \in \mathbb{R}\times\mathcal{W}\times\mathcal{T}$ fixed, the relation

$$\check{\mathbf{v}}_\varepsilon(r, \mathbf{x} + \alpha\, \hat{\mathbf{n}}_t(\mathbf{x}), t) = \check{\mathbf{v}}_\varepsilon(r, \mathbf{x}, t) \tag{5.17}$$

holds identically for α in the open neighborhood of 0 where

$$\hat{\mathbf{s}}_t(\mathbf{x}) = \hat{\mathbf{s}}_t(\mathbf{x} + \alpha\, \hat{\mathbf{n}}_t(\mathbf{x})).$$

We now differentiate with respect to t the identity

$$\mathbf{v}(\mathbf{x}, t) = \check{\mathbf{v}}_\varepsilon(\hat{r}_\varepsilon(\mathbf{x}, t), \mathbf{x}, t), \tag{5.18}$$

so as to obtain

$$\mathbf{v}^\bullet = r_\varepsilon^\bullet\, \partial_r \check{\mathbf{v}}_\varepsilon + \partial_t \check{\mathbf{v}}_\varepsilon, \tag{5.19}$$

that is, recalling (5.6),

$$\mathbf{v}^\bullet = -\varepsilon^{-1}\, v\, \partial_r \check{\mathbf{v}}_\varepsilon + \partial_t \check{\mathbf{v}}_\varepsilon. \tag{5.20}$$

Next, we differentiate (5.18) with respect to $\mathbf{x}$ and obtain

$$\operatorname{grad}\mathbf{v} = \partial_r \check{\mathbf{v}}_\varepsilon \otimes \operatorname{grad} r_\varepsilon + \partial_{\mathbf{x}} \check{\mathbf{v}}_\varepsilon, \tag{5.21}$$

which we rewrite as

$$\operatorname{grad}\mathbf{v} = \varepsilon^{-1} \partial_r \check{\mathbf{v}}_\varepsilon \otimes \mathbf{n} + \partial_{\mathbf{x}} \check{\mathbf{v}}_\varepsilon, \tag{5.22}$$

using (5.4) and (5.8).

We are now in a position to compute both the Laplacian and the curl of $\mathbf{v}$ in terms of its inner representation.

As to the Laplacian, taking the divergence of both sides of (5.22), we find that

$$\Delta \mathbf{v} = \varepsilon^{-1}(\operatorname{grad}(\partial_r \check{\mathbf{v}}_\varepsilon))\,\mathbf{n} + \varepsilon^{-1}(\operatorname{div}\mathbf{n})\,\partial_r \check{\mathbf{v}}_\varepsilon + \operatorname{div}\partial_{\mathbf{x}} \check{\mathbf{v}}_\varepsilon, \tag{5.23}$$

where

$$\operatorname{grad}(\partial_r \check{\mathbf{v}}_\varepsilon) = -\varepsilon^{-1}\partial_{rr}\check{\mathbf{v}}_\varepsilon \otimes \mathbf{n} + \partial_{\mathbf{x}}\partial_r \check{\mathbf{v}}_\varepsilon. \tag{5.24}$$

But, since differentiation of (5.17) with respect to r and α yields

$$(\partial_{\mathbf{x}}(\partial_r \check{\mathbf{v}}_\varepsilon))\mathbf{n} = \mathbf{0}, \tag{5.25}$$

substitution of (5.14) and (5.24) into (5.23) yields

$$\Delta \mathbf{v} = \varepsilon^{-2}\partial_{rr}\check{\mathbf{v}}_\varepsilon - \varepsilon^{-1}\partial_r \check{\mathbf{v}}_\varepsilon + \operatorname{div}\partial_{\mathbf{x}}\check{\mathbf{v}}_\varepsilon + r_\varepsilon(\mathbf{L}\cdot\operatorname{grad}\mathbf{n})\,\partial_r\check{\mathbf{v}}_\varepsilon. \tag{5.26}$$

As to the curl, again from (5.22) we deduce that

$$\operatorname{curl}\mathbf{v} = \varepsilon^{-1}\mathbf{n}\times\partial_r\check{\mathbf{v}}_\varepsilon + (\partial_{\mathbf{x}}\check{\mathbf{v}}_\varepsilon - (\partial_{\mathbf{x}}\check{\mathbf{v}}_\varepsilon)^T)_\times, \tag{5.27}$$

where $\mathbf{W}_\times$ denotes the axial vector associated to the skew tensor $\mathbf{W}$.

5.3. Estimates. Equations (5.22), (5.20), (5.26), and (5.27) yield the following estimates:

$$\mathbf{v}^{\bullet} = -\varepsilon^{-1}\,v\,\partial_r\check{\mathbf{v}}_\varepsilon + O(1), \tag{5.28}$$

$$\operatorname{grad}\mathbf{v} = \varepsilon^{-1}\partial_r\check{\mathbf{v}}_\varepsilon\otimes\mathbf{n} + O(1), \tag{5.29}$$

$$\Delta\mathbf{v} = \varepsilon^{-2}\partial_{rr}\check{\mathbf{v}}_\varepsilon - \varepsilon^{-1}k\partial_r\check{\mathbf{v}}_\varepsilon + O(1), \tag{5.30}$$

$$\operatorname{curl}\mathbf{v} = \varepsilon^{-1}\mathbf{n}\times\partial_r\check{\mathbf{v}}_\varepsilon + O(1), \tag{5.31}$$

$$\operatorname{div}\mathbf{v} = \varepsilon^{-1}\partial_r\check{\mathbf{v}}_\varepsilon\cdot\mathbf{n} + O(1). \tag{5.32}$$

6. Matched asymptotics. For convenience, let us repeat here the system (4.9)–(4.10):

$$\begin{gathered}\varepsilon^2\,(\tau\,\dot{\mathbf{m}} + \mathbf{m}\times\dot{\mathbf{m}}) = \mathbf{m}\times\big(\varepsilon^2\Delta\mathbf{m} + (\mathbf{e}\cdot\mathbf{m})\mathbf{e} + \varepsilon(\mathbf{h}+\mathbf{h}_{\text{ext}})\big),\\ \operatorname{curl}\mathbf{h} = \mathbf{0},\quad \operatorname{div}(\mathbf{h}+\mathbf{m}) = 0.\end{gathered} \tag{6.1}$$

We recall that the independent variables in (6.1) are the dimensionless variables $\tilde{\mathbf{x}}$ and $\tilde{t}$ defined in (4.1); to lighten our notation, we omit the superposed tildes until, in subsection 6.4, we return to the original space and time variables.

We assume that there is a positive constant $\bar{\varepsilon}$ such that for each $\varepsilon\in(0,\bar{\varepsilon})$, the system (6.1) has a solution $(\mathbf{m}_\varepsilon,\ \mathbf{h}_\varepsilon)$ defined on $\mathbb{R}^3\times\mathcal{T}$. We also assume that the surface $\mathcal{S}_t$ introduced in the previous section splits $\mathbb{R}^3$ into a pair of disjoint regions $\mathcal{D}_t^+$ and $\mathcal{D}_t^-$ having $\mathcal{S}_t$ as their common boundary, and that, at each time $t\in\mathcal{T}$, the limits

$$\begin{aligned}\mathbf{m}_0(\mathbf{x},t) &= \lim_{\varepsilon\to 0}\mathbf{m}_\varepsilon(\mathbf{x},t),\\ \mathbf{h}_0(\mathbf{x},t) &= \lim_{\varepsilon\to 0}\mathbf{h}_\varepsilon(\mathbf{x},t)\end{aligned} \tag{6.2}$$

exist at each point $\mathbf{x}\in\mathbb{R}^3\setminus\mathcal{S}_t$, with

$$\mathbf{m}_0(\mathbf{x},t) = \begin{cases} -\mathbf{e} & \text{if } \mathbf{x}\in\mathcal{D}_t^-,\\[2ex] +\mathbf{e} & \text{if } \mathbf{x}\in\mathcal{D}_t^+,\end{cases} \tag{6.3}$$

and with $\mathbf{h}_0(\cdot,t)$ a smooth field on $\mathbb{R}^3\setminus\mathcal{S}_t$. Furthermore, we let

$$\begin{aligned}\check{\mathbf{m}}_\varepsilon(r,\mathbf{x},t)&=\mathbf{m}_\varepsilon(\hat{\mathbf{s}}(\mathbf{x},t)+\varepsilon\, r\,\hat{\mathbf{n}}(\mathbf{x},t),t),\\ \check{\mathbf{h}}_\varepsilon(r,\mathbf{x},t)&=\mathbf{h}_\varepsilon(\hat{\mathbf{s}}(\mathbf{x},t)+\varepsilon\, r\,\hat{\mathbf{n}}(\mathbf{x},t),t)\end{aligned}\tag{6.4}$$

be the inner representations of $\mathbf{m}_\varepsilon$ and $\mathbf{h}_\varepsilon$ for each $\varepsilon\in(0,\bar{\varepsilon})$ (see definition (5.16)), and we assume that the limits

$$\begin{aligned}\check{\mathbf{m}}_0(r,\mathbf{x},t)&=\lim_{\varepsilon\to 0}\check{\mathbf{m}}_\varepsilon(r,\mathbf{x},t),\\ \check{\mathbf{h}}_0(r,\mathbf{x},t)&=\lim_{\varepsilon\to 0}\check{\mathbf{h}}_\varepsilon(r,\mathbf{x},t)\end{aligned}\tag{6.5}$$

exist for each $(r,\mathbf{x},t)\in\mathbb{R}\times(\mathbb{R}^3\setminus\mathcal{S}_t)\times\mathcal{T}$. The following *matching conditions* will soon prove crucial to our developments:

$$\begin{aligned}\lim_{r\to\pm\infty}\check{\mathbf{m}}_0(r,\mathbf{x},t)&=\lim_{d\to 0\pm}\mathbf{m}_0(\hat{\mathbf{s}}_t(\mathbf{x})+d\,\hat{\mathbf{n}}_t(\mathbf{x}),t),\\ \lim_{r\to\pm\infty}\check{\mathbf{h}}_0(r,\mathbf{x},t)&=\lim_{d\to 0\pm}\mathbf{h}_0(\hat{\mathbf{s}}_t(\mathbf{x})+d\,\hat{\mathbf{n}}_t(\mathbf{x}),t)\end{aligned}\tag{6.6}$$

for each $(\mathbf{x},t)\in\mathcal{W}\times\mathcal{T}$. The reader may consult [7] or [5, 8] for justifications of (6.6). As to a motivation, we note that, for each fixed pair $(\mathbf{x},t)\in\mathcal{W}\times\mathcal{T}$, the inner representation of the field $\mathbf{m}_\varepsilon$ can be written in the alternative forms

$$\check{\mathbf{m}}_\varepsilon(r,\mathbf{x},t)=\mathbf{m}_\varepsilon(\hat{\mathbf{s}}_t(\mathbf{x})+\varepsilon r\,\hat{\mathbf{n}}_t(\mathbf{x}),t)=\mathbf{m}_\varepsilon(\hat{\mathbf{s}}_t(\mathbf{x})+d\,\hat{\mathbf{n}}_t(\mathbf{x}),t)=\check{\mathbf{m}}_\varepsilon\left(\frac{d}{\varepsilon},\mathbf{x},t\right),$$

where the real variables r and d satisfy

$$r^{-1}\,d=\varepsilon;$$

hence, both r^{-1} and d tend to null with ε at a smaller rate than ε itself — both as $\varepsilon^{1/2}$, say.

Finally, we recall that all solutions of the Gilbert equation must have constant modulus. For an expansion such as

$$\mathbf{m}_\varepsilon=\mathbf{m}_0+\varepsilon\,\mathbf{m}_1+o(\varepsilon)$$

to be consistent with the requirement that

$$|\mathbf{m}_\varepsilon|=1,$$

the approximations of $\mathbf{m}_\varepsilon$ must be such that

$$|\mathbf{m}_0|=1,\quad \mathbf{m}_0\cdot\mathbf{m}_1=0,\quad \text{etc.}\tag{6.7}$$

6.1. The zeroth-order magnetization field. For each $(\mathbf{x},t)\in\mathcal{W}\times\mathcal{T}$, the relations (6.5) and the estimates (5.28) and (5.30) yield

$$\begin{aligned}\mathbf{m}_\varepsilon(\mathbf{s}_t(\mathbf{x})+\varepsilon\, r\,\mathbf{n}_t(\mathbf{x}))&=\check{\mathbf{m}}_0(r,\mathbf{x},t)+O(\varepsilon),\\ \mathbf{h}_\varepsilon(\mathbf{s}_t(\mathbf{x})+\varepsilon\, r\,\mathbf{n}_t(\mathbf{x}))&=\check{\mathbf{h}}_0(r,\mathbf{x},t)+O(\varepsilon),\\ \dot{\mathbf{m}}_\varepsilon(\mathbf{s}_t(\mathbf{x})+\varepsilon\, r\,\mathbf{n}_t(\mathbf{x}))&=-\varepsilon^{-1}v\,\partial_r\check{\mathbf{m}}_0(r,\mathbf{x},t)+O(1),\\ \Delta\mathbf{m}_\varepsilon(\mathbf{s}_t(\mathbf{x})+\varepsilon\, r\,\mathbf{n}_t(\mathbf{x}))&=\varepsilon^{-2}\partial_{rr}\check{\mathbf{m}}_0(r,\mathbf{x},t)+O(\varepsilon^{-1}).\end{aligned}\tag{6.8}$$

Substituting (6.8) into the scaled Gilbert equation $(6.1)_1$, and letting $\varepsilon\downarrow 0$, we obtain the following ODE for $\check{\mathbf{m}}_0(\cdot,\mathbf{x},t)$:

$$\check{\mathbf{m}}_0\times(\partial_{rr}\mathbf{m}_0+(\mathbf{e}\otimes\mathbf{e})\check{\mathbf{m}}_0)=\mathbf{0}\quad\text{in }\mathbb{R};\tag{6.9}$$

the field $\check{\mathbf{m}}_0$ must be unimodular to satisfy $(6.7)_1$, and it must comply with the conditions at infinity resulting from the assumption (6.3) and the matching condition $(6.6)_1$, namely,

$$\lim_{r\to\pm\infty} \check{\mathbf{m}}_0(r) = \pm\mathbf{e}. \tag{6.10}$$

To construct a solution to the nonlinear boundary-value problem (6.9)–(6.10), we look for smooth solutions $\mathbf{v} : \mathbb{R} \to S^2$ to the problem

$$\begin{cases} \mathbf{v}\times\mathbf{D}\mathbf{v} = \mathbf{0} \quad \text{in } \mathbb{R}, \quad \mathbf{D}\mathbf{v} := \mathbf{v}'' + (\mathbf{e}\cdot\mathbf{v})\mathbf{e}, \\ \lim\limits_{r\to\pm\infty} \mathbf{v}(r) = \pm\mathbf{e}. \end{cases} \tag{6.11}$$

(Here a superscript prime denotes differentiation.)

We preliminarily note that if $\mathbf{v}(r)$ is a solution of (6.11), then the field

$$\mathbf{v}_{\rho,\varphi}(r) := \mathbf{R}_{\mathbf{e}}(\varphi)\mathbf{v}(r-\rho) \tag{6.12}$$

is a solution as well, whatever the translation $\rho \in \mathbb{R}$ and whatever the angle $\varphi \in (-\pi, +\pi)$ of the rotation $\mathbf{R}_{\mathbf{e}}(\varphi)$ about the easy axis.[2] The identities

$$\mathbf{v}'\times\mathbf{D}\mathbf{v} + \mathbf{v}\times\mathbf{D}\mathbf{v}' = \mathbf{0} \tag{6.13}$$

and

$$(\mathbf{e}\times\mathbf{v})\times\mathbf{D}\mathbf{v} + \mathbf{v}\times\mathbf{D}(\mathbf{e}\times\mathbf{v}) = \mathbf{0} \tag{6.14}$$

obtain by differentiating with respect to ρ and ϕ, respectively, the identity

$$\mathbf{v}_{\rho,\phi}\times\mathbf{D}\mathbf{v}_{\rho,\phi} = \mathbf{0}, \tag{6.15}$$

and evaluating the resulting expression at $\rho = \phi = 0$.[3]

We show in the appendix that smooth solutions to problem (6.11) have the form

$$\mathbf{v}(r) = \tanh(r-\rho)\mathbf{e} + \frac{1}{\cosh(r-\rho)}\,\mathbf{t}, \quad \rho\in\mathbb{R}, \ \ \mathbf{t}\cdot\mathbf{e} = 0 \tag{6.16}$$

(compare with the Landau–Lifshitz and Walker solutions). Accordingly, we assign to the zeroth-order magnetization field inside the domain wall the form

$$\check{\mathbf{m}}_0(r,\mathbf{x},t) = \tanh(r-\rho(\mathbf{s}_t(\mathbf{x}),t))\,\mathbf{e} + \frac{1}{\cosh(r-\rho(\mathbf{s}_t(\mathbf{x}),t))}\,\mathbf{t}(\mathbf{s}_t(\mathbf{x}),t) \tag{6.17}$$

with $\rho(\cdot,t)$ and $\mathbf{t}(\cdot,t)$ two smooth fields on $\mathcal{S}_t$, the latter being unimodular and orthogonal to the easy axis. Not only does this vector field solve problem (6.9)–(6.10) but it also satisfies two identities corresponding, respectively, to (6.13) and (6.14):

$$\begin{aligned} &\partial_r\check{\mathbf{m}}_0\times\mathbf{D}\check{\mathbf{m}}_0 + \check{\mathbf{m}}_0\times\mathbf{D}\,\partial_r\check{\mathbf{m}}_0 = \mathbf{0}, \\ &(\mathbf{e}\times\check{\mathbf{m}}_0)\times\mathbf{D}\check{\mathbf{m}}_0 + \check{\mathbf{m}}_0\times\mathbf{D}(\mathbf{e}\times\check{\mathbf{m}}_0) = \mathbf{0}, \end{aligned} \tag{6.18}$$

[2] Recall that

$$\mathbf{R}_{\mathbf{e}}(\varphi) = \mathbf{I} + \sin\varphi\,\mathbf{E} + (1-\cos\varphi)(\mathbf{e}\otimes\mathbf{e} - \mathbf{I}), \quad \mathbf{E}_\times = \mathbf{e}.$$

Needless to say, whatever ρ and φ,

$$|\mathbf{v}_{\rho,\varphi}(r)| = |\mathbf{v}(r-\rho)| \quad \text{for all } r\in\mathbb{R}.$$

[3] In fact, relation (6.13) also follows from differentiating the first of (6.11) with respect to r; a completely analogous relation holds for $\mathbf{v}''$, $\mathbf{v}'''$, etc.

where now

$$\mathbf{D}\mathbf{v} = \partial_{rr}\mathbf{v} + (\mathbf{e}\cdot\mathbf{v})\mathbf{v} \quad \text{for } \mathbf{v} = \check{\mathbf{v}}(r,\mathbf{x},t). \tag{6.19}$$

In addition, as is easy to verify on the basis of (6.17),

$$\lim_{r\to\pm\infty} \partial_r \check{\mathbf{m}}_0(r,\mathbf{x},t) = \mathbf{0}, \quad \lim_{r\to\pm\infty} \partial_{rr} \check{\mathbf{m}}_0(r,\mathbf{x},t) = \mathbf{0}. \tag{6.20}$$

6.2. The zeroth-order stray field. We now use the Maxwell equations to compute $\check{\mathbf{h}}_0$, the lowest-order term of the stray field inside the domain wall.

In view of $(6.5)_2$, (5.31), and (5.32), we have that

$$\begin{aligned} &\operatorname{curl} \mathbf{h}_\varepsilon = \varepsilon^{-1}\mathbf{n}\times\partial_r\check{\mathbf{h}}_0 + O(1),\\ &\operatorname{div} \mathbf{h}_\varepsilon = \varepsilon^{-1}\partial_r\check{\mathbf{h}}_0\cdot\mathbf{n} + O(1). \end{aligned} \tag{6.21}$$

Substituting (6.21) into (4.10) and letting $\varepsilon \downarrow 0$, we get

$$\partial_r\check{\mathbf{h}}_0\times\hat{\mathbf{n}} = \mathbf{0}, \quad \partial_r(\check{\mathbf{h}}_0 + \check{\mathbf{m}}_0)\cdot\hat{\mathbf{n}} = 0. \tag{6.22}$$

The associated boundary conditions are

$$\lim_{r\to\pm\infty} \check{\mathbf{h}}_0(r,\mathbf{s},t) = \lim_{d\to 0\pm} \mathbf{h}_0(\mathbf{s} + d\,\hat{\mathbf{n}}(\mathbf{s},t),t) \tag{6.23}$$

(cf. $(6.6)_2$).

The solution of (6.22)–(6.23) is

$$\check{\mathbf{h}}_0(r,\mathbf{s},t) = \langle\!\langle \mathbf{h}_0 + (\mathbf{m}_0\cdot\mathbf{n})\mathbf{n}\rangle\!\rangle(\mathbf{s},t) - (\check{\mathbf{m}}_0(r,\mathbf{s},t)\cdot\hat{\mathbf{n}}(\mathbf{s},t))\hat{\mathbf{n}}(\mathbf{s},t); \tag{6.24}$$

moreover, the jump conditions

$$[\![\mathbf{h}_0\times\hat{\mathbf{n}}]\!](\mathbf{s},t) = \mathbf{0}, \quad [\![(\mathbf{h}_0 + \mathbf{m}_0)\cdot\hat{\mathbf{n}}]\!](\mathbf{s},t) = \mathbf{0} \tag{6.25}$$

hold at all points $\mathbf{s}\in\mathcal{S}_t$. Here, for φ a field on $\mathbb{R}^3\times(0,T)$, we denote by

$$\langle\!\langle\varphi\rangle\!\rangle(\mathbf{s},t) = \lim_{d\to 0}\frac{1}{2}\left(\varphi(\mathbf{s}+d\mathbf{n},t) + \varphi(\mathbf{s}-d\mathbf{n},t)\right) \tag{6.26}$$

and

$$[\![\varphi]\!](\mathbf{s},t) = \lim_{d\to 0+}\left(\varphi(\mathbf{s}+d\mathbf{n},t) - \varphi(\mathbf{s}-d\mathbf{n},t)\right), \tag{6.27}$$

respectively, the *mean value* and the *jump* of φ at point $\mathbf{s}\in\mathcal{S}_t$. We note here for later use the identity

$$[\![\varphi\psi]\!] = \langle\!\langle\varphi\rangle\!\rangle[\![\psi]\!] + [\![\varphi]\!]\langle\!\langle\psi\rangle\!\rangle, \tag{6.28}$$

whence

$$[\![\varphi\varphi]\!] = 2\,\langle\!\langle\varphi\rangle\!\rangle[\![\varphi]\!]. \tag{6.29}$$

6.3. The first-order magnetization field. We now assume that the inner representation of $\mathbf{m}_\varepsilon$ admits a regular expansion up to the first order:

$$\check{\mathbf{m}}_\varepsilon(r,\mathbf{x},t) = \check{\mathbf{m}}_0(r,\mathbf{x},t) + \varepsilon\check{\mathbf{m}}_1(r,\mathbf{x},t) + o(\varepsilon). \tag{6.30}$$

As a consequence of this assumption and (6.10), we have, among other things, that

$$\lim_{r\to\pm\infty} \check{\mathbf{m}}_1(r,\mathbf{x},t) = \mathbf{0}. \tag{6.31}$$

Furthermore, we can replace relation $(6.8)_4$ by the sharper estimate

$$\Delta\mathbf{m}_\varepsilon = \varepsilon^{-2}\partial_{rr}\check{\mathbf{m}}_0 - \varepsilon^{-1}(k\,\partial_r\mathbf{m}_0 - \partial_{rr}\mathbf{m}_1) + O(1). \tag{6.32}$$

Substituting (6.32) into the scaled Gilbert equation $(6.1)_1$, and using (6.9), we obtain

$$\varepsilon\left(\check{\mathbf{m}}_1\times\mathbf{D}\check{\mathbf{m}}_0 + \check{\mathbf{m}}_0\times\mathbf{D}\check{\mathbf{m}}_1 + \check{\mathbf{b}}_0\right) + O(\varepsilon^2) = \mathbf{0}, \tag{6.33}$$

where $\mathbf{D}$ is defined as in (6.19) and

$$\check{\mathbf{b}}_0 := \sigma v\,\partial_r\check{\mathbf{m}}_0 + \check{\mathbf{m}}_0\times(\mathbf{h}_{\rm ext} + \check{\mathbf{h}}_0 + (v-k)\partial_r\check{\mathbf{m}}_0); \tag{6.34}$$

note that $\check{\mathbf{b}}_0\cdot\check{\mathbf{m}}_0 = 0$. Hence, for each fixed $(\mathbf{x},t)\in\mathcal{W}\times\mathcal{T}$, the field $\check{\mathbf{m}}_1(\cdot,\mathbf{x},t)$ must satisfy the ODE

$$\check{\mathbf{m}}_0\times\mathbf{D}\check{\mathbf{m}}_1 + \check{\mathbf{m}}_1\times\mathbf{D}\check{\mathbf{m}}_0 + \check{\mathbf{b}}_0 = \mathbf{0} \tag{6.35}$$

as well as the boundary conditions at infinity expressed by (6.31). We now derive two *solvability conditions* for the linear problem ruled by (6.35) and (6.31).

First, taking the dot product of both sides of (6.35) with $\check{\mathbf{m}}_0\times\partial_r\check{\mathbf{m}}_0$, we find[4]

$$\partial_r\check{\mathbf{m}}_0\cdot\mathbf{D}\check{\mathbf{m}}_1 - (\partial_r\check{\mathbf{m}}_0\cdot\check{\mathbf{m}}_1)(\check{\mathbf{m}}_0\cdot\mathbf{D}\check{\mathbf{m}}_0) + \check{\mathbf{m}}_0\times\partial_r\check{\mathbf{m}}_0\cdot\check{\mathbf{b}}_0 = 0,$$

or rather

$$(\mathbf{A}_0\check{\mathbf{m}}_1 + \check{\mathbf{b}}_0\times\check{\mathbf{m}}_0)\cdot\partial_r\check{\mathbf{m}}_0 = 0, \tag{6.36}$$

where

$$\mathbf{A}_0 := \mathbf{D} - (\check{\mathbf{m}}_0\cdot\mathbf{D}\check{\mathbf{m}}_0)\mathbf{I}. \tag{6.37}$$

With the use of (6.20) and (6.31), an integration by parts over the real line yields

$$\int_{-\infty}^{+\infty}(\mathbf{A}_0\check{\mathbf{m}}_1)\cdot\partial_r\check{\mathbf{m}}_0\,dr = \int_{-\infty}^{+\infty}(\mathbf{A}_0\check{\mathbf{m}}_0)\cdot\partial_r\check{\mathbf{m}}_1\,dr. \tag{6.38}$$

[4]We exploit the fact that $\check{\mathbf{m}}_0$ is orthogonal both to $\partial_r\check{\mathbf{m}}_0$ (because $|\check{\mathbf{m}}_0| = 1$) and to $\check{\mathbf{m}}_1$ (because of $(6.7)_2$) and apply two consequences of the vectorial identity

$$(\mathbf{a}\times\mathbf{b})\cdot(\mathbf{d}\times\mathbf{c}) = (\mathbf{a}\cdot\mathbf{d})(\mathbf{b}\cdot\mathbf{c}) - (\mathbf{d}\cdot\mathbf{b})(\mathbf{a}\cdot\mathbf{c}),$$

namely,

$$(\mathbf{a}\times\mathbf{b})\cdot(\mathbf{a}\times\mathbf{c}) = \mathbf{b}\cdot\mathbf{c}, \text{ for } \mathbf{a} \text{ unimodular and orthogonal either to } \mathbf{b} \text{ or to } \mathbf{c}$$

and

$$(\mathbf{a}\times\mathbf{b})\cdot(\mathbf{d}\times\mathbf{c}) = -(\mathbf{d}\cdot\mathbf{b})(\mathbf{a}\cdot\mathbf{c}), \text{ for } \mathbf{a} \text{ orthogonal to } \mathbf{d}.$$

But,

$$\mathbf{A}_0 \partial_r \check{\mathbf{m}}_0 \cdot \check{\mathbf{m}}_1 = \mathbf{0}, \tag{6.39}$$

as is proven by taking the dot product of both sides of $(6.18)_1$ with $\check{\mathbf{m}}_0 \times \check{\mathbf{m}}_1$. Thus, integration of (6.36) gives

$$\int_{-\infty}^{+\infty} \check{\mathbf{m}}_0 \times \check{\mathbf{b}}_0 \cdot \partial_r \check{\mathbf{m}}_0 \, dr = 0. \tag{6.40}$$

Equation (6.40) is the first of the solvability conditions we are after. The other obtains quite similarly, by taking the dot product of both sides of (6.35) with $\check{\mathbf{m}}_0 \times \mathbf{e}$, integrating the resulting scalar equation over the real line, and using the identity $(6.18)_2$; all in all,

$$\int_{-\infty}^{+\infty} (\check{\mathbf{m}}_0 \times \check{\mathbf{b}}_0) \cdot (\check{\mathbf{m}}_0 \times \mathbf{e}) \, dr = 0. \tag{6.41}$$

In (6.40) and (6.41) the fields $\check{\mathbf{m}}_0$, $\partial_r \check{\mathbf{m}}_0$ are specified by (6.17) and the field $\check{\mathbf{b}}_0$ by (6.34). An explicit computation, where use is made of the fact that both $\partial_r \check{\mathbf{m}}_0$ and $\check{\mathbf{b}}_0$ are orthogonal to $\check{\mathbf{m}}_0$ and of the identity (6.29), gives (6.40) the form

$$g(\hat{v}(\mathbf{s},t) - \hat{k}(\mathbf{s},t)) + \hat{f}\,(\mathbf{s},t) = 0, \tag{6.42}$$

where the coefficient g, a constant, is defined to be

$$g = \int_{-\infty}^{+\infty} |\partial_r \check{\mathbf{m}}_0|^2 \, dr \tag{6.43}$$

and where the forcing term is

$$\hat{f}(\mathbf{s},t) = (\mathbf{h}_{\rm tot} \cdot [\![\mathbf{m}_0]\!])(\mathbf{s},t), \quad \mathbf{h}_{\rm tot} := \langle\!\langle \mathbf{h}_0 \rangle\!\rangle + \mathbf{h}_{\rm ext}. \tag{6.44}$$

Likewise, (6.41) takes the form

$$(\mathbf{e} \cdot [\![\mathbf{m}_0]\!])\, \tau\, \hat{v} + \mathbf{e} \times \mathbf{t} \cdot \Big(\pi \mathbf{h}_{\rm tot} + (\hat{\mathbf{n}} \otimes \hat{\mathbf{n}})(\pi \langle\!\langle \mathbf{m}_0 \rangle\!\rangle - g\, \mathbf{t}) \Big) = 0. \tag{6.45}$$

Now, with the use of (6.17), it is easy to evaluate the integral in (6.43) and find

$$g = 2. \tag{6.46}$$

Moreover, it follows from assumption (6.3) that the jump and mean of $\mathbf{m}_0$ across the domain wall are, respectively,

$$[\![\mathbf{m}_0]\!] = 2\mathbf{e}, \quad \langle\!\langle \mathbf{m}_0 \rangle\!\rangle = \mathbf{0}. \tag{6.47}$$

Consequently, we rewrite (6.42) and, respectively, (6.45) as follows:

$$\begin{aligned} &\hat{v}(\mathbf{s},t) - \hat{k}(\mathbf{s},t) + \hat{h}(\mathbf{s},t) = 0, \quad \hat{h}(\mathbf{s},t) := \mathbf{e} \cdot \mathbf{h}_{\rm tot}(\mathbf{s},t); \\ &2\tau\, \hat{v} + \mathbf{e} \times \mathbf{t} \cdot \big(\pi \mathbf{h}_{\rm tot} - 2\, (\hat{\mathbf{n}} \otimes \hat{\mathbf{n}}) \mathbf{t} \big) = 0. \end{aligned} \tag{6.48}$$

6.4. The evolution equations of the domain wall. The last step of our derivation consists in scaling equations (6.48) back to the original variables $\mathbf{x}$ and t. This is achieved by replacing v and k, respectively, by $L^{-1}\,T\,v$ and $L\,k$, so as to obtain

$$\nu^{-1}v - \sigma k + h = 0 \tag{6.49}$$

and

$$2\tau\nu^{-1}v + (\mathbf{e}\times\mathbf{t})\cdot\big(\pi\mathbf{h}_{\rm tot} - 2\,(\hat{\mathbf{n}}\otimes\hat{\mathbf{n}})\mathbf{t}\big) = 0, \tag{6.50}$$

where

$$\nu := \frac{1}{\mu}\left(\frac{\alpha}{\beta}\right)^{\frac{1}{2}}, \quad \sigma := (\alpha\beta)^{\frac{1}{2}}. \tag{6.51}$$

(It follows from (4.6) and (4.8) that $L = (\alpha\beta)^{1/2}$ and $T = \beta\mu$.)

Equation (6.49) determines the evolution of the domain wall $\mathcal{S}_t$, whose motion appears to be driven by its own current mean curvature and by a generally *nonlocal* forcing term. It is precisely this last feature that distinguishes (6.49) from the formally identical equations that describe the motion of sharp grain boundaries or solidification fronts according to their *mobility* ν and *surface tension* σ, under the *driving force* $f := -h$ (see, e.g., [10, 11]). From $(6.44)_2$ and $(6.48)_1$ we have that the forcing term h has the form

$$h(\mathbf{s},t) = \mathbf{e}\cdot(\langle\!\langle\mathbf{h}_0\rangle\!\rangle + \mathbf{h}_{\rm ext})(\mathbf{s},t),$$

where, given $\mathcal{S}_t$, the stray field $\mathbf{h}_0$ is the solution of the system of PDEs

$$\operatorname{curl}\mathbf{h}_0 = \mathbf{0}, \quad \operatorname{div}\mathbf{h}_0 = 0 \qquad \text{in } \mathbb{R}^3 - \mathcal{S}_t,$$

subject to the jump conditions

$$[\![\mathbf{h}_0]\!]\cdot\hat{\mathbf{n}} = 2\mathbf{e}\cdot\hat{\mathbf{n}}, \quad [\![\mathbf{h}_0]\!]\times\hat{\mathbf{n}} = \mathbf{0} \qquad \text{on } \mathcal{S}_t.$$

While (6.49) determines v, (6.50) determines $\mathbf{t}$ and hence provides information on the zeroth-order magnetization inside the domain wall. Note that the requirement that (6.50) admits a solution sets a restriction on the possible values of v. Remarkably, this restriction yields Walker's breakdown velocity when flat walls are considered.

6.5. Comparison with Walker's solution. With a view toward using Walker's solution as a benchmark for our theory, we consider the case when $\mathcal{S}_t$ is a plane parallel to the easy axis $\mathbf{e}$, with unit normal $\mathbf{c}_1$. Just as in subsection 3.2, we assume that the spatial dependence of all fields of interest is only through the first coordinate x_1, and that the easy axis coincides with $\mathbf{c}_3$; we also assume that the external field has the form (3.10), and we maintain the prescription (3.3) that the stray field $\mathbf{h}$ vanishes at $x_1 = \pm\infty$, which implies, in the present case, that $\mathbf{h}$ is everywhere null.

Under these circumstances, (6.49) becomes

$$v = \nu H = \frac{1}{\mu}\left(\frac{\alpha}{\beta}\right)^{\frac{1}{2}} H, \tag{6.52}$$

which agrees with (3.15) for $\beta \gg 1$. Moreover, with the use of the representation $(3.7)_2$ for $\mathbf{t}$, (6.50) becomes

$$v = -\nu\tau^{-1}\cos\varphi\sin\varphi. \tag{6.53}$$

Combining (6.52) and (6.53), we recover both (3.13) and the bounds (3.11) on H. Thus, even in this respect, the predictions of our theory are consistent with Walker's.

7. Concluding remarks. In saturated ferromagnets, neither magnetic domains nor domain walls have an observable, definite physical substance: their presence and shape do not seem to reflect differences in composition or texture. Thus, in principle, domain structures should correspond to specific classes of solutions to certain initial and boundary value problems based on the Gilbert equation: for rigid ferromagnets, there should be no need for any further balance principle.[5]

This is why we chose to derive our evolution equation for a domain wall from the Gilbert equation, by way of matched asymptotics. Admittedly, with such an approach, the desired result is anticipated by the very start, in that one does not prove, but assumes, that a domain wall is well described as a smooth and smoothly evolving surface. Given this, the governing equations follow, under assumptions on scaling exponents that, no matter how reasonable and well motivated, are best justified *a posteriori* by the inspection of the phenomenology they encompass.[6]

Needless to say, our results are not applicable to situations where the Gilbert equation would not hold (e.g., they would not be applicable to unsaturated ferromagnets, for which materials, however, the concept of magnetic domain would *stricto sensu* be in the need of a reformulation) or to situations that the Gilbert equation ignores, such as the *pinning* of domain walls. (Indeed, it is only after an evolution equation for domain walls is made available that a discussion of pinning can be started; the same applies to a discussion of *domain-wall junctions*.)

There have been few previous attempts to model domain-wall evolution, among which we mention the work of Slonczewsky [26] (see also [17]) and the works of Jiang [15], Maugin and Fomethe (see [19], where reference is given to other related papers by the same authors and by Maugin, alone and with different coworkers), and James [14]. This is no place for a detailed analysis of the conceptual differences between our approach and those in these papers; for such an analysis, we refer the reader to our forthcoming paper [24]. Over and beyond their differences in method and scope, what is derived in all the papers by Jiang, Maugin and Fomethe, and James is a formula for the driving force; direct geometric contributions to wall motion, such as the curvature term in (6.49), are absent (they are not in Slonczewsky's work). The other ingredient of all models in [15, 19, 14] is a functional relation

$$v = \mathcal{V}(f) \tag{7.1}$$

between normal velocity and driving force, a relation that is postulated, not derived, and is regarded as constitutive. Since wall motion is viewed as an essentially dissipative phenomenon, the form of $\mathcal{V}$ is restricted by the requirement that

$$fv \geq 0. \tag{7.2}$$

In view of the above, we can compare our formula for the driving force,

$$f = -(\langle\!\langle \mathbf{h}_0 \rangle\!\rangle + \mathbf{h}_{\text{ext}}) \cdot [\![\mathbf{m}_0]\!], \tag{7.3}$$

[5] The form of the Gilbert equation we study is consistent with thermodynamics [21]. Yet, it could happen that a dissipation principle would be of use to separate physically significant evolutions from others having only a mathematical life [3].

[6] Different governing equations would follow from scaling assumptions other than those we made. For example, $\dot{\mathbf{m}}$ appears in two addenda of the Gilbert equation (2.5), namely, $\gamma^{-1}\dot{\mathbf{m}}$ and $\mu\mathbf{m}\times\dot{\mathbf{m}}$. There are then two different characteristic times built into the equation, one for each of these terms. As the second of (4.8) makes clear, our scaling selects the "slow" time $\beta\,\mu$ and therefore excludes any account in the domain-wall evolution of the "jerking term" $\mu\mathbf{m}\times\dot{\mathbf{m}}$.

with those derived by Jiang [15], Maugin and Fomethe [19], and James [14], so specialized as to ignore mechanical deformation. We find that (7.3) agrees with Jiang's (3.26) and with the combination of Maugin and Fomethe's (4.28) and (6.2); instead, James' (92) reads, in our notations,

$$f = -[\![(\mathbf{h}_0 + \mathbf{h}_{\text{ext}}) \cdot \mathbf{m}_0]\!], \tag{7.4}$$

or rather, with the use of (6.28),

$$f = -(\langle\!\langle \mathbf{h}_0 \rangle\!\rangle + \mathbf{h}_{\text{ext}}) \cdot [\![\mathbf{m}_0]\!] - [\![\mathbf{h}_0]\!] \cdot \langle\!\langle \mathbf{m}_0 \rangle\!\rangle. \tag{7.5}$$

Appendix. In this section, we address the problem of finding a C^2-solution to the boundary value problem

$$\begin{aligned} &\mathbf{v} \times (\mathbf{v}'' + (\mathbf{v} \cdot \mathbf{e})\mathbf{e}) = \mathbf{0} \quad \text{in } \mathbb{R}, \\ &\lim_{r \to \pm\infty} \mathbf{v}(r) = \pm\mathbf{e}, \end{aligned} \tag{7.6}$$

with the constraint

$$|\mathbf{v}| = 1. \tag{7.7}$$

We begin by noting that an admissible solution of (7.6) must satisfy the following first-order differential conditions:

$$\begin{aligned} &\mathbf{v}' \cdot \mathbf{v} = 0; \\ &\mathbf{v}' \cdot (\mathbf{v} \times \mathbf{e}) = 0; \\ &|\mathbf{v}'|^2 + (\mathbf{v} \cdot \mathbf{e})^2 = \kappa^2, \text{ a constant.} \end{aligned} \tag{7.8}$$

The first of (7.8) is a straightforward consequence of (7.7). To obtain $(7.8)_2$, take the scalar product of both sides of $(7.6)_1$ with $\mathbf{e}$, use the identity $\mathbf{v} \times \mathbf{v}'' = (\mathbf{v} \times \mathbf{v}')'$, and then integrate the resulting equation with the aid of the boundary conditions $(7.6)_2$. As to $(7.8)_3$, by taking the scalar product of both sides of $(7.6)_1$ with $\mathbf{v} \times \mathbf{v}'$, obtain

$$\mathbf{v}' \cdot \mathbf{v}'' + (\mathbf{v} \cdot \mathbf{e})(\mathbf{v} \cdot \mathbf{e})' = 0, \tag{7.9}$$

then integrate.

Every solution of (7.6)–(7.7) can be given a provisional partial representation as

$$\mathbf{v}(r) = \alpha(r)\mathbf{e} + \beta(r)\mathbf{t}(r), \tag{7.10}$$

provided that, for all $r \in \mathbb{R}$,

(i) $\mathbf{t}(r)$ is a unimodular vector orthogonal to the easy axis:

$$|\mathbf{t}(r)| = 1, \quad \mathbf{t}(r) \cdot \mathbf{e} = 0; \tag{7.11}$$

(ii) the scalars $\alpha(r)$, $\beta(r)$ satisfy

$$\alpha^2(r) + \beta^2(r) = 1 \tag{7.12}$$

and, moreover,

$$\lim_{r \to \pm\infty} \alpha(r) = \pm 1, \quad \lim_{r \to \pm\infty} \beta(r) = 0. \tag{7.13}$$

Note that taking

$$|\alpha(r)| \equiv 1, \quad \beta(r) \equiv 0 \tag{7.14}$$

would not yield a continuous solution to problem (7.6)–(7.7). Needless to say, the functions α and β we introduce and manipulate here have nothing to do with, respectively, the exchange modulus and the anisotropy modulus, which everywhere else in this paper we have denoted by those same Greek letters.

With this representation for the solutions, the differential relations (7.8) become

$$\begin{aligned} &\alpha'\alpha + \beta'\beta = 0; \\ &\beta^2 \mathbf{t} \times \mathbf{t}' = \mathbf{0}; \\ &\alpha'^2 + \beta'^2 + \beta^2 |\mathbf{t}'|^2 + \alpha^2 = \kappa^2. \end{aligned} \tag{7.15}$$

A consequence of $(7.15)_1$, (7.12), and $(7.13)_2$ is that

$$(1 - \beta^2)\alpha'^2 = \beta^2 \beta'^2. \tag{7.16}$$

On the other hand, it follows from $(7.15)_2$ and $(7.11)_1$ that

$$\beta^2 \mathbf{t}' = \mathbf{0}, \tag{7.17}$$

so that $(7.15)_3$ becomes

$$\alpha'^2 + \beta'^2 + \alpha^2 = \kappa^2. \tag{7.18}$$

Substituting (7.16) in (7.18) multiplied by $(1 - \beta^2)$, we obtain

$$\beta'^2 + (1 - \beta^2)^2 = (1 - \beta^2)\kappa^2. \tag{7.19}$$

Taking the limits of (7.19) for $r \to \pm\infty$ in the light of $(7.13)_2$, we see that β'^2 tends to $(\kappa^2 - 1)$, a constant that must be null, for otherwise $\beta(r)$ could not have finite limits as $r \to \pm\infty$ and thus, in particular, it could not satisfy $(7.13)_2$. Then, since

$$\kappa^2 = 1,$$

(7.19) becomes

$$\beta'^2 = \beta^2(1 - \beta^2), \tag{7.20}$$

while $(7.15)_3$ and (7.20) imply that

$$\alpha'^2 = (1 - \alpha^2)^2. \tag{7.21}$$

It follows from (7.20) that β can never vanish. Indeed, were $\beta(r_0) = 0$ for some $r_0 \in \mathbb{R}$, then

$$\beta^2(r) = \int_{r_0}^{r} 2\beta\beta' \leq \int_{r_0}^{r} (\beta^2 + \beta'^2) \leq 2 \int_{r_0}^{r} \beta^2,$$

where the last inequality is a consequence of (7.20). With this, an application of the Gronwall lemma would lead to the unacceptable conclusion that $\beta(r) \equiv 0$ in $\mathbb{R}$ (cf. $(7.14)_2$).

But, if $\beta(r)$ is never null, then (7.17) implies that $\mathbf{t}(r)$ is a constant vector. Furthermore, $\alpha^2(r) < 1$ for all $r \in \mathbb{R}$, and (7.21) reduces to the ODE for α[7]

$$\alpha' = 1 - \alpha^2, \tag{7.22}$$

whose solutions have the form

$$\alpha(r) = \tanh(r - \rho), \quad \rho \in \mathbb{R}, \tag{7.23}$$

and satisfy the boundary conditions $(7.13)_1$.

Acknowledgments. We gratefully acknowledge the valuable comments of two referees.

REFERENCES

[1] F. Alouges and A. Soyeur, *On global weak solutions for Landau-Lifshitz equations: Existence and nonuniqueness*, Nonlinear Anal., 18 (1992), pp. 1071–1084.
[2] G. Bertotti, *Hysteresis in Magnetism*, Academic Press, New York, 1998.
[3] M. Bertsch, P. Podio-Guidugli, and V. Valente, *On the dynamics of deformable ferromagnets* I. *Global weak solutions for soft ferromagnets at rest*, Ann. Mat. Pura Appl. (4), 179 (2001), pp. 331–360.
[4] W. F. Brown, *Micromagnetics*, Krieger, Huntington, NY, 1963.
[5] G. Caginalp and P. C. Fife, *Dynamics of layered interfaces arising from phase boundaries*, SIAM J. Appl. Math., 48 (1988), pp. 506–518.
[6] A. DeSimone and P. Podio-Guidugli, *On the continuum theory of deformable ferromagnetic solids*, Arch. Rational Mech. Anal., 136 (1996), pp. 201–233.
[7] W. Eckhaus, *Asymptotic Analysis of Singular Perturbations*, North–Holland, Amsterdam, 1979.
[8] P. C. Fife, *Dynamics of Internal Layers and Diffusive Interfaces*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 53, SIAM, Philadelphia, 1988.
[9] T. L. Gilbert, *A Lagrangian formulation of the gyromagnetic equation of the magnetization fields*, Phys. Rev., 100 (1955), p. 1243.
[10] M. E. Gurtin, *Thermomechanics of Evolving Phase Boundaries in the Plane*, Oxford University Press, Oxford, UK, 1993.
[11] M. E. Gurtin, *Configurational Forces as Basic Concepts of Continuum Physics*, Springer-Verlag, Berlin, 2000.
[12] A. Hubert, *Theorie der Domänenwände in Geordneten Medien*, Springer-Verlag, Berlin, 1974.
[13] A. Hubert and R. Schäfer, *Magnetic Domains*, Springer-Verlag, Berlin, Heidelberg, New York, 2000.
[14] R. D. James, *Configurational forces in magnetism with application to the dynamics of a small-scale ferromagnetic shape memory cantilever*, Contin. Mech. Thermodyn., 14 (2002), pp. 55–86.
[15] Q. Jiang, *On the driving traction on a surface of discontinuity within a continuum in the presence of electromagnetic fields*, J. Elasticity, 34 (1994), pp. 1–21.
[16] L. D. Landau and E. Lifshitz, *On the theory of dispersion of magnetic permeability in ferromagnetic bodies*, Phys. Z. Sowjetunion, 8 (1935), pp. 135–169.
[17] A. P. Malozemoff and J. C Slonczewsky, *Magnetic Domain Walls in Bubble Materials*, Academic Press, New York, 1979.
[18] G. A. Maugin, *Continuum Mechanics of Electromagnetic Solids*, North–Holland, Amsterdam, 1988.
[19] G. A. Maugin and A. Fomethe, *Phase-transition fronts in deformable ferromagnets*, Meccanica, 32 (1997), pp. 347–362.

[7] The alternative

$$\alpha' = \alpha^2 - 1$$

would force α to be everywhere nonincreasing, a feature incompatible with the prescribed boundary conditions.

[20] F. Pistella and V. Valente, *Numerical study of the appearance of singularities in ferromagnets*, Adv. Math. Sci. Appl., 12 (2002), pp. 803–816.
[21] P. Podio-Guidugli, *On dissipation mechanisms in micromagnetics*, Eur. Phys. J. B, 19 (2001), pp. 417–424.
[22] P. Podio-Guidugli, *Concepts and problems in dynamic micromagnetics*, in Proceedings of CanCNSM 2002, Vancouver, BC, Canada, 2002.
[23] P. Podio-Guidugli and G. Tomassetti, *On the steady motions of a flat domain wall in a ferromagnet*, Eur. Phys. J. B, 26 (2002), pp. 191–198.
[24] P. Podio-Guidugli and G. Tomassetti, forthcoming, 2004.
[25] P. Podio-Guidugli and V. Valente, *Existence of global-in-time weak solutions to a modified Gilbert equation*, Nonlinear Anal., 47 (2001), pp. 147–158
[26] J. C. Slonczewsky, *Dynamics of magnetic domain walls*, Int. J. Magnetism, 2 (1972), pp. 85–97.
[27] G. Tomassetti, *Dynamics of Domain Walls in Ferromagnets*, Ph.D. thesis, Università di Roma TorVergata, Rome, Italy, 2002.
[28] G. Tomassetti, *Curved domain walls in ferromagnets*, in Proceedings of the 6th Congress of the Italian Society for Applied and Industrial Mathematics (SIMAI 2002), Chia, Italy, 2002.
[29] A. Visintin, *On Landau-Lifshitz' equation in ferromagnetism*, Japan J. Appl. Math., 2 (1985), pp. 69–84.
[30] L. R. Walker, manuscript. See J. F. Dillon, Jr., *A treatise on magnetism*, Vol. III, G. T. Rado and H. Suhl, eds., Academic Press, New York, 1963, pp. 450–453.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 1907–1932

NUMERICAL SOLUTION OF THE CAUCHY PROBLEM FOR THE STATIONARY SCHRÖDINGER EQUATION USING FADDEEV'S GREEN FUNCTION*

MASARU IKEHATA† AND SAMULI SILTANEN‡

Abstract. Numerical solution of the Cauchy problem for the stationary Schrödinger equation in a bounded two-dimensional domain is discussed. The solution algorithm is based on the properties of Faddeev's Green function. Numerical examples with computer-simulated data are presented, including an application to the inverse potential problem of electrocardiography.

Key words. Cauchy problem, stationary Schrödinger equation, exponentially growing solution, Faddeev's Green function, electrocardiography

AMS subject classifications. 15A15, 15A09, 15A23, 65N21

DOI. 10.1137/S0036139903424916

1. Introduction.

The Cauchy problem for an elliptic equation is an ill-posed problem appearing in engineering, medical imaging, and geophysics. One important application is to recover the stationary temperature inside a given body from the temperature and heat flux on the boundary of the body. Another application is the inverse problem of electrocardiography, or determination of electric voltage potential on the surface of the heart from measurements on the skin.

We consider the Cauchy problem for the stationary Schrödinger equation. Let $n = 2, 3$ and let $\Omega \subset \mathbb{R}^n$ be a bounded connected domain with Lipschitz boundary. Let $u \in H^2(\Omega)$ satisfy

$$(-\Delta + V)u = 0 \qquad \text{in } \Omega, \tag{1}$$

where $V = V(x)$ is a known, essentially bounded and complex-valued function. We denote by ν the unit outward normal vector field to $\partial\Omega$. Given a nonempty open subset $\Gamma \subset \Omega$, the pair

$$\left(u|_\Gamma, \frac{\partial u}{\partial \nu}\Big|_\Gamma \right)$$

is called the Cauchy data of u on Γ. It is well known that the Cauchy data of u on Γ uniquely determines u in Ω. See [15] for recent uniqueness and stability results of Cauchy problems for general partial differential equations. We are interested in finding an analytic formula and a regularized algorithm for calculating the value of u at a given point in Ω.

In the case $V = 0$, (1) becomes the Laplace equation. In two dimensions the Cauchy problem for the Laplace equation is equivalent to the corresponding Cauchy

*Received by the editors March 20, 2003; accepted for publication (in revised form) October 24, 2003; published electronically August 4, 2004.

http://www.siam.org/journals/siap/64-6/42491.html

†Department of Mathematics, Faculty of Engineering, Gunma University, Kiryu 376-8515, Japan (ikehata@sv1.math.sci.gunma-u.ac.jp). The research of this author was partially supported by Grant-in-Aid for Scientific Research (C)(2) (15540154) of the Japan Society for the Promotion of Science.

‡Instrumentarium Corporation, Imaging Division, P.O. Box 20, FIN-04301 Tuusula, Finland (samuli.siltanen@iki.fi). The research of this author was supported by Grant-in-Aid for JSPS Fellows (00002757) of the Japan Society for the Promotion of Science.

problem for the Cauchy–Riemann system of equations provided Ω is simply connected. Carleman [4] gave an explicit formula for calculating the value of the solution of the Cauchy–Riemann system of equations from the Cauchy data on a part of the boundary of a domain having a special shape. In [10] Goluzin and Krylov established a generalization of the formula in the simply connected domain. We refer the reader to [1, 20, 36] for other formulae and related results in complex analysis.

In the higher-dimensional case Yarmukhamedov [34, 35] gave explicit formulae of the Carleman type for the Laplace equation for special Ω and Γ. His result covers also the case when V is a constant function [37]. For constant coefficient partial differential equations several formulae of the Carleman type are described in Tarkhanov [29]. The approach is based on the uniqueness of the Cauchy problem, or equivalently, the Runge approximation property of the governing equations. The common point of their methods is the construction of special fundamental solutions $\Phi_\tau(x, y)$ for the governing equation that depend on a large parameter τ and have the following property: for a fixed $y \in \Omega$ the Cauchy data of $\Phi_\tau(\,\cdot\,, y)$ on $\partial\Omega \setminus \Gamma$ decay as $\tau \longrightarrow \infty$. Following M. M. Lavrent'ev [19], we call those fundamental solutions Carleman functions for the governing equation, Ω and Γ. Explicit construction of Carleman functions for (1) for general Ω and Γ in three dimensions is an interesting open problem.

In [13] the first author gave a formula of the Carleman type for (1) for general V and particular Ω and Γ. Here we present its minor modification given in [14]. The set Ω is the intersection of a convex open set with the half-space $x_n > 0$, and Γ is the part of $\partial\Omega$ satisfying $x_n > 0$. Let us describe the result in the two-dimensional case. The construction of the formula is divided into three steps: first, given $y \in \Omega$, let $D \subset \Omega \cap \{x_2 < y_2\}$ be the interior of a triangle with vertex at y. Second, construct the exponentially growing solution v_τ of Sylvester and Uhlmann [28] for the Schrödinger equation

$$(-\Delta + \tilde{V})v_\tau = \chi_D e^{\tau(x_2-y_2)+i\tau x_1} \quad \text{in } \mathbb{R}^2, \tag{2}$$

where $\tilde{V}$ is the zero extension of V outside Ω. Then the restriction of v_τ to Ω satisfies the equation

$$(-\Delta + V)v_\tau = \chi_D e^{\tau(x_2-y_2)+i\tau x_1} \quad \text{in } \Omega. \tag{3}$$

Third, establish the following asymptotic behavior as $\tau \to \infty$:

$$\int_D e^{\tau(x_2-y_2)+i\tau x_1} u(x)dx \sim \frac{C_D}{2\tau^2} e^{i\tau y_1} u(y), \tag{4}$$

where C_D is a nonzero constant. A combination of (3) and (4) yields the following formula for the solution u of (1):

$$u(y) = \lim_{\tau\to\infty} u_\tau(y), \tag{5}$$

where

$$u_\tau(y) := \frac{2\tau^2 e^{-i\tau y_1}}{C_D} \int_\Gamma \left(\frac{\partial u}{\partial \nu} v_\tau - \frac{\partial v_\tau}{\partial \nu} u \right) d\sigma(x). \tag{6}$$

The construction of exponentially growing solutions is based on the properties of Faddeev's Green function [8],

$$G_\zeta(x) := \frac{e^{i\zeta\cdot x}}{(2\pi)^2} \int_{\mathbb{R}^2} \frac{e^{ix\cdot\xi}}{|\xi|^2 + 2\zeta\cdot\xi} d\xi, \qquad \zeta \in \mathbb{C}^2 \setminus 0, \quad \zeta\cdot\zeta = 0. \tag{7}$$

Therefore one may consider (5) as a new application of G_ζ in addition to inverse boundary value problems and inverse scattering problems (see [31] for the problems).

We present a numerical implementation of (6) in dimension two. The ill-posed Cauchy problem is regularized by choosing τ small enough for the numerical computation to be robust against noise in the data. The main computational task is numerical evaluation of the exponentially growing solutions and their normal derivatives. For this we present an improvement of the algorithm for G_ζ given in [26] and write the derivatives of G_ζ in terms of itself and explicit formulae. The exponentially growing solutions and their derivatives can be computed combining the above with the algorithm introduced in [22] (a modification of the fast Lippmann–Schwinger equation solver of Vainikko [32, 24]). These algorithms have independent interest in the fields of electrical impedance tomography and inverse scattering.

We review some earlier numerical works on the Cauchy problem for the elliptic equation. The constant coefficient case has been studied by Leitão [21], Berntsson and Eldén [2], Cheng et al. [6], Kabanikhin and Karchevsky [16], and Háo and Lesnic [12]. The method of quasi reversibility proposed by Lattés and Lions [18] covers the variable coefficient case, and Klibanov and Santosa [17] gave an explicit estimate of the convergence rate. However, in the proof of the convergence, the uniqueness of the Cauchy problem is essential.

The present solution algorithm does not require uniqueness of the Cauchy problem for the convergence proof, and its implementation does not involve solution of boundary value problems. The computational effort is divided into two parts: first, evaluation of $v_\tau|_\Gamma$ and $\partial v_\tau/\partial\nu|_\Gamma$ for given y, V, Ω, and τ and, second, evaluation of u_τ for given Cauchy data. The second computation is very fast since it is essentially linear filtering of the data. The method can thus be applied to real-time monitoring of fixed targets with changing Cauchy data.

This paper is organized as follows. In section 2 we give details of the reconstruction formula. In section 3 we discuss the stability of our method when applied to noisy data. In section 4 we describe some properties of Faddeev's Green function and show how to evaluate it numerically. In section 5 we describe a numerical implementation of (6). We illustrate the algorithm in sections 6 and 7 by numerical examples using computer-simulated noisy data.

2. Background of the method. Throughout the paper we assume that Ω is the intersection of the open unit disc $B = \{x \in \mathbb{R}^2 \,|\, |x| < 1\}$ with the half-plane $\{x \in \mathbb{R}^2 \,|\, x_2 > t\}$ with $-1 < t < 1$ and that Cauchy data is given on $\Gamma = \{x \in \partial B \,|\, x_2 > t\}$. There is no loss of generality with this simplification of the geometry of Ω since any simply connected domain with a smooth boundary can be conformally mapped to the case when $t = 0$ (see Figure 2.1). However, in section 7 we will consider the case when $t \neq 0$.

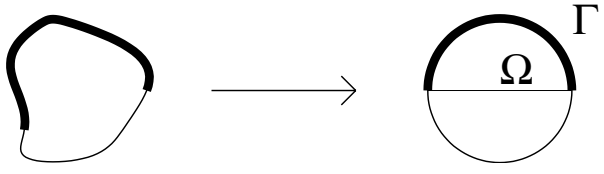


FIG. 2.1. *Conformal mapping of a domain onto the upper half of the unit disc.*

Let u be an $H^2(\Omega)$ solution of the stationary Schrödinger equation

$$-\Delta u + Vu = 0 \quad \text{in } \Omega, \tag{8}$$

where V belongs to the following class.

DEFINITION 2.1. *The potential V is admissible if V is C^2 in each component of $\Omega \setminus c$, where $c = \cup_{j=1}^{J} c_j$ with $c_j \subset \Omega$ compact, piecewise C^1 curves for which $c_i \cap c_j$ is a discrete set if $i \neq j$.*

The set D mentioned in the introduction is defined as follows.

DEFINITION 2.2. *Given $y = (y_1, y_2) \in \Omega$, let $L, p, q \in \mathbb{R}$ satisfy $0 < L \leq y_2 - t$ and*

$$-\sqrt{1-(y_2-L)^2} < y_1 + p < y_1 + q < \sqrt{1-(y_2-L)^2}. \tag{9}$$

We call the interior of the triangle with vertices

$$y = (y_1, y_2),\ y' = (y_1 + p, y_2 - L),\ y'' = (y_1 + q, y_2 - L) \tag{10}$$

a triangular patch D at y (see Figure 2.2). Although not explicitly indicated, D depends on the point y and the parameters L, p, q. D is an open subset of Ω and satisfies $D \subset \{x \,|\, x_2 < y_2\}$.

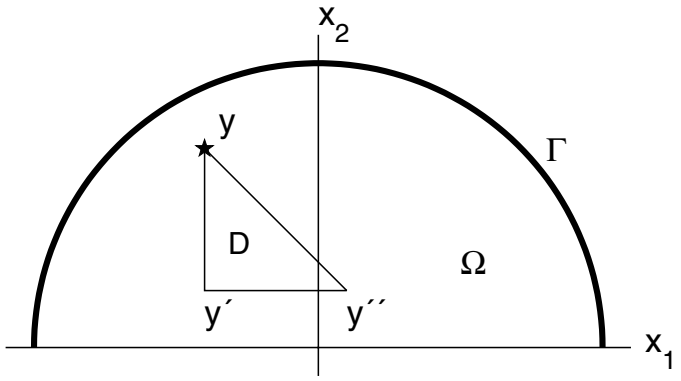


FIG. 2.2. *Geometry of the problem for $t = 0$. Domain Ω is the intersection of the unit disc with the upper half-space $x_2 > 0$. The set $\Gamma \subset \partial\Omega$ is drawn as a thick curve. The reconstruction point y is marked with a star, and one possible choice for the triangular patch D is drawn below y.*

In what follows, we take for simplicity $t = 0$.

Let χ_D denote the characteristic function of D. Let $\tilde{V}$ denote the zero extension of V outside Ω. By Sylvester and Uhlmann [28], for large $\tau \gg 1$ there exists the unique solution w_τ of the integral equation

$$\begin{aligned} w_\tau(x) + \int_{\mathbb{R}^2} g_\tau(x-z)\{\tilde{V}(z) - \chi_D(z)\} w_\tau(z) dz \\ = -\int_{\mathbb{R}^2} g_\tau(x-z)\{\tilde{V}(z) - \chi_D(z)\} dz \end{aligned} \tag{11}$$

such that for $-1 < \delta < 0$

$$\|w_\tau\|_\delta \equiv \left(\int_{\mathbb{R}^2} |w_\tau(x)|^2 (1+|x|^2)^\delta dx \right)^{\frac{1}{2}} = O\left(\frac{1}{\tau}\right). \tag{12}$$

Here g_τ is defined for any $\tau > 0$ by

$$g_\tau(x) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \frac{e^{ix\cdot\xi} d\xi}{|\xi|^2 + 2\tau(\xi_1 - i\xi_2)}. \tag{13}$$

Note that g_τ satisfies $\{\Delta + 2i\tau(\partial_1 - i\partial_2)\} g_\tau(x) + \delta(x) = 0$ in $\mathbb{R}^2$.

For any $\tau > 0$ the function

$$v_\tau^0(x) = e^{\tau(x_2 - y_2)} e^{i\tau x_1}, \quad x \in \mathbb{R}^2,$$

is harmonic and has the following properties:

If $x_2 > y_2$, then $|v_\tau^0|$ is exponentially growing as $\tau \longrightarrow \infty$.

If $x_2 < y_2$, then $|v_\tau^0|$ is exponentially decaying as $\tau \longrightarrow \infty$.

We see that

$$v_\tau^0(x) g_\tau(x) = e^{-\tau y_2} G_{(\tau, -i\tau)}(x), \tag{14}$$

where $G_{(\tau,-i\tau)}$ is Faddeev's Green function (7). Then one knows that the functions $v'_\tau \equiv v_\tau^0(1 + w_\tau)$ become the solutions of the equation

$$-\Delta v'_\tau + \tilde{V} v'_\tau = \chi_D v'_\tau \quad \text{in } \mathbb{R}^2$$

and satisfy $v'_\tau \sim v_\tau^0$ as $\tau \to \infty$ in the sense that (12) holds. Define

$$v_\tau = v'_\tau|_\Omega. \tag{15}$$

Since $v'_\tau \in H^2_{\mathrm{loc}}(\mathbb{R}^2)$, v_τ is an $H^2(\Omega)$ solution of the equation

$$-\Delta v_\tau + V v_\tau = \chi_D v_\tau \quad \text{in } \Omega.$$

Now, by Theorem 2.1 of [13]

$$u(y) = \lim_{\tau\to\infty} u_\tau(y) = \lim_{\tau\to\infty} \frac{2\tau^2 e^{-i\tau y_1}}{C_D} \int_\Gamma \left(\frac{\partial u}{\partial \nu} v_\tau - \frac{\partial v_\tau}{\partial \nu} u \right) d\sigma(x), \tag{16}$$

where

$$C_D := \frac{2L(q-p)}{(L - ip)(L - iq)}. \tag{17}$$

The proof uses the estimate $\|w_\tau\|_{L^\infty(\Omega)} = O(\frac{1}{\tau})$ which comes from (2.11) of Proposition 2.3 in [27] and the well-known fact that the growth rate of $\|w_\tau\|_{H^2(\Omega)}$ with respect to τ is at most algebraic.

Note that one can give a simpler choice of v_τ appearing in (16). This is done in [14]. More precisely, for large $\tau \gg 1$ there exists the unique solution w'_τ of the integral equation

$$w'_\tau(x) + \int_{\mathbb{R}^2} g_\tau(x - z) \tilde{V}(z) w'_\tau(z) dz = \int_{\mathbb{R}^2} g_\tau(x - z) \chi_D(z) dz \tag{18}$$

such that for $-1 < \delta < 0$ we have $\|w'_\tau\|_\delta = O(1/\tau)$. Then

$$v''_\tau = v_\tau^0 w'_\tau \tag{19}$$

satisfies the equation $-\Delta v''_\tau + \tilde{V} v''_\tau = \chi_D v_\tau^0$ in $\mathbb{R}^2$. A trivial modification of the proof of Theorem 2.1 of [13] shows that (16) holds with the choice

$$v_\tau = v''_\tau|_\Omega. \tag{20}$$

In this case we use only the algebraic growth of $\|w'_\tau\|_{H^2(\Omega)}$ with respect to τ. Hereafter we consider v_τ given by (20) and not by (15).

3. Stability of the method. In this section we consider the case when the Cauchy data of u on Γ contains noise. Let $M > 0$ satisfy

$$\|V\|_{L^\infty(\Omega)} \leq M.$$

Given $y \in \Omega$ choose a triangular patch D at y according to Definition 2.2. Fix $\delta \in]-1, 0[$. Using a perturbation argument, (2.7) in Proposition 2.1 in [23], and the argument made for the proof of (2.53) in Lemma 2.11 of [23], one obtains the unique solvability of (18). More precisely, there exist positive constants $C_1(M)$ and $C_2(M)$ (independent of y and D) such that for $\tau > C_1(M)$ the equation (18) has a unique solution w'_τ satisfying the estimate

$$\tau\|w'_\tau\|_\delta + \|\nabla w'_\tau\|_\delta + \tau^{-1}\sum_{i,j=1}^{2}\|\partial_i\partial_j w'_\tau\|_\delta \leq C_2(M). \tag{21}$$

For v_τ given by (20) and $(f, g) \in L^2(\Gamma) \times L^2(\Gamma)$ define

$$S_\tau(f, g)(y) = \frac{2\tau^2 e^{-i\tau y_1}}{C_D}\int_\Gamma \left(gv_\tau - \frac{\partial v_\tau}{\partial \nu} f\right) d\sigma.$$

Let $E = (E_1, E_2) \in L^2(\Gamma) \times L^2(\Gamma)$ be additive noise on the Cauchy data on Γ. Denote $\|E\| = (\|E_1\|^2_{L^2(\Gamma)} + \|E_2\|^2_{L^2(\Gamma)})^{1/2}$.

The problem is to calculate an approximate value of $u(y)$ from

$$S_\tau\left(u|_\Gamma + E_1, \left.\frac{\partial u}{\partial \nu}\right|_\Gamma + E_2\right)(y)$$

with $\tau > C_1(M)$ when $\|E\|$ is small. One cannot choose extremely large τ since such a selection enlarges the effect of noise. The suitable choice of τ is just the problem of regularizing the formula (16).

In order to describe a result quantitatively and show the effect of the choice of D we prepare two lemmas.

LEMMA 3.1. *Assume that u belongs to the space of Hölder continuous functions $C^{0,\theta}(\overline{D})$ with $0 < \theta \leq 1$. Then for all $\tau > 0$ we have*

$$\begin{aligned}&\left|\tau^2 e^{-\tau(y_2+iy_1)}\int_D u(x)e^{\tau(x_2+ix_1)}dx - \frac{C_D}{2}u(y)\right| \\ &\leq \frac{q-p}{L}\|u\|_{C^{0,\theta}(\overline{D})}\left\{(\tau L+1)e^{-\tau L} + \left(\frac{\operatorname{diam} D}{L}\right)^\theta \frac{C_\theta}{\tau^\theta}\right\},\end{aligned} \tag{22}$$

where C_θ is a positive constant depending only on θ, and q, p, L are as in Definition 2.2.

Proof. See [13, Lemma 2]. □

LEMMA 3.2. *Let $0 < \epsilon < 1$ and y satisfy $y_2 > \epsilon$. There exists such a positive constant $C_{M,\epsilon}$ depending on M and ϵ that for any $u \in H^2(\Omega)$ and v_τ with $\tau > C_1(M)$ given by (20) the following estimate holds:*

$$\left|\tau^2 e^{-i\tau y_1}\int_{\partial\Omega\setminus\Gamma}\left(\frac{\partial u}{\partial \nu}v_\tau - \frac{\partial v_\tau}{\partial \nu}u\right)d\sigma\right| \leq C_{M,\epsilon}\|u\|_{H^2(\Omega)}\tau^3 e^{-\frac{\tau\epsilon}{2}}. \tag{23}$$

Proof. Using (19), (20), (21), and the estimate $|e^{\tau(x_2-y_2)}| \le e^{-\tau\varepsilon/2}$ for $x \in \Omega_\varepsilon$ we obtain

$$\|v_\tau\|_{H^2(\Omega_\epsilon)} \le C'_M \tau e^{-\frac{\tau\epsilon}{2}}, \tag{24}$$

where $\Omega_\epsilon = \{x \in \Omega \,|\, 0 < x_2 < \epsilon/2\}$ and C'_M is a positive constant independent of ϵ. Since $\partial\Omega_\epsilon$ is Lipschitz, we have the following consequence of a general trace theorem [11]: for any $\phi \in H^2(\Omega_\varepsilon)$ we have

$$\|\nabla\phi\|_{L^2(\partial\Omega_\varepsilon)} + \|\phi\|_{L^2(\partial\Omega_\varepsilon)} \le C_\varepsilon \|\phi\|_{H^2(\Omega_\varepsilon)}. \tag{25}$$

Combining (24) and (25) yields (23). ∎

Now we discuss the problem mentioned above. Integration by parts yields

$$\tau^2 e^{-i\tau y_1} \int_\Gamma \left(\frac{\partial u}{\partial \nu} v_\tau - \frac{\partial v_\tau}{\partial \nu} u \right) d\sigma = \tau^2 e^{-\tau(y_2+iy_1)} \int_D u(x) e^{\tau(x_2+ix_1)} dx + \tau^2 e^{-i\tau y_1} \int_{\partial\Omega\setminus\Gamma} \left(\frac{\partial u}{\partial \nu} v_\tau - \frac{\partial v_\tau}{\partial \nu} u \right) d\sigma.$$

Recalling (6), we rewrite

$$\frac{C_D}{2} u_\tau(y) = \frac{C_D}{2} u(y) + \left\{ \tau^2 e^{-\tau(y_2+iy_1)} \int_D u(x) e^{\tau(x_2+ix_1)} dx - \frac{C_D}{2} u(y) \right\} + \tau^2 e^{-i\tau y_1} \int_{\partial\Omega\setminus\Gamma} \left(\frac{\partial u}{\partial \nu} v_\tau - \frac{\partial v_\tau}{\partial \nu} u \right) d\sigma.$$

This together with (22) and (23) yields

$$|u_\tau(y) - u(y)| \frac{|C_D|}{2} \le \frac{q-p}{L} \|u\|_{C^{0,\theta}(\overline{D})} \left\{ (\tau L + 1) e^{-\tau L} + \left(\frac{\operatorname{diam} D}{L} \right)^\theta \frac{C_\theta}{\tau^\theta} \right\} + C_{M,\epsilon} \|u\|_{H^2(\Omega)} \tau^3 e^{-\frac{\tau\epsilon}{2}}. \tag{26}$$

This is an error estimate of the formula (16), and the order of the convergence is $O(\tau^{-\theta})$ as $\tau \longrightarrow \infty$.

Write

$$S_\tau \left(u|_\Gamma + E_1, \left. \frac{\partial u}{\partial \nu} \right|_\Gamma + E_2 \right)(y) = u_\tau(y) + S_\tau(E_1, E_2)(y).$$

From (17) one has

$$|C_D| = \frac{2L(q-p)}{\sqrt{L^2+p^2}\sqrt{L^2+q^2}} \le 2. \tag{27}$$

Recalling (19) and (20), from (21) we have

$$\|v_\tau\|_{H^2(\Omega)} \le C'_M \tau e^{\tau(1-y_2)}, \qquad \tau > C_1(M), \tag{28}$$

where C'_M is a positive constant. Using (27), (28), and the trace theorem we see that there exists a positive constant C''_M such that

$$\left| S_\tau \left(u|_\Gamma + E_1, \left. \frac{\partial u}{\partial \nu} \right|_\Gamma + E_2 \right)(y) - u_\tau(y) \right| \frac{|C_D|}{2} \le C''_M \|E\| \tau^3 e^{\tau(1-y_2)}. \tag{29}$$

Let $A > 0$ satisfy

$$\|u\|_{H^2(\Omega)} \leq A. \tag{30}$$

By the Sobolev imbedding theorem, one can choose a positive constant C'_θ depending on $0 < \theta < 1$ such that for all $v \in H^2(\Omega)$

$$\|v\|_{C^{0,\theta}(\overline{\Omega})} \leq C'_\theta \|v\|_{H^2(\Omega)}. \tag{31}$$

Now from (26), (29), (30), and (31) we obtain

$$\begin{aligned}
&\sup_{\|E\| \leq \eta} \left| S_\tau \left(u|_\Gamma + E_1, \left. \frac{\partial u}{\partial \nu} \right|_\Gamma + E_2 \right)(y) - u(y) \right| |C_D| \\
&\quad \leq \frac{q-p}{L} C'_\theta A \left\{ (\tau L + 1) e^{-\tau L} + \left(\frac{\operatorname{diam} D}{L} \right)^\theta \frac{C_\theta}{\tau^\theta} \right\} \\
&\quad + C_{M,\epsilon} A \tau^3 e^{-\frac{\tau \epsilon}{2}} + C''_M \eta \tau^3 e^{\tau(1-y_2)}.
\end{aligned} \tag{32}$$

The last term of this right-hand side estimates the speed of enlarging the effect of noise. We choose a suitable $\tau > C_1(M)$ depending on η in such a way that for this τ the right-hand side converges to zero as $\eta \longrightarrow 0$. There should be several choices of τ. Here we ignore the exponential decaying terms in the right-hand side of (32) and consider minimizing the remaining term $f(\tau; \eta)$ with respect to $\tau > C_1(M)$:

$$f(\tau; \eta) = \frac{\alpha}{\tau^\theta} + \beta \eta \tau^3 e^{\tau(1-y_2)},$$

where

$$\alpha = \frac{q-p}{L} C'_\theta A \left(\frac{\operatorname{diam} D}{L} \right)^\theta C_\theta; \qquad \beta = C''_M.$$

Since $\lim_{\tau \longrightarrow 0} f(\tau; \eta) = \infty$ and $\lim_{\tau \longrightarrow \infty} f(\tau; \eta) = \infty$, $f(\tau; \eta)$ attains its minimum value in a point in the interval $]0, \infty[$. The point has to satisfy the equation $f'(\tau; \eta) = 0$. This is equivalent to the equation

$$\tau^{\theta+3} \{3 + (1 - y_2)\tau\} e^{\tau(1-y_2)} = \frac{\alpha \theta}{\beta \eta}. \tag{33}$$

This equation has a unique positive solution and can be written as

$$\tau = \tau \left(\frac{\alpha \theta}{\beta \eta}, y_2 \right) = \frac{1}{1 - y_2} w \left(\frac{\alpha \theta}{\beta \eta} (1 - y_2)^{\theta+3} \right),$$

where $w = w(s), s > 0$, is the unique positive solution of the equation

$$w^{\theta+3}(3 + w) e^w = s. \tag{34}$$

If $\tau(\alpha\beta/\beta\eta, y_2) \leq C_1(M)$, then $f(\tau; \eta)$ does not attain its greatest lower bound in the interval $]C_1(M), \infty[$. So we assume that the magnitude of the noise η satisfies

$$\tau \left(\frac{\alpha \theta}{\beta \eta}, y_2 \right) > C_1(M).$$

This is equivalent to the inequality

$$C_1(M)^{\theta+3}\{3+(1-y_2)C_1(M)\}e^{C_1(M)(1-y_2)} < \frac{\alpha\theta}{\beta\eta},$$

that is,

$$\eta < \frac{\alpha\theta C_1(M)^{-(\theta+3)}e^{-C_1(M)(1-y_2)}}{\beta\{3+(1-y_2)C_1(M)\}}. \tag{35}$$

From (33) we have

$$\min_{\tau>C_1(M)} f(\tau;\eta) = f\left(\tau\left(\frac{\alpha\theta}{\beta\eta},y_2\right);\eta\right) = \frac{\alpha}{\tau(\frac{\alpha\theta}{\beta\eta},y_2)^{\theta}}\left\{1+\frac{\theta}{3+(1-y_2)\tau(\frac{\alpha\theta}{\beta\eta},y_2)}\right\}; \tag{36}$$

for $\tau = \tau(\alpha\theta/\beta\eta, y_2)$,

$$e^{-\tau L} = \left(\frac{\beta\eta}{\alpha\theta}\right)^{\frac{L}{1-y_2}}\{\tau^{\theta+3}(3+(1-y_2)\tau)\}^{\frac{L}{1-y_2}}, \tag{37}$$

$$e^{-\tau\epsilon/2} = \left(\frac{\beta\eta}{\alpha\theta}\right)^{\frac{\epsilon/2}{1-y_2}}\{\tau^{\theta+3}(3+(1-y_2)\tau)\}^{\frac{\epsilon/2}{1-y_2}}. \tag{38}$$

It is easy to see that, from (34), we have $w(s)\sim\log s$ as $s\longrightarrow\infty$, and one concludes that, as $\eta\longrightarrow 0$,

$$\tau\left(\frac{\alpha\theta}{\beta\eta},y_2\right) \sim \frac{1}{1-y_2}\log\left\{\frac{\alpha\theta}{\beta\eta}(1-y_2)^{\theta+3}\right\}. \tag{39}$$

Therefore the order of blowing up of τ is $|\log\eta|$ as $\eta\longrightarrow 0$. Moreover, from (34) one knows that $w(s)\sim(s/3)^{1/(\theta+3)}$ as $s\longrightarrow 0$. Then, for fixed η that satisfies the condition

$$\eta < \min_{y_2>\epsilon}\frac{\alpha\theta C_1(M)^{-(\theta+3)}e^{-C_1(M)(1-y_2)}}{\beta\{3+(1-y_2)C_1(M)\}} = \frac{\alpha\theta C_1(M)^{-(\theta+3)}e^{-C_1(M)(1-\epsilon)}}{\beta\{3+(1-\epsilon)C_1(M)\}},$$

we obtain, as $y_2\longrightarrow 1$,

$$\tau\left(\frac{\alpha\theta}{\beta\eta},y_2\right)\sim\left(\frac{\alpha\theta}{3\beta\eta}\right)^{\frac{1}{\theta+3}}.$$

Note that from (36), (37), (38), and (39) we obtain, for $\tau=\tau(\alpha\theta/\beta\eta,y_2)$,

$$\sup_{\|E\|\le\eta}\left|S_\tau\left(u|_\Gamma+E_1,\left.\frac{\partial u}{\partial\nu}\right|_\Gamma+E_2\right)(y)-u(y)\right||C_D| = O(|\log\eta|^{-\theta})$$

as $\eta\longrightarrow 0$. This is a regularized formula of (16).

It should be noted that the above type of argument for choosing τ is due to Lavrent'ev [19]. Therein he gave a regularization of Carleman's original formula.

We remark also that the conformal mapping depicted in Figure 2.1 deforms the potential in such a way that the bound M can become very large. This in turn makes the error estimates worse and can have an impact on the quality of the numerical solution.

4. Faddeev's Green function.

4.1. Definitions and basic properties. Consider the differential operator $\Delta_\zeta := \Delta + 2i\zeta \cdot \nabla$ with $\zeta \in \mathbb{C}^2 \setminus 0$ satisfying $\zeta \cdot \zeta = 0$. Any such ζ can be written in the form $\zeta = (k, \pm ik)$ for some $k \in \mathbb{C} \setminus 0$. We consider fundamental solutions

$$g_k^\pm(x) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \frac{e^{ix\cdot\xi}}{|\xi|^2 + 2k(\xi_1 \pm i\xi_2)} d\xi, \qquad k \in \mathbb{C} \setminus 0, \tag{40}$$

satisfying

$$-\Delta_{(k,\pm ik)} g_k^\pm(x) = \Big(-\Delta - 2ik\Big(\frac{\partial}{\partial x_1} \pm i\frac{\partial}{\partial x_2}\Big)\Big) g_k^\pm(x) = \delta(x). \tag{41}$$

Then Faddeev's Green function (7) takes the form

$$G_{(k,\pm ik)}(x) = e^{ik(x_1 \pm ix_2)} g_k^\pm(x), \qquad k \in \mathbb{C} \setminus 0. \tag{42}$$

We see from (40) that the two types of fundamental solutions are related by

$$g_k^-(x) = \overline{g_{\bar k}^+(-x)}. \tag{43}$$

Moreover, coordinate changes in (40) give the following symmetries:

$$g_k^+(x) = g_1^+(kx), \qquad g_k^+(x) = \overline{g_{\bar k}^+(-\bar x)}, \qquad g_k^+(x) = e_{-k}(x)\overline{g_k^+(x)}, \tag{44}$$

where $e_{-k}(x) = \exp(-i(kx + \bar k \bar x))$. It is easy to see from (13), (43), and (44) that

$$g_\tau(x) = g_\tau^-(x) = \overline{g_\tau^+(-x)} = \overline{g_1^+(-\tau x)}. \tag{45}$$

4.2. Numerical evaluation of g_1^+. We improve here the algorithm for g_1^+ given in [25, 26]. Divide the plane into disjoint regions $D_1, \dots, D_7$ as in Figure 4.1. We describe how to numerically evaluate $g_1^+(x)$ accurately in each region.

In region $D_1 = \{|x| \le R_1\}$ with $R_1 = 5.5$ we use formulae (3.10) and (3.12) of [3]:

$$g_1^+(x) = -\frac{e^{-ix}}{4\pi}\Big(2\gamma + \log|x|^2 + \sum_{n=1}^{\infty} \frac{(ix)^n + (-i\bar x)^n}{nn!}\Big), \tag{46}$$

where $\gamma \approx 0.577215665$ is the Euler–Mascheroni constant. The infinite sum in (46) is truncated at $n = 23$.

In region D_3 we use formula (82) of [25]:

$$g_1(x) = \frac{e^{-ix_1}}{2\pi} \mathrm{Re}\left[-e^{ix_1} \sum_{j=0}^{N} \frac{j!}{(ix)^{j+1}} + \frac{(N+1)! e^{ix_1}}{(-x)^{N+1}} \int_0^\infty \frac{e^{-t(x_1+ix_2)}}{(t-i)^{N+2}} dt \right]. \tag{47}$$

We use $N = 6$ and implement the one-dimensional integration of the exponentially decaying integrand with Gaussian quadrature.

For region D_2 we modify (47) using residue calculus:

$$\int_0^\infty \frac{e^{-ix_2 t - x_1 t}}{(t-i)^{N+2}} dt = (1+i) \int_0^\infty \frac{e^{-is(x_2+x_1) + s(x_2 - x_1)}}{(s + is - i)^{N+2}} ds. \tag{48}$$

This modification ensures exponential decay of the integrand.

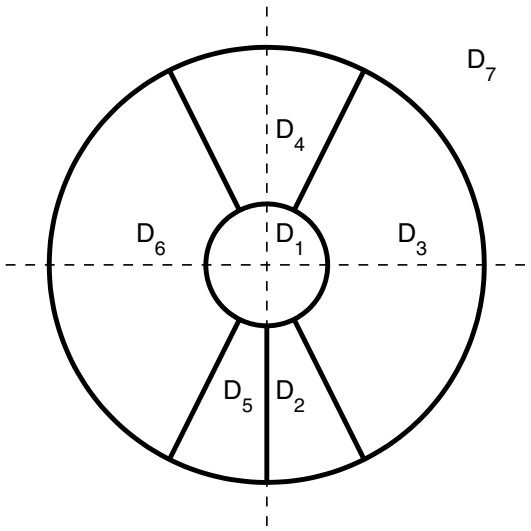


FIG. 4.1. *Computational regions $D_1, \dots, D_7$ dividing the plane into disjoint parts. The radii of the two circles are $R_1 = 5.5$ and $R_2 = 25$. The slopes of the skew lines dividing regions are either 2 or -2. It is irrelevant for the algorithm how the boundary points are divided between the regions.*

For region D_4 we modify (47) using residue calculus:

$$\int_0^\infty \frac{e^{-ix_2t-x_1t}}{(t-i)^{N+2}}dt = -i\int_0^\infty \frac{e^{-x_2s+ix_1s}}{(-is-i)^{N+2}}ds. \tag{49}$$

Again, the modified integrand decays exponentially.

For regions D_5 and D_6 we use the reflectional symmetry

$$g_1^+(-x_1, x_2) = \overline{g_1^+(x_1, x_2)} \tag{50}$$

and the algorithms for reflected regions D_2 and D_3 described above.

In region $D_7 = \{|x| \geq R_2\}$ with $R_2 = 25$ we set $N = 9$ and ignore the term with the integral in (47).

Let us comment on the choice of the radii R_1, R_2. The choice of R_1 is a trade-off: if R_1 is small, then only a few terms are needed in the truncated power series (46) to achieve desired accuracy, but on the other hand, the numerical integrations in formulae (47), (48), and (49) require many quadrature points to achieve the same accuracy. The choice $R_1 = 5.5$ gives a good balance but is not proven to be optimal. For radii $R_1 > 1$ formula (46) leads to faster computation and less memory consumption than the previous approach based on the Poisson kernel used in [25, 26]. The choice of R_2 is a similar trade-off between accuracy and computational speed.

4.3. Derivatives. As shown in [25], we can write derivatives of $g_k^\pm$ as follows.

LEMMA 4.1. *Define the functions $g_k^\pm(x)$ by (40) for $k \in \mathbb{C} \setminus 0$. Then*

$$\frac{\partial g_k^+}{\partial x_1}(x) = -\frac{1}{4\pi x} - \frac{e_{-k}(x)}{4\pi \bar{x}} - ikg_k^+(x), \tag{51}$$

$$\frac{\partial g_k^+}{\partial x_2}(x) = +\frac{1}{4\pi i x} - \frac{e_{-k}(x)}{4\pi i \bar{x}} + kg_k^+(x), \tag{52}$$

$$\frac{\partial g_k^-}{\partial x_1}(x) = -\frac{1}{4\pi\bar{x}} - \frac{e_{-\bar{k}}(x)}{4\pi x} - ikg_k^-(x), \tag{53}$$

$$\frac{\partial g_k^-}{\partial x_2}(x) = -\frac{1}{4\pi i\bar{x}} + \frac{e_{-\bar{k}}(x)}{4\pi ix} - kg_k^-(x), \tag{54}$$

where $e_{-k}(x) = \exp(-i(kx + \bar{k}\bar{x}))$, $x = x_1 + ix_2$, *and* $\bar{x} = x_1 - ix_2$.

Proof. By (43) and (44) it is enough to consider g_1^- and apply the chain rule.

Denote $\partial = (\partial/\partial x_1 - i\partial/\partial x_2)/2$. Let us compute $\partial g_1^-(x)$. By (40) we have

$$\partial g_1^-(x) = \frac{i}{2}\frac{1}{(2\pi)^2}\int_{\mathbb{R}^2}\frac{e^{ix\cdot\xi}}{\xi_1 + i\xi_2 + 2}d\xi = \frac{ie^{-2ix_1}}{2(2\pi)^2}\int_{\mathbb{R}^2}\frac{e^{ix\cdot\xi}}{\xi_1 + i\xi_2}d\xi. \tag{55}$$

Furthermore,

$$\frac{2}{i}\frac{1}{(2\pi)^2}\int_{\mathbb{R}^2}\frac{e^{ix\cdot\xi}}{\xi_1 + i\xi_2}d\xi = \frac{1}{\pi(x_1 + ix_2)}. \tag{56}$$

Combining (55) and (56) we get

$$\partial g_1^-(x) = -\frac{e^{-i2x_1}}{4\pi(x_1 + ix_2)}. \tag{57}$$

Next we determine $\bar{\partial} g_1^-(x)$. Combining (57) and (44) gives

$$-\frac{e^{i2x_1}}{4\pi(x_1 - ix_2)} = \overline{\partial g_1^-(x)} = \bar{\partial}\,\overline{g_1^-(x)} = \bar{\partial}e^{i(x+\bar{x})}g_1^-(x) = ie^{i2x_1}g_1^-(x) + e^{i2x_1}\bar{\partial}g_1^-(x),$$

and we have

$$\bar{\partial}g_1^-(x) = -\frac{1}{4\pi(x_1 - ix_2)} - ig_1^-(x). \tag{58}$$

Now formulae (57) and (58) yield the claim for $\tau = 1$ since $\partial_1 g_1^- = \partial g_1^- + \bar{\partial}g_1^-$ and $\partial_2 g_1^- = -i(\bar{\partial}g_1^- - \partial g_1^-)$. ∎

We remark that with formulae (51)–(54) any derivatives of Faddeev's Green functions can be written in terms of the Green functions themselves and explicit expressions. For instance,

$$\bar{\partial}G_{(\tau,-i\tau)}(x) = \bar{\partial}[e^{i\tau\bar{x}}g_\tau^-(x)] = e^{i\tau\bar{x}}[i\tau g_\tau^-(x) + \bar{\partial}g_\tau^-(x)] = -\frac{e^{i\tau(x_1 - ix_2)}}{4\pi(x_1 - ix_2)}. \tag{59}$$

This indicates a relationship between Faddeev's Green function and Fok–Kuni's Carleman function in the complex domain [9].

5. Numerical solution of the Cauchy problem. We discuss step by step the numerical implementation of formula (6) with fixed $y \in \Omega$.

5.1. Integration on Γ. We must choose a numerical quadrature for Γ. This is a collection of points $x^{(k)} \in \Gamma$ with $k = 1, \ldots, K$ and corresponding weights $w^{(k)}$ satisfying

$$\int_\Gamma f d\sigma \approx \sum_{k=1}^{K} w^{(k)} f(x^{(k)}). \tag{60}$$

Suitable choices are, e.g., Simpson's rule or Gaussian quadrature.

5.2. Discussion of data. We must evaluate the Cauchy data $u|_\Gamma$ and $\frac{\partial u}{\partial \nu}|_\Gamma$ on the quadrature points $x^{(k)} \in \Gamma$. How this can be done depends on the way the data are given in a particular application. We discuss here one possibility for evaluating the trace; the normal derivative can be treated similarly.

Assume that our knowledge of $u|_\Gamma$ is a finite collection of noisy point samples:

$$m_j := u(\tilde{z}^{(j)}) + \varepsilon_j, \qquad \tilde{z}^{(j)} \in \Gamma, \quad j = 1, \dots, J_0, \tag{61}$$

where ε_j for $j = 1, \dots, J_0$ are independent Gaussian, real-valued, zero-mean random variables with standard deviation $\sigma > 0$.

Define $z^{(k)} = (\cos\theta_k, \sin\theta_k) \in \Gamma$ with $\theta_k = (k-1)\pi/(J-1)$ with $k = 1, \dots, J$ and $J \geq J_0$. Assume that the data points $\tilde{z}^{(j)}$ are included in the evaluation points:

$$\tilde{z}^{(j)} = z^{(k_j)}, \qquad j = 1, \dots, J_0, \quad 1 \leq k_j \leq J. \tag{62}$$

Next we approximate $u(z^{(k)})$ under the a priori assumption that u is smooth.

Denote by $U = [U_1, \dots, U_J]^T = [u(z^{(1)}), \dots, u(z^{(J)})]^T$ the unknown values and by $m = [m_1, \dots, m_J]^T$ the measured data. We use Tikhonov regularization [30] with second derivative penalty. That is, we solve the optimization problem

$$\hat{U} := \arg \min_U \{\|\mathcal{R}U - m\|_2^2 + \alpha \|\mathcal{D}U\|_2^2\}. \tag{63}$$

The first term in the penalty functional (63) describes how well U fits the data m. The matrix $\mathcal{R}$ implements (62): each row of $\mathcal{R}$ has all zeros except the entry 1 in the k_jth column. The second term in (63) expresses our a priori knowledge on u: we know that u is smooth, so we take the matrix $\mathcal{D} : \mathbb{R}^J \to \mathbb{R}^{J-2}$ to be the second-order difference matrix

$$\mathcal{D}(U)_k = \frac{1}{(\Delta\theta)^2}(U_{k+1} - 2U_k + U_{k-1}), \qquad k = 2, \dots, J-1.$$

The parameter $\alpha > 0$ is the regularization parameter: the greater α is, the stronger we require smoothness from the reconstruction. It is practical to write (63) in the stacked form as explained by Varah [33]:

$$\begin{bmatrix} \mathcal{R} \\ \sqrt{\alpha}\mathcal{D} \end{bmatrix} U = \begin{bmatrix} m \\ 0 \end{bmatrix}. \tag{64}$$

The regularized solution $\hat{U}$ is the least squares solution of (64).

Finally, we interpolate the values $u(x^{(k)})$ at the quadrature points in (60) with spline interpolation from the recovered values $u(z^{(k)})$. Under the smoothness assumption this does not produce significant error.

5.3. Choosing the triangle D. We need a systematic choice for $D = D(y)$. The analysis in section 3 suggests the following:

1. Better results are expected in the domain $y_2 \geq 1/2$ if one chooses $L = y_2$. This is because $L \geq 1 - y_2$ and the convergence rate of (37) is better than Hölder.
2. The constant $|C_D|$ should be as large as possible because the inverse of $|C_D|$ enlarges the error (32).
3. Note that $\operatorname{diam} D/L \geq 1$ and cannot be small.

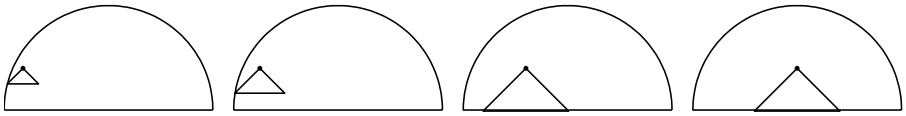


FIG. 5.1. *Triangular patches D corresponding to some y.*

We take vertices of the triangle enclosing D to be

$$(y_1, y_2), \qquad (y_1 - L, y_2 - L), \qquad (y_1 + L, y_2 - L), \tag{65}$$

with $L > 0$ taken as large as possible while still having $D \subset \Omega$. See Figure 5.1.

Now $p = -L$ and $q = L$, so

$$C_D = \frac{2L(q-p)}{(L-ip)(L-iq)} = \frac{4L^2}{L^2(1+i)(1-i)} = 2. \tag{66}$$

Note that in light of (27) this choice of C_D maximizes $|C_D|$.

5.4. Computing exponentially growing solutions. We want to evaluate the function v_τ given by (20) at the points $x^{(k)} \in \Gamma$ for $k = 1, \ldots, K$. We have $v_\tau|_\Gamma = v_\tau^0 w'_\tau|_\Gamma$ with w'_τ solving (18). In the case $V \equiv 0$, solving (18) amounts to computing a convolution. If $V \neq 0$, write (18) in the form

$$[I + g_\tau * (\tilde{V} \cdot\ )] w'_\tau = f, \tag{67}$$

where $f = g_\tau * \chi_D$. A numerical solution method for (67) is described in [22]. It is a modification of Vainikko's fast Lippmann–Schwinger solver [32, section 2]. This method is valid for potentials in the class of Definition 2.1.

Given an integer $m > 1$, the outcome of the solution algorithm is the set

$$\{w'_\tau(x^{(\ell)})\}_{\ell=1}^{M^2},$$

where the evaluation points $x^{(\ell)}$ belong to the Cartesian grid

$$\begin{aligned} &\mathcal{G}_m = \{jh \,|\, j \in \mathbb{Z}_m^2\}, \\ &\mathbb{Z}_m^2 = \{j = (j_1, j_2) \in \mathbb{Z}^2 \,|\, -2^{m-1} \le j_l < 2^{m-1},\ l = 1, 2\}, \end{aligned} \tag{68}$$

where $s > 1$ is a real number, $M = 2^m$, and $h = 2s/M$.

5.5. Computing derivatives of exponentially growing solutions. We need the values $\partial v_\tau / \partial \nu(x^{(k)})$ for $k = 1, \ldots, K$. We show that it is enough to evaluate w'_τ in addition to explicit formulae.

Take $\tau > 0$ and let w'_τ be the solution of $w'_\tau = g_\tau * \chi_D - g_\tau * (\tilde{V} w'_\tau)$. The derivatives $\partial_j w'_\tau$ for $j = 1, 2$ are given by

$$\partial_j w'_\tau = -(\partial_j g_\tau) * (\tilde{V} w'_\tau - \chi_D). \tag{69}$$

Using Lemma 4.1 we get

$$\begin{aligned}
\frac{\partial v_\tau}{\partial \nu} &= e^{-\tau y_2}\frac{\partial(e^{i\tau\bar{x}}w'_\tau)}{\partial \nu} = e^{-\tau y_2}\left[\nu_1\frac{\partial(e^{i\tau\bar{x}}w'_\tau)}{\partial x_1}+\nu_2\frac{\partial(e^{i\tau\bar{x}}w'_\tau)}{\partial x_2}\right]\\
&= e^{\tau(x_2-y_2)}e^{i\tau x_1}[i\tau\nu_1 w'_\tau+\tau\nu_2 w'_\tau-(\nu_1(\partial_1 g_\tau)+\nu_2(\partial_2 g_\tau))*(\tilde{V}w'_\tau-\chi_D)]\\
&= e^{\tau(x_2-y_2)}e^{i\tau x_1}\left[i\tau\nu_1 w'_\tau+\tau\nu_2 w'_\tau\right.\\
&\qquad -\nu_1\left(-\frac{1}{4\pi\bar{x}}-\frac{e_{-\bar{\tau}}(x)}{4\pi x}-i\tau g_\tau\right)*(\tilde{V}w'_\tau-\chi_D)\\
&\qquad \left.-\nu_2\left(-\frac{1}{4\pi i\bar{x}}+\frac{e_{-\bar{\tau}}(x)}{4\pi i x}-\tau g_\tau\right)*(\tilde{V}w'_\tau-\chi_D)\right] \qquad (70)\\
&= \frac{e^{\tau(x_2-y_2)}e^{i\tau x_1}}{4\pi}\left[\left(\nu_1\left(\frac{1}{\bar{x}}+\frac{e^{-i2\tau x_1}}{x}\right)+\nu_2\left(\frac{1}{i\bar{x}}-\frac{e^{-i2\tau x_1}}{ix}\right)\right)*(\tilde{V}w'_\tau-\chi_D)\right],
\end{aligned}$$

where we used formulae (53) and (54) and the real-valuedness of τ. Note the cancellation of four terms containing w'_τ resulting from the identity $w'_\tau = -g_\tau * (\tilde{V}w'_\tau - \chi_D)$.

5.6. Choosing τ. Theoretically, the larger $\tau > 0$ is, the closer $u_\tau(y)$ is to $u(y)$. However, too large τ leads to computations involving exponentially large numbers and numerical instability. This is even more so when the data is noisy. Thus τ must be chosen large enough for the approximation $u(y) \approx u_\tau(y)$ to be accurate enough but small enough to avoid instability. We discussed the optimal choice of τ theoretically in section 3, and in section 6 we study several choices of τ numerically.

6. Numerical results for $V \equiv 0$.

6.1. The model problem. Let $u = \mathrm{Re}((x_1 + ix_2)^4)$ be the harmonic function to be recovered from its Cauchy data on

$$\Gamma = \{x_1 + ix_2 = e^{i\theta} \,|\, 0 < \theta < \pi\} \subset \partial\Omega. \qquad (71)$$

See Figure 6.1 for a contour plot of u in the domain Ω together with plot of trace of u and plot of $\partial u/\partial\nu$.

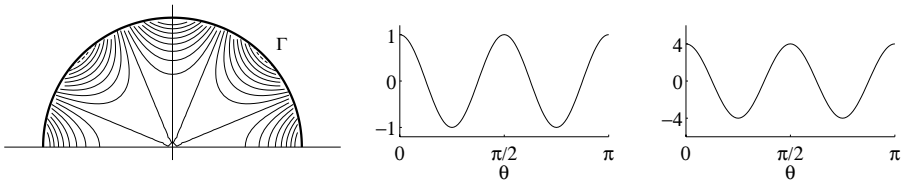


FIG. 6.1. *Left: contour plot of the harmonic function u in the domain Ω. Middle: plot of the trace $u|_\Gamma$ as function of angular parameter θ. Right: plot of normal derivative $\partial u/\partial\nu|_\Gamma$.*

6.2. Details of implementation.

Step 1*: Integration on* Γ. According to a given y, we divide Γ into three intervals:

$$0 < \theta < \tilde{\theta}, \qquad \tilde{\theta} < \theta < \pi - \tilde{\theta}, \qquad \pi - \tilde{\theta} < \theta < \pi.$$

Here $0 < \tilde{\theta} < \pi/2$ is chosen so that, roughly, the largest values of the integrand are in the interval containing $\pi/2$. We take $\tilde{\theta} = (70/360)\cdot 2\pi$. We choose K_0 Gaussian

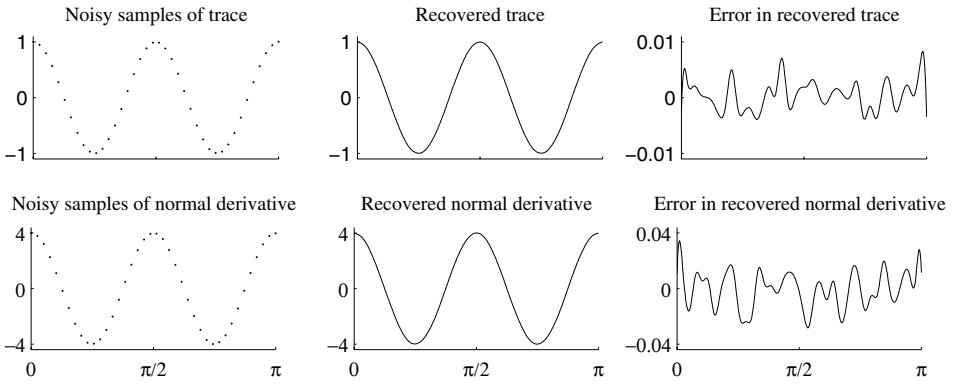


FIG. 6.2. *Simulation of noisy Cauchy data of u satisfying Laplace's equation. In each plot, the abscissa is the angular parameter θ for the curve $\Gamma = \{e^{i\theta} \,|\, 0 < \theta < \pi\}$.*

quadrature points for each interval leading to a quadrature rule of $K = 3K_0$ evaluation points for Γ.

Step 2*: Evaluation of data.* The Cauchy data of u are given by explicit formulae:

$$u|_\Gamma(\theta) = \cos 4\theta, \qquad \left.\frac{\partial u}{\partial \nu}\right|_\Gamma (\theta) = 4\cos 4\theta. \tag{72}$$

We produce simulated noisy data following the discussion in section 5.2. Set $\tilde{z}^{(j)} = (\cos\phi_j, \sin\phi_j)$ with $\phi_j = (j-1)\pi/(J_0 - 1)$ with $j = 1, \ldots, J_0 = 40$. We compute noisy samples as

$$u(\tilde{z}^{(j)}) + 0.003\,\varepsilon_j, \qquad \frac{\partial u}{\partial \nu}(\tilde{z}^{(j)}) + 0.012\,\varepsilon'_j, \tag{73}$$

where ε_j and ε'_j are normally distributed independent random numbers with standard deviation $\sigma = 1$. See Figure 6.2.

To recover the smooth data using Tikhonov regularization, we take

$$\theta_k = (k-1)\pi/(J-1), \qquad z^{(k)} = (\cos\theta_k, \sin\theta_k) \text{ for } k = 1, \ldots, J = 391.$$

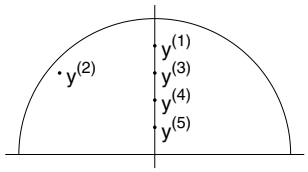
The result of solving (64) with regularization parameter $\alpha = 4$ is shown in Figure 6.2. The choice of α was based on visual inspection. The L^2 norm for the noise introduced in section 3 is $E \approx 0.02$.

Step 3*: Choosing the triangle D and computing C_D.* We implement the choice given in section 5.3. For any y, we start by $L = y_2$. Generally, this leads to $D \not\subset \Omega$. Then, we replace L with $L/2$ so many times that $D \subset \Omega$. Then we replace L with $L + 0.01$ as many times as possible while still having $D \subset \Omega$. We have $C_D = 2$.

Step 4*: Evaluation of v_τ.* With fixed y and given choice of $D = D(y)$ and $\tau = \tau(y)$, we substitute $V \equiv 0$ into (18):

$$w'_\tau(x) = \int_{\mathbb{R}^2} g_\tau(x-z)\chi_D(z)dz = \int_D g_\tau(x-z)dz.$$

So we need to integrate over D to find $w'_\tau(x)$ for a given $x \in \Gamma$. We use Gaussian product quadrature with $\tilde{K}_0^2 = \tilde{K}$ evaluation points. As indicated in section 4, we have available a numerical algorithm for g_τ, so Step 4 is complete.



$$y^{(1)} = (0, 4/5)$$
$$y^{(2)} = (-7/10, 3/5)$$
$$y^{(3)} = (0, 3/5)$$
$$y^{(4)} = (0, 2/5)$$
$$y^{(5)} = (0, 1/5)$$

FIG. 6.3. *Special points for examining the convergence of* $u_\tau(y^{(j)})$ *to* $u(y^{(j)})$.

Step 5*: Evaluation of* $\partial v_\tau/\partial\nu$. We have $v_\tau(x) = e^{-\tau y_2}e^{i\tau\bar{x}}w'_\tau(x)$, where $w'_\tau(x) = g_\tau * \chi_D$. Compute

$$\frac{\partial v_\tau}{\partial x_1} = e^{-\tau y_2}e^{i\tau\bar{x}}\left(i\tau w'_\tau(x) + \frac{\partial w'_\tau}{\partial x_1}\right). \tag{74}$$

Further, using Lemma 4.1,

$$\frac{\partial w'_\tau}{\partial x_1} = \frac{\partial g_\tau}{\partial x_1} * \chi_D = \left(-\frac{1}{4\pi\bar{x}} - \frac{e_{-\tau}(x)}{4\pi x}\right) * \chi_D - i\tau w'_\tau. \tag{75}$$

A combination of (74) and (75) yields

$$\frac{\partial v_\tau}{\partial x_1} = -e^{-\tau y_2}e^{i\tau\bar{x}}\left(\frac{1}{4\pi\bar{x}} + \frac{e_{-\tau}(x)}{4\pi x}\right) * \chi_D. \tag{76}$$

We can compute $\partial v_\tau/\partial x_2$ in a similar fashion. Note that on Γ the normal vector ν takes the simple form $\nu(x) = (x_1, x_2)$. Thus we get

$$\left.\frac{\partial v_\tau}{\partial\nu}\right|_\Gamma = -\frac{e^{-\tau y_2}e^{i\tau\bar{x}}}{4\pi}\Big(\frac{x_1}{\bar{x}} + \frac{x_1 e_{-\tau}(x)}{x} + \frac{x_2}{i\bar{x}} - \frac{x_2 e_{-\tau}(x)}{ix}\Big) * \chi_D. \tag{77}$$

Step 6*: Choosing* τ. We want to find a suitable τ experimentally. So we will compute u_τ with τ varying in the interval $[10, 80]$. We study convergence of u_τ to u at the points $y^{(1)}, \ldots, y^{(5)}$ given in Figure 6.3. We plot $u_\tau(y^{(j)})$ for $j = 1, \ldots, 5$ as functions of τ in Figure 6.4. Note that numerical instability occurs when τ is large. This is due to finite precision of the computation and the exponential functions appearing in the reconstruction formula.

6.3. Results for ideal data. We now have a complete numerical algorithm for u_τ. Since it is numerically impossible to compute u_τ for y_2 close to zero, we choose the computational reconstruction domain as

$$\Omega' = \left\{y \in \Omega \,|\, y_2 \geq \frac{1}{8}\right\}. \tag{78}$$

We compute $u_\tau(y)$ in Ω' for $\tau = 10, 40, 70$ on a collection of 1382 evaluation points inside the upper half of the unit disc. For integration on Γ we choose $K = 360$ quadrature points, and for integration on D we take a product rule with $\tilde{K} = 25^2 = 625$ points. We show the functions u, u_{20}, u_{40}, and u_{70} in Figure 6.5.

We see that the quality of the reconstruction varies depending on y_2 and τ. In particular, we observe that each plot with fixed τ has a region of acceptable reconstruction always containing a neighborhood of the point $(0, 1)$. Furthermore, when τ grows, the region of acceptable reconstruction shrinks, but the quality of

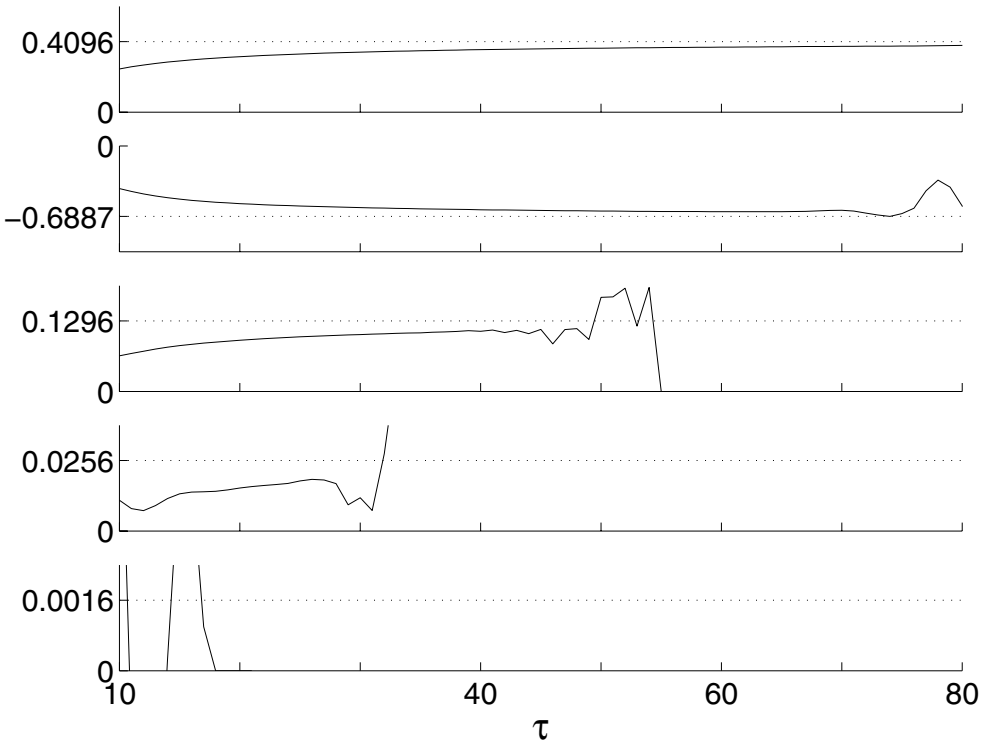


FIG. 6.4. *Convergence study for the Laplace equation. From top to bottom: plot of the real part of the function* $\tau \mapsto u_\tau(y^{(j)})$ *for* $j = 1, 2, 3, 4, 5$. *The theoretical limit values for these functions when* $\tau \to \infty$ *are marked in the plots. Note that numerical instability makes the computation inaccurate when* τ *grows. Divergence appears with smaller* τ *values for those reconstruction points that are deeper inside* Ω.

TABLE 6.1

Relative errors in u_τ *computed from ideal Cauchy data of a solution* u *to Laplace's equation. Left: relative* L^2 *errors* $E_2(t,\tau)$. *Right: relative* L^∞ *errors* $E_\infty(t,\tau)$. *In these tables, "—" stands for "greater than* 1000%." *For definitions of* E_2 *and* E_∞, *see* (79).

	$t = 1/8$	1/2	7/8
$\tau = 10$	53%	47%	45%
40	923%	16%	14%
70	—	—	9%

	$t = 1/8$	1/2	7/8
$\tau = 10$	91%	91%	33%
40	—	46%	10%
70	—	—	6%

reconstruction in the acceptable region is better than with smaller τ. For quantitative examination of this property we introduce the following norms for measuring the error of reconstructions. Given $0 < t < 1$ and $\tau > 0$, we consider the relative errors

$$E_2(t,\tau) = \frac{\|u - u_\tau\|_{L^2(\Omega_t)}}{\|u\|_{L^2(\Omega_t)}}, \qquad E_\infty(t,\tau) = \frac{\|u - u_\tau\|_{L^\infty(\Omega_t)}}{\|u\|_{L^\infty(\Omega_t)}}, \tag{79}$$

where $\Omega_t = \{y \in \Omega \,|\, y_2 > t\}$. The errors are given in Table 6.1.

For one choice of τ, the computation takes about 4 hours with MATLAB 6.5 running on a desktop PC computer with an Intel Pentium IV 2.8 GHz processor and 1 GB memory. In practical applications the collection of recovery points y and a good

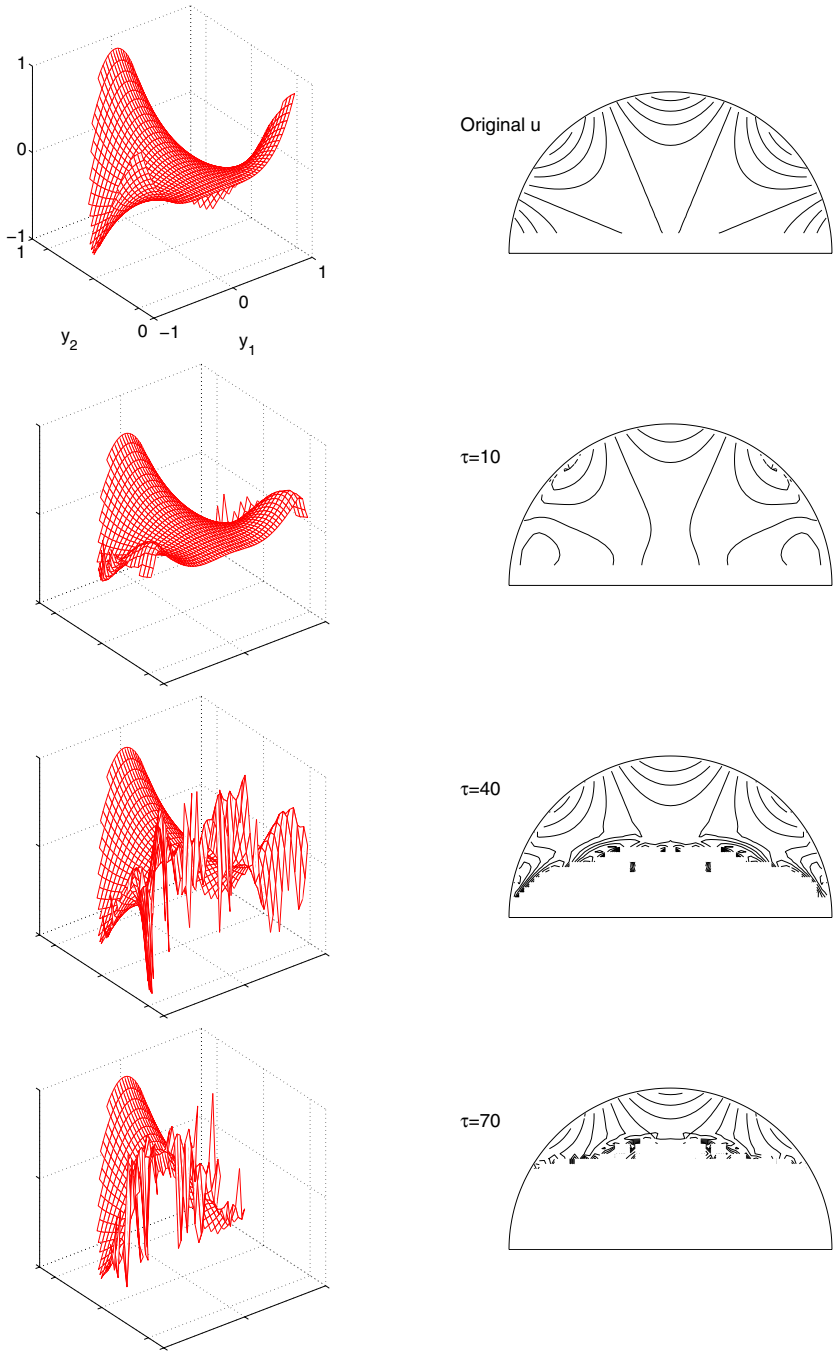


FIG. 6.5. *Results for the Laplace equation and ideal data. Top left: mesh plot of the original harmonic function u. Top right: contour plot of u. Similarly, from top to bottom, we show mesh and contour plots of the reconstructions u_{10}, u_{40}, and u_{70}. The axis limits are the same in all mesh plots, allowing easy comparison. We do not plot any function values greater than* 1 *in absolute value since numerical instability causes extremely large (incorrect) values in the reconstructions, and visualizing these values would obscure the acceptable parts of the reconstructions. Note that the greater τ is, the better the reconstruction is for points (y_1, y_2) with y_2 near* 1.

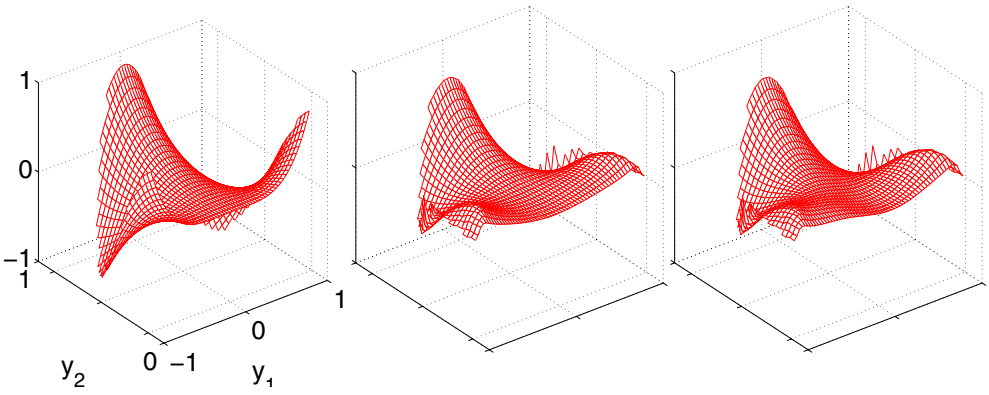


FIG. 6.6. *Results for the Laplace equation and noisy data. Left: mesh plot of the original harmonic function u. Middle: recovery of u with spatially varying τ from data with noise level* 0.3%. *Right: recovery of u with spatially varying τ from data with noise level* 0.6%.

choice for τ can be fixed. Then traces $v_\tau|_{\partial\Omega}$ and the normal derivatives $\partial v_\tau/\partial\nu|_{\partial\Omega}$ according to each y need be computed only once and saved. Then recovery of $u(y)$ with given Cauchy data takes in this case only a couple of seconds.

It can be seen in Figure 6.5 that near those parts of the boundary that are almost parallel to the y_2 axis, the quality of the reconstruction is bad. This is related to the smaller triangular patch D used there, leading to slower convergence.

6.4. Results for noisy data. We compute the functions u_7, u_{10}, and u_{12} using noisy Cauchy data. The results are similar to the nonnoisy case, the main difference being that the region of acceptable results shrinks with considerably smaller τ values. To achieve a uniform level of regularization, we choose τ as a function of y as follows:

$$\tau = \tau(y) = 6 + 6y_2^3,$$

so deep inside Ω we use a smaller value of τ, leading to less oscillation. Since the τ values used were relatively small, we did not need so many quadrature points for numerical integration. For integration on Γ we choose $K = 36$ quadrature points, and for integration on D we take a product rule with $\tilde{K} = 7^2 = 49$ points.

For the result, see the middle plot in Figure 6.6. The recovered solution has 38% relative $L^2(\Omega')$ error. To examine the robustness of our method against noise, we produce noisy data with double standard deviation in the random errors in (73). We repeat the recovery process, leading to a result having 41% relative $L^2(\Omega')$ error; see the rightmost plot in Figure 6.6. The relative difference between the two reconstructions in $L^2(\Omega')$ is only 6% although the noise level was doubled. The computation of each of the two reconstructions took 5 minutes.

7. Numerical results for $V \neq 0$. We present a two-dimensional example roughly modelling the inverse potential problem in cardiology.

7.1. The Cauchy problem in cardiology. A beating heart produces an electrical field inside the body, and the resulting voltage distribution can be measured with electrodes placed on the skin. This is called electrocardiography (ECG). The

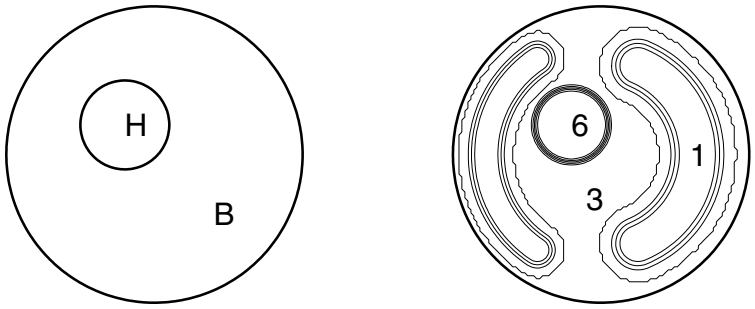


FIG. 7.1. *Left: the domain of the conductivity equation is the annulus* $\tilde{\Omega} = B \setminus H$. *Right: contour plot of the twice differentiable conductivity distribution on* B.

inverse potential problem of cardiology is of clinical interest: *given the ECG measurements and the conductivity distribution of the body, what is the voltage potential on the surface of the heart?*

We present a two-dimensional quasi-static model of the above inverse problem. Let the unit disc $B = \{x \in \mathbb{R}^2 \,|\, |x| < 1\}$ model a cross section of human thorax, and assume that the heart is located on the disc H with center at $(-0.2, 0.2) \in B$ and radius 0.3. Further, we model the electrical conductivity $\gamma : B \to \mathbb{R}$ of the body with a strictly positive $C^2(\overline{B})$ function taking value 6 in the heart, 1 in the lungs, and 3 in the background. These values approximate the tissue conductivities during perfusion. See Figure 7.1.

Electric current inside the heart results in the following boundary value problem for the electric voltage potential $\tilde{u}$ in the annulus $\tilde{\Omega} = B \setminus H$:

$$\nabla \cdot \gamma \nabla \tilde{u} = 0 \text{ in } \tilde{\Omega}, \quad \tilde{u}|_{\partial H} = f, \quad \left.\frac{\partial \tilde{u}}{\partial \nu}\right|_{\partial B} = 0, \tag{80}$$

where we assumed that the outer boundary ∂B is perfectly insulated.

We create our example by setting $f(x_1, x_2) = (x_1 + 0.5)(x_2 - 0.2)$ in (80), qualitatively resembling a voltage distribution depicted on page 386 in [7]. We solve the elliptic boundary value problem (80) with the finite element solver of MATLAB's PDE toolbox using 14848 triangles in the domain $\tilde{\Omega}$. See Figure 7.2.

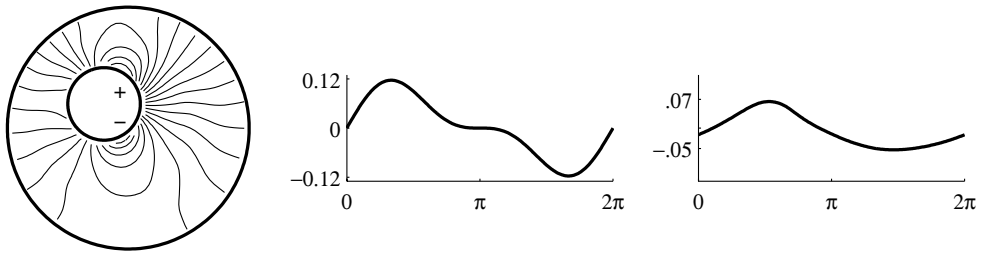


FIG. 7.2. *Left: contour plot of the solution* $\tilde{u}$ *of the conductivity equation* (80). *Middle: voltage potential* $\tilde{u}|_{\partial H}$ *on the surface of the heart as a function of the angular variable* θ *corresponding to the parametrization* $\partial H = \{(-0.2 + 0.3\cos\theta, 0.2 + 0.3\sin\theta) \in \mathbb{R}^2 \,|\, 0 \le \theta < 2\pi\}$. *Right: voltage potential* $\tilde{u}|_{\partial B}$ *as a function of the angular variable* θ *corresponding to the parametrization* $\partial B = \{(\cos\theta, \sin\theta) \in \mathbb{R}^2 \,|\, 0 \le \theta < 2\pi\}$. *Axis limits in the two plots are the same to allow quantitative comparison.*

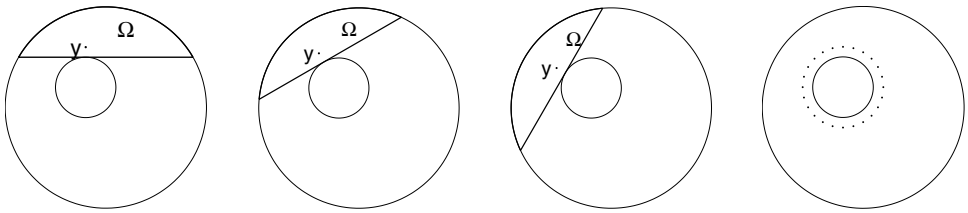


FIG. 7.3. *In the rightmost picture are* 24 *reconstruction points on the circle* S *defined in* (83) *at distance* 0.1 *from the surface of the heart. For each reconstruction point we choose a domain* Ω *that coincides after rotation with a canonical domain described in section* 2.

Application of the techniques of this paper requires transforming the conductivity equation (80) into the Schrödinger equation. Set

$$V(x) = \frac{\Delta\sqrt{\gamma(x)}}{\sqrt{\gamma(x)}}. \tag{81}$$

Since $\gamma \in C^2(\overline{B})$, we have $V \in C^0(\tilde{\Omega}) \subset L^\infty(\tilde{\Omega})$. (The norm of the particular potential we use is approximately $\|V\|_{L^\infty(\tilde{\Omega})} \approx 293$, so our example is not a small perturbation of the harmonic case.) It is straightforward to check that $u := \gamma^{1/2}\tilde{u}$ satisfies the equation

$$(-\Delta + V)u = 0 \qquad \text{in } \tilde{\Omega}. \tag{82}$$

Because $\gamma \equiv 3$ in a neighborhood of ∂B, we know the Cauchy data of u on ∂B:

$$u|_{\partial B} = \sqrt{3}\tilde{u}|_{\partial B}, \qquad \left.\frac{\partial u}{\partial \nu}\right|_{\partial B} = 0.$$

Equation (82) is valid only outside the heart. We choose a collection of computational domains Ω as shown in Figure 7.3; each of these domains coincides after rotation with a canonical domain described in section 2. We cannot choose our recovery points right at the surface of the heart because the set D would then be empty, so we choose 24 points on the circle S given by

$$S = \{(-0.2 + 0.4\cos\theta, 0.2 + 0.4\sin\theta) \in \mathbb{R}^2 \,|\, 0 \le \theta < 2\pi\}. \tag{83}$$

Thus we reconstruct the voltage at distance 0.1 from the surface of the heart. See Figure 7.3.

7.2. Details of implementation. We assume that the domain Ω is (possibly after rotation) of the canonical form with $-1 < t < 1$ described in section 2. The Neumann data of u vanish, we need only compute

$$u_\tau(y) = -\frac{2\tau^2 e^{-i\tau y_1}}{C_D} \int_\Gamma \frac{\partial v_\tau}{\partial \nu} u \, d\sigma(x). \tag{84}$$

Step 1*: Integration on* Γ. We choose K Gaussian quadrature points on Γ. There is no need to divide Γ into subintervals as done in section 6.2 since we use so small a value of τ that the integrand is roughly of the same order of magnitude throughout Γ.

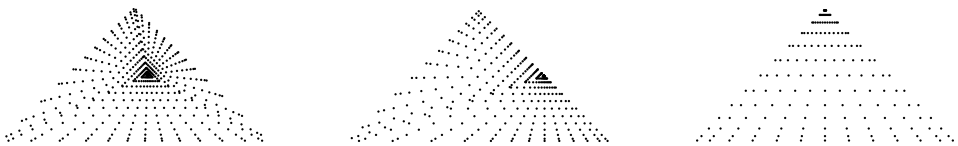


FIG. 7.4. *Product Gaussian quadratures on the triangular patch D with 13×13 points in polar coordinates. Left: the origin is inside D. Middle: the origin is on the boundary of D. Right: the origin is at the corner of D.*

Step 2*: Evaluation of data.* The experimental setup in [7] uses 24 electrodes with 2% noise level. We simulate that measurement with

$$\tilde{u}(\tilde{z}^{(j)}) + 0.0022\,\varepsilon_j,$$

where $\tilde{z}^{(j)} = (\cos\phi_j, \sin\phi_j)$ with $\phi_j = (j-1)2\pi/J_0$ with $j = 1, \ldots, J_0 = 24$ and ε_j are normally distributed independent random numbers with standard deviation 1.

We use Tikhonov regularization to recover a smooth approximation to the actual voltage. In the notation of section 5.2, we have $J_0 = 24$, $J = 144$, and $\alpha = 2$. Since we reconstruct the trace on the full circle, we include requirement of periodicity into the regularization. Relative $L^2(\partial\Omega)$ error in the reconstruction of the trace $\tilde{u}|_{\partial\Omega}$ is 0.033, and relative $L^\infty(\partial\Omega)$ error is 0.032.

Step 3*: Choosing the triangle D and computing C_D.* We take $L = 0.1$ and use the choice given in section 5.3 leading to $C_D = 2$.

Step 4*: Evaluation of v_τ.* We need to solve the Lippmann–Schwinger-type equation $[I + g_\tau * (\tilde{V}\cdot\,)]w'_\tau = f$ as explained in section 5.4. The problem is the evaluation of

$$f(x^{(\ell)}) = (g_\tau * \chi_D)(x^{(\ell)}) = \int_D g_\tau(x^{(\ell)} - y)dy, \quad \ell = 1, 2, \ldots, M^2. \tag{85}$$

Since $g_\tau(x)$ has a logarithmic singularity at $x = 0$, numerical integration in (85) becomes problematic when $x^{(\ell)}$ belongs to D or is close to the boundary ∂D. We overcome this problem by writing the integral in polar coordinates and using product Gaussian quadrature; due to the product measure $rdrd\phi$ the integrand is bounded and continuous since $\lim_{r\to 0} r\log r = 0$. We need only to go through the tedious task of writing the integration domain as a function of ϕ and $r(\phi)$. We do not bore the reader with the details of dividing the algorithm into 19 subcases and performing the necessary trigonometric calculations but instead show some resulting quadrature points in Figure 7.4.

Step 5*: Evaluation of $\partial v_\tau/\partial\nu$.* From (70) we see that the normal derivative of v_τ appearing in (84) is given by

$$\begin{aligned}
&-\frac{e^{\tau(x_2-y_2)}e^{i\tau x_1}}{4\pi}\left[\left(\nu_1\left(\frac{1}{\bar{x}} + \frac{e^{-i2\tau x_1}}{x}\right) + \nu_2\left(\frac{1}{i\bar{x}} - \frac{e^{-i2\tau x_1}}{ix}\right)\right) * \chi_D\right] \\
&+\frac{e^{\tau(x_2-y_2)}e^{i\tau x_1}}{4\pi}\left[\left(\nu_1\left(\frac{1}{\bar{x}} + \frac{e^{-i2\tau x_1}}{x}\right) + \nu_2\left(\frac{1}{i\bar{x}} - \frac{e^{-i2\tau x_1}}{ix}\right)\right) * \tilde{V}w'_\tau\right] \\
&\equiv I_1 + I_2.
\end{aligned}$$

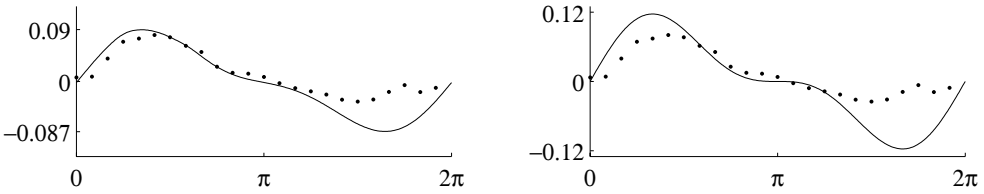


FIG. 7.5. *Left: true voltage potential at distance* 0.1 *from the boundary of the heart is plotted as a solid line. Reconstructed voltage potential values are plotted as dots. See Figure* 7.3 *for the reconstruction points. Right: true voltage potential at the boundary of the heart (solid line) and the same reconstructed voltage potentials as in the left plot (dots). The abscissa in both plots is the angular parameter of the circle around the heart.*

Note that I_2 vanishes when $V = 0$, and therefore I_2 can be seen as a correction term compensating for nonzero V. The computation of I_1 was already described for the case $V = 0$.

Given an integer $m > 1$, the outcome of Step 4 is the set $\{w'_\tau(x^{(\ell)})\}_{\ell=1}^{M^2}$, where the evaluation points $x^{(\ell)}$ belong to the grid (68). It is then natural to implement the integral in I_2 simply with the midpoint rule.

Step 6*: Choosing* τ*.* We take $\tau = 4, 6, 8, 10, 12$.

7.3. Results. To compute (84) we choose $K = 70$ for integrating over Γ, and for all integrations over D we take $\tilde{K} = 15^2 = 225$ quadrature points. We take $m = 7$, or $M = 128$ in the Lippmann–Schwinger solver. We compute u_τ with $\tau = 4, 6, 8, 10, 12$ and find that the reconstructions with $\tau > 6$ are oscillatory, and $\tau = 6$ gives a better result than $\tau = 4$. We thus choose $\tau = 6$.

The plot on the left in Figure 7.5 shows the superposition of reconstructed voltage potential $\gamma^{-1/2}u_6$ and the actual potential *on the circle* S *containing the reconstruction points*. We find that the maximum relative absolute error of the reconstruction is 86%. Diagnostically, the most interesting part of the reconstruction is the angular interval $0 \le \theta \le \pi$. In this interval, the maximum and average relative absolute errors are 25% and 10%, respectively.

However, we are interested in the voltage potential at the boundary of the heart. We simply consider our reconstruction of the voltage on S to be an approximation to the voltage on ∂H. The plot on the right in Figure 7.5 shows a comparison of these two quantities. Maximum relative absolute error in the reconstruction as compared to the voltage potential at the boundary of the heart is 1.07. In the interval $0 \le \theta \le \pi$, the maximum and average relative absolute errors are 43% and 21%, respectively.

The computation took 4 hours.

7.4. Discussion. Unlike in many works on the inverse potential problem of ECG, such as [7], we do not assume that the tissue between the skin and the surface of the heart is homogeneous. If the electric conductivity of the body is known, e.g., by electrical impedance tomography [5], our method thus allows more accurate modelling of the problem.

The worst-case performance of our algorithm is not impressive: the maximum relative error is 1.07. However, this worst error appears near the posterior surface of the heart (facing the back), which is far away from the boundary. The anterior surface of the heart (facing the chest) is diagnostically more important. Relative error on the

anterior surface, defined as $0 \leq \theta \leq \pi$ in the notation of Figure 7.2, is on average 21% and at most 43%. This result is somewhat better than the 30%–50% error reported in [7], where conductivity was taken to be constant.

Instead of quantitative reconstruction of the voltage, we might want to know the location of the maximum voltage potential on the anterior surface of the heart. The true maximum appears at $\theta_0 = 1.05$ (in radians), and the reconstruction attains its maximum at $\tilde{\theta}_0 = 1.31$. The error in the reconstructed angle is 15 degrees.

The main advantage of our method is modelling the conductivity, and the main source of error is the inherent problem that we cannot recover the voltage at the surface of the heart but slightly away from it. As mentioned in the introduction, there are other methods capable of dealing with nonconstant conductivities and additionally providing reconstruction right at the surface of the heart. However, those methods typically involve solution of boundary value problems, which is computationally intensive. Our reconstruction method is very fast after the initial computational load, and it could thus be better suited for real-time monitoring. Also, modelling the movement of a beating heart for the solution of boundary value problems is difficult, and our approach of reconstructing a little bit away from the heart might be considered an advantage.

Our tissue model assumes that the conductivity is differentiable although in reality it is discontinuous, but since many regularized electrical impedance tomography reconstructions produce a differentiable approximation to the conductivity, this is perhaps not so serious. The most obvious drawback of the presented algorithm is the two-dimensional approximation. However, the theory behind our method covers the three-dimensional case, and a similar algorithm can be designed in three dimensions. This is left for a future study.

REFERENCES

[1] L. Aizenberg, *Carleman's Formulas in Complex Analysis*, Kluwer Academic Publishers, London, 1993.

[2] F. Berntsson and L. Eldén, *Numerical solution of a Cauchy problem for the Laplace equation*, Inverse Problems, 17 (2001), pp 839–853.

[3] M. Boiti, J. P. Leon, M. Manna, and F. Pempinelli, *On a spectral transform of a KdV-like equation related to the Schrödinger operator in the plane*, Inverse Problems, 3 (1987), pp. 25–36.

[4] T. Carleman, *Les Fonctions Quasi Analytiques*, Gauthier-Villars, Paris, 1926.

[5] M. Cheney, D. Isaacson, and J. C. Newell, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[6] J. Cheng, Y. C. Hon, T. Wei, and M. Yamamoto, *Numerical computation of a Cauchy problem for Laplace equation*, ZAMM Z. Angew. Math. Mech., 81 (2001), pp. 665–674.

[7] P. Colli-Franzone, L. Guerri, S. Tentoni, C. Viganotti, S. Baruffi, S. Spaggiari, and B. Taccardi, *A mathematical procedure for solving the inverse potential problem of electrocardiography. Analysis of the time-space accuracy from in vitro experimental data*, Math. Biosci., 77 (1985), pp. 353–396.

[8] L. D. Faddeev, *Increasing solutions of the Schrödinger equation*, Sov. Phys. Dokl., 10 (1966), pp. 1033–1035.

[9] V. A. Fok and F. M. Kuni, *On the introduction of a "suppressing" function in dispersion relation*, Dokl. Akad. Nauk SSSR, 127 (1959), pp. 1195–1198 (in Russian).

[10] M. G. Goluzin and I. V. Krylov, *A generalized Carleman formula and its application to analytic continuation of functions*, Mat. Sb., 4 (1933), pp. 144–149.

[11] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[12] D. N. Háo and D. Lesnic, *The Cauchy problem for Laplace's equation via the conjugate gradient method*, IMA J. Appl. Math., 65 (2000), pp. 199–217.

[13] M. Ikehata, *Exponentially growing solutions and the Cauchy problem*, Appl. Anal., 78 (2001), pp. 79–95.

[14] M. Ikehata, *The enclosure method and its applications*, in Analytic Extension Formulas and Their Applications, S. Saitoh, N. Hayashi, and M. Yamamoto, eds., Int. Soc. Anal. Appl. Comput. 9, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 87–103.
[15] V. Isakov, *On uniqueness and stability of the Cauchy problem*, Contemp. Math., 209 (1997), pp. 131–146.
[16] S. I. Kabanikhin and A. L. Karchevsky, *Optimizational method for solving the Cauchy problem for an elliptic equation*, J. Inverse Ill-Posed Probl., 3 (1995), pp. 21–46.
[17] M. V. Klibanov and F. Santosa, *A computational quasi-reversibility method for Cauchy problems for Laplace's equation*, SIAM J. Appl. Math., 51 (1991), pp. 1653–1675.
[18] R. Lattés and J.-L. Lions, *The Method of Quasi-Reversibility: Applications to Partial Differential Equations*, translated from the French and edited by R. Bellman, Elsevier, New York, 1969.
[19] M. M. Lavrent'ev, *On the Cauchy problem for the Laplace equation*, Izv. Akad. Nauk. SSSR Ser. Mat., 20 (1956), pp. 819–842.
[20] M. M. Lavrent'ev, V. G. Romanov, and S. P. Shishatskii, *Ill-Posed Problems of Mathematical Physics and Analysis*, Transl. Math. Monogr. 64, AMS, Providence, RI, 1986.
[21] A. Leitão, *An iterative method for solving elliptic Cauchy problems*, Numer. Funct. Anal. Optim., 21 (2000), pp. 715–742.
[22] J. L. Mueller and S. Siltanen, *Direct reconstructions of conductivities from boundary measurements*, SIAM J. Sci. Comput., 24 (2003), pp. 1232–1266.
[23] A. I. Nachman, *Reconstructions from boundary measurements*, Ann. Math., 128 (1988), pp. 531–576.
[24] J. Saranen and G. Vainikko, *Periodic Integral and Pseudodifferential Equations with Numerical Approximation*, Springer, Berlin, 2002.
[25] S. Siltanen, *Electrical Impedance Tomography and Faddeev's Green Function*, Ph.D. thesis, Helsinki University of Technology, Helsinki, Finland, 1999.
[26] S. Siltanen, J. Mueller, and D. Isaacson, *An implementation of the reconstruction algorithm of A. Nachman for the 2-D inverse conductivity problem*, Inverse Problems, 16 (2000), pp. 681–699.
[27] Z. Sun and G. Uhlmann, *Recovery of singularities for formally determined inverse problems*, Comm. Math. Phys., 153 (1993), pp. 431–445.
[28] J. Sylvester and G. Uhlmann, *Global uniqueness theorem for an inverse boundary value problem*, Ann. Math., 125 (1987), pp. 153–169.
[29] N. N. Tarkhanov, *The Cauchy Problem for Solutions of Elliptic Equations*, Math. Topics 17, Akademie Verlag, Berlin, 1995.
[30] A. N. Tikhonov, *Solution of incorrectly formulated problems and the regularization method*, Soviet Math. Dokl., 4 (1963), pp. 1035–1038.
[31] G. Uhlmann, *Developments in inverse problems since Calderón's foundational paper*, in Harmonic Analysis and Partial Differential Equations, M. Christ, C. E. Kenig, and C. Sadosky, eds., University of Chicago Press, Chicago, London, 1999, pp. 295–345.
[32] G. Vainikko, *Fast solvers of the Lippmann-Schwinger equation*, in Direct and Inverse Problems of Mathematical Physics (Newark, DE), Int. Soc. Anal. Appl. Comput. 5, R. P. Gilbert, J. Kajiwara, and Y. S. Xu, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 423–440.
[33] J. M. Varah, *A practical examination of some numerical methods for linear discrete ill-posed problems*, SIAM Rev., 21 (1979), pp. 100–111.
[34] Sh. Yarmukhamedov, *On the Cauchy problem for the Laplace equation*, Dokl. Akad. Nauk SSSR, 235 (1977), pp. 281–283 (in Russian).
[35] Sh. Yarmukhamedov, *Harmonic extension of continuous functions defined on a piece of the boundary*, Russian Acad. Sci. Dokl. Math., 46 (1993), pp. 430–434.
[36] Sh. Yarmukhamedov, *Integral representation of a CR-function and its holomorphic continuation*, Dokl. Math., 51 (1995), pp. 253–255.
[37] Sh. Yarmukhamedov, *Continuing solutions to the Helmholtz equation*, Dokl. Math., 56 (1997), pp. 887–890.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 1933–1953

SEMICLASSICAL APPROXIMATION OF ELECTRON-PHONON SCATTERING BEYOND FERMI'S GOLDEN RULE*

C. RINGHOFER†, M. NEDJALKOV‡, H. KOSINA§, AND S. SELBERHERR§

Abstract. We derive a quantum mechanical correction to the semiclassical Fermi golden rule operator for scattering of electrons in a crystal. This correction takes into account the fact that electron-phonon interaction is not instantaneous in quantum mechanics. The corrective term is derived via an oscillatory, i.e., weak, limit in the Levinson equation for large timescales.

Key words. asymptotic analysis, quantum mechanics, Levinson equation, Wigner functions, Fermi's golden rule, Boltzmann equation

AMS subject classifications. 65N35, 65N05

DOI. 10.1137/S0036139903428914

1. Introduction.

It is generally accepted that the dominant collision mechanism for electron transport in crystals is scattering of electrons by phonons, i.e., with vibrations of the crystal lattice. In a semiclassical description this collision mechanism is described by the Fermi golden rule. In the absence of an electric field and in the spatially homogeneous case, the evolution of the effective single electron density function is then given by the Boltzmann equation

$$\begin{aligned} &\text{(a)} \quad \partial_t f(p,t) = Q_{FGR}[f](p,t) \\ &\qquad\qquad := \int dp' [S_{FGR}(p,p') f(p',t) - S_{FGR}(p',p) f(p,t)], \\ &\text{(b)} \quad S_{FGR}(p,p') = [A^- \delta(\varepsilon(p) - \varepsilon(p') - \hbar\omega) + A^+ \delta(\varepsilon(p) - \varepsilon(p') + \hbar\omega)], \end{aligned} \tag{1.1}$$

where p denotes the momentum vector and $\varepsilon(p) = \frac{|p|^2}{2m_*}$ denotes the energy associated with the momentum p. The Fermi golden rule states that during a collision the electron gains or loses an amount $\hbar\omega$ of energy from the crystal lattice by annihilation or generation of a phonon. We remark that the Boltzmann equation (1.1) models instantaneous collisions; i.e., the momentum of an electron changes instantaneously from p' to p during a collision event.

Semiclassical transport theory based on the Boltzmann equation neglects several effects which originate from the quantum mechanical nature of the charge carriers, such as a collisional broadening due to the finite lifetime of the carrier momentum eigenstate, collision retardation, and the intracollisional field effect due to the action of the electric field during the scattering process [11]. To describe these effects a quantum kinetic equation has to be adopted which takes the finiteness of the scattering duration into account. An appropriate kinetic equation describing the interaction of a single electron with the equilibrium phonon system of a semiconductor has been proposed

*Received by the editors June 4, 2003; accepted for publication (in revised form) October 24, 2003; published electronically August 4, 2004. This work was supported by NSF grant DECS-0218008.
http://www.siam.org/journals/siap/64-6/42891.html

†Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804 (ringhofer@asu.edu).

‡Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5506 (mixi@iue.tuwien.ac.at).

§Institute for Microelectronics, Technical University of Vienna, Gusshausstrasse 27-29, A1040 Vienna, Austria (kosina@iue.tuwien.ac.at).

by Levinson [6]. Restricting ourselves to the case of a spatially uniform semiconductor and vanishing electric field, the Levinson equation is of the form

$$
\begin{aligned}
&\text{(a)} \quad \partial_t f(p,t) = Q[f](p,t) \\
&\qquad\qquad := \int_0^t dt' \int dp' [S(p,p',t-t')f(p',t') - S(p',p,t-t')f(p,t')], \\
&\text{(b)} \quad S(p,p',t) = \frac{2VF^2 n}{(2\pi\hbar)^3} \cos\left[\frac{t}{\hbar}(\varepsilon(p) - \varepsilon(p') - \hbar\omega)\right] \\
&\qquad\qquad + \frac{2VF^2(n+1)}{(2\pi\hbar)^3} \cos\left[\frac{t}{\hbar}(\varepsilon(p) - \varepsilon(p') + \hbar\omega)\right], \\
&\qquad \varepsilon(p) = \frac{|p|^2}{2m_*}, \quad n = \frac{1}{\exp(\beta\hbar\omega) - 1}.
\end{aligned} \tag{1.2}
$$

The symbols in (1.2) have the following meaning: $\hbar F$ denotes the electron-phonon interaction matrix element, $\hbar\omega$ the phonon energy, V the normalization volume, m_* the effective electron mass, n the Bose–Einstein distribution, and $\beta = (k_B T)^{-1}$ the inverse temperature of the phonon system. Note that, other than the Fermi golden rule equation (1.1), the Levinson equation (1.2) is nonlocal in time, and the effect of a collision is actually felt for all future times. Therefore the Levinson equation is able to model some of the effects mentioned above, which become increasingly relevant as the dimensions of modern semiconductor devices decrease, and, consequently, fast relaxation processes play a more prominent role. The Levinson equation can be derived from the quantum mechanical many body problem for one electron and an arbitrary number of phonons, i.e., from an infinite system of Schrödinger equations for the wave functions $\psi_n(p, q_1, \ldots, q_n, t)$, where p is the electron momentum vector and the q_j denote the phonon momenta. ψ_n describes the state of the system for one electron and n phonons, and ψ_n is coupled to ψ_{n-1} and ψ_{n+1} via coupling terms in the Fröhlich Hamiltonian, modelling the creation and annihilation of phonons. The function f is then the Wigner function corresponding to the phonon trace of the density matrix. We refer the reader to [1], [3], [4], [5] for an overview of the derivation. The Levinson equation represents the weak coupling limit of this system, which means that only electron-phonon interactions of first order are retained. An interaction starts at some time, say t_1, when one half of the phonon momentum is transferred to the electron, and gets completed at some time t_2, when the second half of the phonon momentum is transferred (see, e.g., [4], [9]). These partial processes capture the emission and absorption of both real and virtual phonons. The weak coupling limit implies that during the period $t_2 - t_1$ of a particular interaction no other interaction can start. In other words, only a sequence of completed interactions is considered. The time between two interactions is determined by the frequency F given by the interaction matrix element, whereas the duration of the interaction depends on the frequency of the lattice vibrations, ω. Therefore, $F \ll \omega$ must hold. A result of the weak coupling limit is that no powers higher than F^2 appear in (1.2)(b). To our knowledge, a completely rigorous mathematical derivation of the Levinson equation from the many body problem for the Fröhlich Hamiltonian is still outstanding. However, for the purpose of this paper, we will assume the Levinson equation (1.2) to be valid.

Remark: The Levinson equation results from an asymptotic expansion of the Fröhlich Hamiltonian for small coupling coefficients [1], [3]. While there obviously is a density matrix formulation of the Levinson equation, which is given by the Fourier

transform of (1.2), the corresponding density matrix cannot simply be written as a superposition of pure state wave functions, and therefore the positive definiteness of the Wigner function f in (1.2) cannot be guaranteed automatically.

The objective of the present paper is to relate (1.2) to the Fermi golden rule (1.1). While much simpler than the original many body equation, the Levinson equation still poses significant challenges because of its nonlocality in time and rapid oscillations due to the presence of the term $t/\hbar$ in the integral kernel in (1.2). It is mentioned in the original work [6] that in the classical limit $\hbar \to 0$ the scattering rate S in (1.2)(b) will be replaced by the Fermi golden rule (cf. [2]).

In the present paper we prove the convergence of the Levinson operator Q in (1.2) to the Fermi golden rule operator Q_{FGR} for large timescales, and more importantly, in addition to the golden rule the first order term in the asymptotic expansion is derived. The result is a correction to the Fermi golden rule, which better reflects the effects of finite collision times. The resulting corrected operator is structurally of the form

$$Q_C[f] = \int dp' [S_C(p, p', \partial_t) f(p', t) - S_C(p', p, \partial_t) f(p, t)], \tag{1.3}$$

where the corrected scattering rate S_C contains the Fermi golden rule rate S_{FGR} and a correction term that involves the time derivative of the density function. However, the corrected operator Q_C in (1.3) is still local in time, in the sense that it is not an integral operator in time, and therefore the resulting transport equation is much simpler to solve than the Levinson equation (1.2). More precisely, we prove that the Levinson operator Q converges weakly, in zeroth order to Q_{FGR}, and in first order to the corrected operator Q_C. The considered regime is one of large timescales, i.e., timescales which are much larger than $1/\omega$, where ω is the frequency with which the lattice vibrates.

This paper is organized as follows. In section 2 we introduce an appropriate dimensionless form of the Levinson equation (1.2) which contains a dimensionless parameter $\lambda = (\omega t_0)^{-1}$, where t_0 is the timescale under consideration. Section 3 contains the asymptotic analysis for $\lambda \to 0$. We prove that $Q = Q_C + o(\lambda)$, $Q_C = Q_{FGR} + \lambda Q_1$ holds in a weak sense, i.e., when integrated against a fixed test function. The main result of the paper, the form of Q_C, is given at the end of section 3 in formula (3.10). Section 4 is devoted to numerical experiments. First the asymptotic result of section 3 is verified. This result states only the weak convergence of Q to Q_C and not the convergence of the solution f of the Levinson equation to the solution of the corresponding transport equation containing Q_C. The approximation of the solution of the transport equation is verified numerically in section 4 as well. The numerical solution of the transport equation involving the operator Q_C in (1.3) is nontrivial because this equation is implicit. We propose a solution method which is amenable to particle discretizations.

2. Scaling. We start by bringing the Levinson equation (1.2) into an appropriate dimensionless form. Choosing scales p_0, t_0 for the momentum p and the time t, and rescaling S by s_0, we set

$$f(p,t) = \frac{1}{p_0^3} f_s(p_s, t_s), \quad S(p, p', t) = s_0 S_s(p_s, p_s', t_s),$$
$$\varepsilon(p) = \frac{p_0^2}{m_*} \varepsilon_s(p_s), \quad p_s = \frac{p}{p_0}, \quad t_s = \frac{t}{t_0},$$

where m_* denotes the effective electron mass, and we obtain

$$\begin{aligned}
&\text{(a)}\quad \partial_{t_s} f_s(p_s,t_s) = s_0 t_0^2 p_0^3 \int_0^{t_s} dt'_s \int dp'_s [S_s(p_s,p'_s,t_s-t'_s) f_s(p'_s,t'_s) \\
&\qquad\qquad - S_s(p'_s,p_s,t_s-t'_s) f_s(p_s,t'_s)], \\
&\text{(b)}\quad S_s(p_s,p'_s,t_s) = \frac{2VF^2 n}{s_0(2\pi\hbar)^3} \cos\left[\frac{t_0 p_0^2 t_s}{m_*\hbar}\left(\varepsilon_s(p_s)-\varepsilon_s(p'_s)-\frac{m_*\hbar\omega}{p_0^2}\right)\right] \\
&\qquad + \frac{2VF^2(n+1)}{s_0(2\pi\hbar)^3} \cos\left[\frac{t_0 p_0^2 t_s}{m_*\hbar}\left(\varepsilon_s(p_s)-\varepsilon_s(p'_s)+\frac{m_*\hbar\omega}{p_0^2}\right)\right],
\end{aligned} \tag{2.1}$$

where, in the case of parabolic bands, $\varepsilon_s(p_s) = \frac{|p_s|^2}{2}$ holds. For the rest of this paper we will restrict ourselves to parabolic bands and assume that the matrix element F of the electron-phonon interaction is constant with respect to momentum.

The parameter s_0 can be chosen more or less freely, since it cancels as soon as (2.1)(b) is inserted into (2.1)(a). We choose $s_0 = \frac{1}{t_0^2 p_0^3}$, which ensures that the resulting equation varies on an $O(1)$ scale in time. The key issue is now to choose an appropriate scale p_0 for the momentum variable. A natural choice is to scale the phonon energy to unity, which gives $p_0^2 = m_*\hbar\omega$. Furthermore, we will consider the Levinson equation on a timescale that is much larger than the timescale on which the lattice vibrates. Therefore we set $t_0 = (\lambda\omega)^{-1}$, where λ denotes a dimensionless parameter. We drop the subscript s from here on and obtain

$$\begin{aligned}
\partial_t f(p,t) &= \int_0^t dt' \int dp' [S(p,p',t-t') f(p',t') - S(p',p,t-t') f(p,t')], \\
S(p,p',t) &= \frac{2VF^2}{\lambda^2(2\pi)^3}\sqrt{\frac{m_*^3}{\hbar^3\omega}}\left(n\cos\left[\frac{t}{\lambda}(\varepsilon(p)-\varepsilon(p')-1)\right]\right. \\
&\qquad \left. + (n+1)\cos\left[\frac{t}{\lambda}(\varepsilon(p)-\varepsilon_s(p')+1)\right]\right).
\end{aligned}$$

Since the scattering rate varies on a timescale of order $\frac{1}{\lambda}$, the amplitude should be of the same order to keep the integral of order $O(1)$, which is obtained by setting

$$\lambda^2 = \frac{2VF^2(n+1)}{(2\pi)^3}\sqrt{\frac{m_*^3}{\hbar^3\omega}}. \tag{2.2}$$

This gives a scaled equation of the form

$$\begin{aligned}
&\text{(a)}\quad \partial_t f(p,t) = Q_\lambda[f](p,t) \\
&\qquad\qquad := \int_0^t dt' \int dp' [S_\lambda(p,p',t-t') f(p',t') - S_\lambda(p',p,t-t') f(p,t')], \\
&(b)\quad S_\lambda(p,p',t) = \sum_{\nu=\pm 1} \frac{a_\nu}{\lambda}\cos\left[\frac{t}{\lambda}(\varepsilon(p)-\varepsilon(p')+\nu)\right], \quad a_{-1} = \frac{n}{n+1}, \quad a_1 = 1.
\end{aligned} \tag{2.3}$$

Thus we consider an asymptotic regime where the quantity λ defined by (2.2) is small, and consider the asymptotic behavior of the collision operator in the Levinson equation for timescales $t_0 = (\lambda\omega)^{-1}$, which are much larger than the scale on which the lattice vibrates.

We conclude this section with the following observation, giving a heuristic argument for the convergence to the Fermi golden rule operator. Changing variables in the integral in (2.3)(a), we obtain

$$Q_\lambda[f](p,t) = \int_0^{t/\lambda} d\tau \int dp' [\lambda S_\lambda(p,p',\lambda\tau) f(p',t-\lambda\tau) - \lambda S_\lambda(p',p,\lambda\tau) f(p,t-\lambda\tau)],$$
$$\lambda S_\lambda(p,p',\lambda\tau) = \sum_{\nu=\pm 1} a_\nu \cos[\tau(\varepsilon(p) - \varepsilon(p') + \nu)].$$

If the term $\lambda S_\lambda(p,p',\lambda\tau)$, which is actually independent of λ, would decay for large τ, we could Taylor-expand the solution f and obtain in zeroth order

$$\partial_t f(p,t) = \int_0^\infty d\tau \int dp' [\lambda S(p,p',\lambda\tau) f(p',t) - \lambda S(p',p,\lambda\tau) f(p,t)],$$

which makes the collision operator local in time. The corresponding scattering rate would then be given by

$$\int_0^\infty \lambda S(p,p',\lambda\tau) d\tau = \sum_{\nu=\pm 1} a_\nu \int_0^\infty \cos[\tau(\varepsilon(p) - \varepsilon(p') + \nu)] d\tau,$$

and the integral over the cosine produces the δ-function in the Fermi golden rule. This heuristic argument has been given in [1], [3], [5]. Although S does not decay for large times, this result still holds, but the limit process is oscillatory; i.e., we have to compute a weak limit for $\lambda \to 0$. The computation of this weak limit is the subject of the present paper.

3. Asymptotics. In this section we derive the asymptotic behavior of the collision operator Q_λ in (2.3) for $\lambda \to 0$ and show that Q_λ indeed converges to the Fermi golden rule operator in the weak sense. More importantly, we are able to derive the first order term in the asymptotic expansion. This enables us to obtain a corrected Fermi golden rule operator which is still local in time, and thus a corrected Boltzmann equation which better reflects the effects of finite collision times. The main result of this section is stated in Theorem 3.2, which gives an asymptotic expression for the Levinson operator Q_λ in (2.3) up to terms of order $o(\lambda)$ in the weak sense. This approximation is still local in time in the sense that it depends only on the values of the density function f at time t and on its time derivative. The first order approximation is, although local in time, only given in a weak sense in p since the scattering rates in the loss term will contain integrals which exist only as principal value. The form of the resulting approximate collision operator is given in formulas (3.7) and (3.10).

Since we are considering a weak limit, we will define, for a given density f, the functional

$$Y_\lambda(f,\psi) = \int_0^\infty dt \int dp \psi(p,t) Q_\lambda[f](p,t)$$

for a smooth test function ψ, and investigate the limiting behavior of $Y_\lambda(f,\psi)$ for $\lambda \to 0$. It will be convenient to rewrite Y_λ using the adjoint of the collision operator. Interchanging the integration variables p and p' in the first part of (2.3)(a) gives

$$\text{(3.1)} \quad \begin{aligned} &\text{(a)} \quad Y_\lambda(f,\psi) = \int_0^\infty dt \int dp [f(p,t) Q_\lambda^{adj}[\psi](p,t)], \\ &\text{(b)} \quad Q_\lambda^{adj}[\psi](p,t) = \int_t^\infty dt' \int dp' [S_\lambda(p',p,t'-t)(\psi(p',t') - \psi(p,t'))], \end{aligned}$$

where we have also interchanged the time variables t, t'. In this form it is easy to see that the collision operator Q_λ conserves mass locally in time since the adjoint operator $Q_\lambda^{adj}(\psi)$ equals zero for test functions ψ which are constant in the momentum direction. The functional Y_λ represents a convolution in time and is therefore best expressed through Fourier transforms. To this end, we extend the definition (3.1)(b) of the adjoint collision operator $Q_\lambda^{adj}[\psi](p,t)$ for negative time. We define the Fourier transforms of the truncated density function f and the test functions by

$$\hat{f}(p,\tau) = \frac{1}{\sqrt{2\pi}} \int dt[H(t)f(p,t)e^{-i\tau t}], \qquad \hat{\psi}(p,\tau) = \frac{1}{\sqrt{2\pi}} \int dt[\psi(p,t)e^{-i\tau t}], \tag{3.2}$$

where $H(t)$ denotes the Heaviside function. From now on all integrals are to be understood as being over the whole real line or all of $\mathbb{R}^3$ unless stated explicitly otherwise. We have the following.

PROPOSITION 3.1. *In terms of the Fourier transform of the truncated density function f and the test function ψ, the functional $Y_\lambda(f,\psi)$ is given by*

$$\text{(a)} \quad Y_\lambda(f,\psi) = \int dp \int d\tau \{\hat{f}^*(p,\tau)\hat{Q}_\lambda^{adj}[\psi](p,\tau)\},$$

$$\text{(b)} \quad \hat{Q}_\lambda^{adj}[\psi](p,\tau) = \frac{1}{2} \sum_{\nu,\sigma=\pm 1} a_\nu \int dp' \left[\frac{\pi}{\lambda} \delta\left(\frac{\sigma}{\lambda} w_\nu(p',p) - \tau \right) + \frac{1}{i(\sigma w_\nu(p',p) - \lambda\tau)} \right] [\hat{\psi}(p',\tau) - \hat{\psi}(p,\tau)], \tag{3.3}$$

where * denotes the complex conjugate and $w_\nu(p',p) = \varepsilon(p') - \varepsilon(p) + \nu$ holds.

Proof of Proposition 3.1. If we define the Fourier transform of the adjoint operator $Q_\lambda^{adj}[\psi]$ by

$$\hat{Q}_\lambda^{adj}[\psi](p,\tau) = \frac{1}{\sqrt{2\pi}} \int dt \int dt' \int dp' H(t'-t) S_\lambda(p',p,t'-t)[\psi(p',t') - \psi(p,t')]e^{-i\tau t}, \tag{3.4}$$

the functional Y_λ becomes

$$Y_\lambda(f,\psi) = \int d\tau \int dp[\hat{f}^*(p,\tau)\hat{Q}^{adj}[\psi](p,\tau)],$$

where from here on * will denote the complex conjugate. In order to compute the Fourier transform of the adjoint collision operator Q^{adj}, we have to essentially compute the Fourier transform of the function $H(t)\cos(ut)$ as a distribution in t. We choose a sufficiently smooth test function $\phi(t)$, which is compactly supported in time, and compute the integral

$$\int dt[H(t)\cos(ut)\phi(t)].$$

We split the integrand into its even and odd parts by writing

$$\cos(ut)\phi(t) = a(u,t) + \partial_t b(u,t),$$

where a and b are both real and even functions in time. These functions are given by

$$a(u,t) = \frac{1}{2}\cos(ut) \sum_{\gamma=\pm 1} \phi(\gamma t), \qquad b(u,t) = -\frac{1}{2} \int_{|t|}^{\infty} ds \left[\cos(us) \sum_{\gamma=\pm 1} \gamma\phi(\gamma s) \right]$$

and are both compactly supported in time as well. Their Fourier transforms in time satisfy

$$\hat{a}(u,\tau) = \frac{1}{4}\sum_{\sigma,\gamma=\pm 1} \hat{p}hi(\gamma\tau - \sigma u), \qquad \hat{b}(u,\tau) = \frac{1}{4i\tau}\sum_{\sigma,\gamma=\pm 1} \gamma\hat{p}hi(\gamma\tau - \sigma u)$$

and are both well defined, also for $\tau \to 0$. Since both a and b are even functions of t, we can extend the integral now over the whole real line and write

$$\begin{aligned}\int dt[H(t)\cos(ut)\phi(t)] &= \int_0^\infty \cos(ut)\phi(t)dt = \frac{1}{2}\int a(u,t)dt - b(u,0)\\ &= \frac{\sqrt{2\pi}}{2}\hat{a}(u,0) - \frac{1}{\sqrt{2\pi}}\int \hat{b}(u,\tau)d\tau.\end{aligned}$$

Inserting the expressions for the Fourier transforms of a and b gives

$$\int dt[H(t)\cos(ut)\phi(t)] = \frac{1}{4\sqrt{2\pi}}\sum_{\sigma,\gamma=\pm 1}\left\{\pi\hat{p}hi(-\sigma u) - \int \frac{\gamma}{i\tau}\hat{p}hi(\gamma\tau - \sigma u)d\tau\right\}.$$

Shifting the integration variable in the second term gives, since the result does not depend on the summation index γ anymore,

$$\int dt[H(t)\cos(ut)\phi(t)] = \frac{1}{2\sqrt{2\pi}}\sum_{\sigma=\pm 1}\left\{\pi\hat{p}hi(-\sigma u) - \int \frac{\hat{p}hi(\tau)}{i(\tau+\sigma u)}d\tau\right\}.$$

Finally, we introduce a δ-function to have a more compact notation, obtaining that

$$\int dt[H(t)\cos(ut)\phi(t)] = \frac{1}{2\sqrt{2\pi}}\sum_{\sigma=\pm 1}\int\left[\pi\delta(\tau+\sigma u) - \frac{1}{i(\tau+\sigma u)}\right]\hat{p}hi(\tau)d\tau$$

holds for all test functions $\phi(t)$ which are compactly supported in t. Thus, in a weak sense, i.e., when integrated against the Fourier transform of compactly supported test functions,

$$[\widehat{H(t)\cos(ut)}](\tau) = \frac{1}{2\sqrt{2\pi}}\sum_{\sigma=\pm 1}\left[\pi\delta(\tau+\sigma u) + \frac{1}{i(\tau+\sigma u)}\right]$$

holds. Equipped with this, we can express the convolution integral for $\hat{Q}^{adj}$ in (3.4) as

$$\begin{aligned}\hat{Q}^{adj}[\psi](p,\tau) &= \sqrt{2\pi}\int dp'\{[H(t)\widehat{S_\lambda(p')},p,t)](-\tau)[\hat{\psi}(p',\tau) - \hat{\psi}(p,\tau)]\}\\ &= \frac{1}{2}\sum_{\nu,\sigma=\pm 1} a_\nu \int dp'\left[\frac{\pi}{\lambda}\delta\left(\frac{\sigma}{\lambda}w_\nu(p',p) - \tau\right)\right.\\ &\qquad\left. + \frac{1}{i(\sigma w_\nu(p',p) - \lambda\tau)}\right][\hat{\psi}(p',\tau) - \hat{\psi}(p,\tau)],\\ w_\nu(p',p) &:= \varepsilon(p') - \varepsilon(p) + \nu. \qquad \square\end{aligned}$$

We now change to energy-angle variables. We make the coordinate transformation in momentum space of the form

$$p \to (\varepsilon(p), p_0), \quad p_0 := \frac{p}{|p|},$$

where p_0 is a vector living only on the unit sphere. Carrying out this coordinate transformation in integrals and directional derivatives in the radial direction for the parabolic band energy $\varepsilon(p) = \frac{|p|^2}{2}$, this means

$$\int f(p)dp = \int_0^\infty d\varepsilon \int dp_0 f(\varepsilon, p_0)\sqrt{2\varepsilon}, \quad \int 1 dp_0 = 4\pi, \quad p^T \nabla_p f(p) = 2\varepsilon \partial_\varepsilon f(\varepsilon, p_0),$$

and the functional $Y_\lambda(f, \psi)$ in (3.3) is given by

$$\text{(a)} \quad Y_\lambda(f,\psi) = \int_0^\infty d\varepsilon \int dp_0 \int d\tau \left\{ \sqrt{2\varepsilon} \hat{f}^*(\varepsilon, p_0, \tau) \hat{Q}_\lambda^{adj}[\psi](\varepsilon, p_0, \tau) \right\},$$

$$\begin{aligned} \text{(b)} \quad \hat{Q}_\lambda^{adj}[\psi](\varepsilon, p_0, \tau) = \frac{1}{2} \sum_{\nu,\sigma=\pm 1} a_\nu \int_0^\infty d\varepsilon' \int dp_0' \\ \times \sqrt{2\varepsilon'} \Bigg[\pi \delta(\varepsilon' - \varepsilon + \nu - \sigma\lambda\tau) \\ + \frac{\sigma}{i(\varepsilon' - \varepsilon + \nu - \sigma\lambda\tau)} \Bigg] [\hat{\psi}(\varepsilon', p_0', \tau) - \hat{\psi}(\varepsilon, p_0, \tau)], \end{aligned} \tag{3.5}$$

where we have made use of the identity $\frac{1}{\lambda}\delta(\frac{z}{\lambda}) = \delta(z)$ and the fact that the δ-function is even. We now give the weak expansion of the collision operator Q_λ as follows.

THEOREM 3.2. *For any fixed test function $\psi(\varepsilon, p_0, t)$ whose Fourier transform in time $\hat{\psi}(\varepsilon, p_0, \tau)$ decays sufficiently fast, the value of the functional $Y_\lambda(f, \psi)$ can be written as*

$$\text{(a)} \quad Y_\lambda(f,\psi) = Y_0(f,\psi) + \lambda Y_1(f,\psi) + o(\lambda), \tag{3.6}$$

with Y_0 and Y_1 given by

$$\begin{aligned} \text{(b)} \quad Y_0(f,\psi) = \sum_{\nu=\pm 1} a_\nu \int_0^\infty d\varepsilon \int dp_0 \int_0^\infty d\varepsilon' \int dp_0' \int_0^\infty dt \\ \pi \delta(\varepsilon' - \varepsilon + \nu) \sqrt{2\varepsilon'}\sqrt{2\varepsilon} f(\varepsilon, p_0, t)[\psi(\varepsilon', p_0', t) - \psi(\varepsilon, p_0, t)], \end{aligned}$$

$$\begin{aligned} \text{(c)} \quad Y_1(f,\psi) = -\sum_{\nu=\pm 1} a_\nu \int_0^\infty d\varepsilon \int dp_0 \int_0^\infty d\varepsilon' \int dp_0' \int_0^\infty dt \\ \ln(|\varepsilon' - \varepsilon + \nu|) \partial_{\varepsilon'} \partial_\varepsilon \left(\sqrt{2\varepsilon'}\sqrt{2\varepsilon} f(\varepsilon, p_0, t) \partial_t [\psi(\varepsilon', p_0', t) - \psi(\varepsilon, p_0, t)] \right). \end{aligned}$$

The proof of Theorem 3.2 is deferred to the end of this section.

Remark: In the usual Cartesian coordinates this means that the collision operator Q_λ is given in weak form by

$$\text{(a)} \quad Q_\lambda[f] = Q_0[f] + \lambda Q_1[f] + o(\lambda), \tag{3.7}$$

$$\text{(b)} \quad Q_0[f](p,t) = \sum_{\nu=\pm 1} \pi a_\nu \int dp' [\delta(\varepsilon - \varepsilon' + \nu) f(p', t) - \delta(\varepsilon' - \varepsilon + \nu) f(p, t)],$$

$$\begin{aligned} \text{(c)} \quad \int dp [\phi(p) Q_1[f](p,t)] = \sum_{\nu=\pm 1} a_\nu \int dp \int dp' \\ \ln(|\varepsilon' - \varepsilon + \nu|) \frac{1}{4\varepsilon\varepsilon'} (p^T \nabla p)((p')^T \nabla_{p'}) \left[\sqrt{4\varepsilon\varepsilon'} \partial_t f(p,t)(\phi(p') - \phi(p)) \right], \end{aligned}$$

where the first order term Q_1 is formulated weakly in the momentum direction only, in order to guarantee that the integrals converge. What remains of the nonlocality in time of the Levinson operator Q_λ in (2.3) is that the operator Q_1 in (3.7)(c) acts on the time derivative of the density function f.

Remark: Theorem 3.2 states only the weak convergence of the Levinson operator (1.2) towards the Fermi golden rule operator (1.1)(b) and not the convergence of solutions of the Levinson equation towards solutions of the corresponding Boltzmann equation. Since solutions of the Boltzmann equation remain nonnegative for nonnegative initial data, a weak convergence result for solutions would actually imply that the Wigner function and its density matrix equivalent would remain nonnegative definite for all time.

Note that Theorem 3.2 holds only for a fixed function f which is independent of λ. However, its validity can be extended by considering a filtered collision operator, since convolution integrals are commutative. If we choose a test function ψ which is of the form $\psi(p,t) = \phi(p)\Gamma(s-t)$ for any s, whose Fourier transform is given by $\hat{\psi}(p,\tau) = \phi(p)\hat{\Gamma}(\tau)^* e^{i\tau s}$, then the Fourier transform of the convolution kernel can be transferred onto the Fourier transform of f, and (3.3)(a) reads

$$Y_\lambda(f_\lambda,\psi) = \sqrt{2\pi}\int dp \int d\tau\{\hat{f}^*_\lambda(p,\tau)\hat{\Gamma}^*(\tau)\hat{Q}^{adj}_\lambda[\phi(p)\delta(t-s)](p,\tau)\}.$$

Theorem 3.2 will still hold as long as the function $\hat{f}^*_\lambda(p,\tau)\hat{\Gamma}(\tau)$ decays sufficiently fast in the variable τ. This means that, even for a function which is oscillating rapidly in time, the filtered operator

$$Q^F_\lambda[f_\lambda](p,t) = \int \Gamma(t-s)Q_\lambda[f_\lambda](p,s)ds \tag{3.8}$$

will satisfy

$$Q^F_\lambda[f_\lambda] = Q_0[f^F_\lambda] + \lambda Q_1[f^F_\lambda] + o(\lambda) \tag{3.9}$$

pointwise in t, where the filtered signal $f^F_\lambda(p,t)$ is given by

$$\hat{f}^F_\lambda(p,\tau) = \sqrt{2\pi}\hat{\Gamma}(\tau)\hat{f}_\lambda(p,\tau), \qquad f^F_\lambda(p,t) = \int \Gamma(t-s)H(s)f_\lambda(p,s)ds.$$

The unscaled equation. Finally, we reverse the scaling of section 2 and write the corrected Fermi golden rule in dimensional variables. Undoing the scaling and choosing the strong form gives the following corrected Boltzmann equation:

$$\partial_t f(p,t) = \int dp' S_0(p,p')f(p',t) - \kappa_0(p)f(p,t) + \int dp' S_1(p,p')\partial_t f(p',t) - \partial_t\kappa_1(p)f(p,t). \tag{3.10}$$

The transition rates S_j and the out-scattering rates κ_j are

$$S_0(p,p') = \frac{V}{(2\pi\hbar)^3}\sum_{\nu=\pm1}\frac{2\pi}{\hbar}M^2\left(n+\frac{1}{2}+\frac{\nu}{2}\right)\delta(\varepsilon(p)-\varepsilon(p')+\nu\hbar\omega),$$

$$S_1(p,p') = \frac{V}{(2\pi\hbar)^3}\sum_{\nu=\pm1}2M^2\left(n+\frac{1}{2}+\frac{\nu}{2}\right)\frac{1}{(\varepsilon(p)-\varepsilon(p')+\nu\hbar\omega)^2},$$

$$\kappa_j(p) = \int dp' S_j(p',p), \quad j=0,1,$$

where $M = \hbar F$ holds, and it should again be pointed out that the strong form of the collision operator is purely formal; i.e., the integral in the out-scattering rate κ_1 is actually infinite, and the first order term has to be formulated in the weak form (3.7).

We conclude this section with the proof of Theorem 3.2.

Proof of Theorem 3.2. We start by writing (3.5) in a more compact form as

(3.11)
$$Y_\lambda(f,\psi) = \frac{1}{2}\sum_{\nu,\sigma=\pm 1} a_\nu \int d\varepsilon \int dp_0 \int d\varepsilon' \int dp_0' \int d\tau A_\sigma'(\varepsilon' - \varepsilon + \nu - \sigma\lambda\tau) B(\varepsilon,\varepsilon',p_0,p_0',\tau)\sqrt{2\varepsilon'}^{+},$$

where we have formally extended the integrals with respect to the energy variables ε, ε' over the whole real line and denote by $\sqrt{z}^{+}$ the truncated root; i.e., $\sqrt{z}^{+} = 0$ for $z < 0$ holds. This notation will simplify the further derivation. Here the function $A(u)$, its derivative $A'(u)$, and B are given by

(3.12)
$$\text{(a)}\quad A_\sigma(u) = \pi H(u) - i\sigma \ln(|u|), \qquad A_\sigma'(u) = \pi\delta(u) - \frac{i\sigma}{u},$$
$$\text{(b)}\quad B(\varepsilon,\varepsilon',p_0,p_0',\tau) = \sqrt{2\varepsilon}^{+}\hat{f}^{*}(\varepsilon,p_0,\tau)[\hat{\psi}(\varepsilon',p_0',\tau) - \hat{\psi}(\varepsilon,p_0,\tau)].$$

Shifting the ε' variable in (3.11) gives

$$Y_\lambda(f,\psi) = \frac{1}{2}\sum_{\nu,\sigma=\pm 1} a_\nu \int d\varepsilon \int dp_0 \int d\varepsilon' \int dp_0' \int d\tau A_\sigma'(\varepsilon' - \varepsilon + \nu) B(\varepsilon,\varepsilon' + \sigma\lambda\tau,p_0,p_0',\tau)\sqrt{2(\varepsilon' + \sigma\lambda\tau)}^{+}.$$

In principle, we are now going to Taylor-expand the function B with respect to the variable ε'. This is admissible since the variable ε' only appears in the argument of the test function $\hat{\psi}$ in the definition of B, and therefore the function $\partial_{\varepsilon'}B$ decays sufficiently fast in ε' as well. However, care has to be taken with the various singularities appearing in the integrals. We remove the singularity in the function A_σ' by integrating by parts with respect to ε and obtain

$$Y_\lambda(f,\psi) = \frac{1}{2}\sum_{\nu,\sigma=\pm 1} a_\nu \int d\varepsilon \int dp_0 \int d\varepsilon' \int dp_0' \int d\tau \; A_\sigma(\varepsilon' - \varepsilon + \nu)\partial_\varepsilon B(\varepsilon,\varepsilon' + \sigma\lambda\tau,p_0,p_0',\tau)\sqrt{2(\varepsilon' + \sigma\lambda\tau)}^{+}.$$

Now we Taylor-expand the function B with respect to the variable ε' and write

(3.13)
$$\partial_\varepsilon B(\varepsilon,\varepsilon' + \sigma\lambda\tau,p_0,p_0',\tau) = \partial_\varepsilon B(\varepsilon,\varepsilon',p_0,p_0',\tau) + \sigma\lambda\tau\partial_\varepsilon\partial_{\varepsilon'}B(\varepsilon,\varepsilon',p_0,p_0',\tau) + O(\lambda^2).$$

This is admissible since the function B is compactly supported in the variable ε'. Next we formally expand the volume element $\sqrt{2\varepsilon'}^{+}$ and write

$$\sqrt{2(\varepsilon' + \sigma\lambda\tau)}^{+} = \sqrt{2\varepsilon'}^{+} + \frac{\sigma\lambda\tau H(\varepsilon')}{\sqrt{2\varepsilon'}} + \frac{\lambda\sigma\tau}{(\varepsilon')^{\alpha}}R_\alpha(\varepsilon',\lambda\sigma\tau) \tag{3.14}$$

for some α, which of course is, at this point, only a definition for the remainder term R_α. Inserting (3.13) and (3.14) into the definition for $Y_\lambda(f,\psi)$, and neglecting the $O(\lambda^2)$ terms in (3.13), gives

$$Y_\lambda = Y_0 + \lambda Y_1 + \lambda Y_{2\lambda} + O(\lambda^2)$$

with

$$\begin{aligned}
&Y_0(f,\psi)\\
&= \frac{1}{2}\sum_{\nu,\sigma=\pm 1} a_\nu \int d\varepsilon \int dp_0 \int d\varepsilon' \int dp_0' \int d\tau A'_\sigma(\varepsilon'-\varepsilon+\nu)\sqrt{2\varepsilon'}^{+} B(\varepsilon,\varepsilon',p_0,p_0',\tau),\\
&Y_1(f,\psi)\\
&= \frac{1}{2}\sum_{\nu,\sigma=\pm 1} a_\nu \int d\varepsilon \int dp_0 \int d\varepsilon' \int dp_0' \int d\tau A_\sigma(\varepsilon'-\varepsilon+\nu)\sigma\tau\partial_{\varepsilon'}\partial_\varepsilon(\sqrt{2\varepsilon'}^{+} B(\varepsilon,\varepsilon',p_0,p_0',\tau)),\\
&Y_{2\lambda}(f,\psi)\\
&= \frac{1}{2}\sum_{\nu,\sigma=\pm 1} a_\nu \int d\varepsilon \int dp_0 \int d\varepsilon' \int dp_0' \int d\tau A_\sigma(\varepsilon'-\varepsilon+\nu)\frac{\sigma\tau}{(\varepsilon')^\alpha}\\
&\qquad \times \partial_\varepsilon B(\varepsilon,\varepsilon'+\sigma\lambda\tau,p_0,p_0',\tau)R_\alpha(\varepsilon',\lambda\sigma\tau).
\end{aligned}$$

Inserting the definition of the function A_σ from (3.12), we see that odd terms in σ will cancel in Y_0, and the even terms in σ will cancel in Y_1, giving

$$\begin{aligned}
&Y_0(f,\psi)\\
&= \sum_{\nu=\pm 1} a_\nu \int d\varepsilon \int dp_0 \int d\varepsilon' \int dp_0' \int d\tau\pi\delta(\varepsilon'-\varepsilon+\nu)\sqrt{2\varepsilon'}^{+} B(\varepsilon,\varepsilon',p_0,p_0',\tau),\\
&Y_1(f,\psi)\\
&= -\sum_{\nu=\pm 1} a_\nu \int d\varepsilon \int dp_0 \int d\varepsilon' \int dp_0' \int d\tau \ln(|\varepsilon'-\varepsilon+\nu|)i\tau\partial_{\varepsilon'}\partial_\varepsilon\left(\sqrt{2\varepsilon'}^{+} B(\varepsilon,\varepsilon',p_0,p_0',\tau)\right).
\end{aligned}$$

Reversing the Fourier transforms in time gives (3.6)(b,c). The term Y_0 produces the Fermi golden rule, and the term Y_1 the $O(\lambda)$ correction to it. It remains to estimate the term $Y_{2\lambda}$. Since the singularity in the integrand A_σ in ε' is only logarithmic, the integrals will converge for $\alpha < 1$. It therefore remains to choose α such that R_α remains uniformly bounded in ε'; i.e., we can write

(3.15)
$$|Y_{2\lambda}(f,\psi)| \le const \max\{|R_\alpha(\varepsilon',\lambda\sigma\tau)|,\ 0\le\varepsilon'<\infty,\ |\tau|\le K\} \quad \text{for } 0<\alpha<1,$$

where we only have to consider a finite range for τ, since the test function $\hat{\psi}$ can be assumed to be compactly supported. Thus we have to estimate the term

$$\max\{|R_\alpha(\varepsilon',z)|,\ 0\le\varepsilon'<\infty,\ |z|\le\lambda K\}.$$

According to (3.14), R_α is given by

$$R_\alpha(\varepsilon',z) = (\varepsilon')^\alpha\left[\frac{\sqrt{2(\varepsilon'+z)}^{+}-\sqrt{2\varepsilon'}^{+}}{z} - \frac{H(\varepsilon')}{\sqrt{2\varepsilon'}}\right]$$

or

$$R_\alpha(\varepsilon', z) = \begin{pmatrix} (\varepsilon')^\alpha \left[\dfrac{-\sqrt{2\varepsilon'}}{z} - \dfrac{1}{\sqrt{2\varepsilon'}} \right] & \text{for } 0 \le \varepsilon' \le \max\{-z, 0\} \\ (\varepsilon')^\alpha \left[\dfrac{\sqrt{2(\varepsilon'+z)} - \sqrt{2\varepsilon'}}{z} - \dfrac{1}{\sqrt{2\varepsilon'}} \right] & \text{for } \max\{-z, 0\} \le \varepsilon' \end{pmatrix}, \tag{3.16}$$

where the first row is relevant only for $z < 0$. If we choose $\alpha > \frac{1}{2}$, then we can estimate

$$(\varepsilon')^\alpha \left| \frac{-\sqrt{2\varepsilon'}}{z} - \frac{1}{\sqrt{2\varepsilon'}} \right| \le \sqrt{2}(-z)^{\alpha-1/2} + \frac{1}{\sqrt{2}}(-z)^{\alpha-1/2} = O(\lambda^{\alpha-1/2}) \quad \text{for } 0 \le \varepsilon' \le -z,$$

which takes care of the first row of (3.16). To estimate the second row of (3.16), we rewrite the expression as

$$\begin{aligned} &\left| (\varepsilon')^\alpha \left[\frac{\sqrt{2(\varepsilon'+z)} - \sqrt{2\varepsilon'}}{z} - \frac{1}{\sqrt{2\varepsilon'}} \right] \right| \\ &\quad = \frac{(\varepsilon')^\alpha}{\sqrt{2\varepsilon'}} \left| \frac{1 - \sqrt{1+\frac{z}{\varepsilon'}}}{1+\sqrt{1+\frac{z}{\varepsilon'}}} \right| = \frac{|z|^{\alpha-1/2}}{\sqrt{2}} \left| \frac{z}{\varepsilon'} \right|^{1/2-\alpha} \left| \frac{1 - \sqrt{1+\frac{z}{\varepsilon'}}}{1+\sqrt{1+\frac{z}{\varepsilon'}}} \right| \\ &\quad \le \frac{|z|^{\alpha-1/2}}{\sqrt{2}} \max_{-1 \le x < \infty} \left\{ |x|^{1/2-\alpha} \left| \frac{1-\sqrt{1+x}}{1+\sqrt{1+x}} \right| \right\} = O(\lambda^{\alpha-1/2}) \quad \text{for } \alpha > \frac{1}{2}. \end{aligned}$$

Thus, in summary, $\max\{|R_\alpha(\varepsilon', \lambda\sigma\tau)|, 0 \le \varepsilon' \le K, |\tau| \le K\} = O(\lambda^{\alpha-1/2})$ will hold for any $\alpha > \frac{1}{2}$, and because of (3.15), $Y_{2\lambda}(f, \psi) = O(\lambda^{\alpha-1/2})$ will hold for any $\frac{1}{2} < \alpha < 1$. Therefore $\lambda Y_{2\lambda}$ is actually a term of order $o(\lambda)$, although not of order $O(\lambda^2)$, and can be neglected in the first order approximation. Inserting the definition (3.12)(b) for the function B into Y_0, Y_1 and reversing the Fourier transforms gives the result. ∎

4. Numerical results. In this section we verify the asymptotic analysis of the previous section numerically. This verification will consist of two parts. The first part is concerned directly with the weak approximation of the operator Q_λ by $Q_0 + \lambda Q_1$, i.e., with the verification of Theorem 3.2. The more interesting question is of course in what sense the solution of the zero field Levinson equation (1.2) is approximated by the solution of the corresponding approximate equation. To answer this question rigorously we would need some form of stability or entropy estimate for the Levinson equation (1.2). This will be the subject of future work. Nevertheless, the second part of this section is devoted to a numerical study of this question, i.e., a numerical comparison of the solution of the Levinson equation to the solution of an appropriate approximate problem based on the result in Theorem 3.2.

Discretization of the collision operators. For reasons of computational simplicity, we choose a finite difference discretization of the involved collision operators Q_0, Q_1, and Q_λ. While the discretization of the full collision operator Q_λ in (2.3) and the zero order term Q_0 in (3.7)(b) (the Fermi golden rule) by finite differences is straightforward, some care has to be taken when discretizing the first order term Q_1 in (3.7)(c), since it is only formulated in a weak sense. This means that the corresponding strong formulation of the operator Q_1 will contain diverging integrals.

Integrating (3.6)(c) by parts to obtain the strong version of Q_1 gives

$$Q_1[f](p,t) = \int dp'[S_1(p,p')\partial_t f(p',t) - S_1(p',p)\partial_t f(p,t)]$$

with the first order scattering cross section S_1 given by

$$S_1(p,p') = \sum_{\nu=\pm 1} \frac{a_\nu}{(\varepsilon(p) - \varepsilon(p') + \nu)^2}, \tag{4.1}$$

and the resulting integral will be infinite in the strong formulation. We therefore discretize the first order operator Q_1 in a weak difference form. We start by choosing a mesh in energy and time direction of the form

$$M_\varepsilon = \{\varepsilon : \varepsilon = j\Delta\varepsilon,\ j = 0,1,\dots\}, \quad \Delta\varepsilon = \frac{1}{K}, \qquad M_t = \{t : t = n\Delta t,\ n = 0,1,\dots\},$$

where we choose $\Delta\varepsilon$ conveniently in such a way that the emission/absorption energy, which is equal to unity in our scaling, is an integer multiple of the mesh size. The density function f can be assumed to be a function of the energy only, so $f = f(\varepsilon,t)$ holds. Using parabolic bands ($\varepsilon = \frac{|p|^2}{2}$), integrals with respect to the momentum p are approximated by

$$\int f(p,t)dp \approx \Delta\varepsilon \sum_{j=0}^{\infty} f(j\Delta\varepsilon,t)dp(j\Delta\varepsilon), \quad dp(\varepsilon) := 4\pi\sqrt{2\varepsilon}.$$

The full collision operator Q_λ in (2.3) is now approximated by

$$\begin{aligned} Q_\lambda[f](j\Delta\varepsilon, n\Delta t) := \Delta t\Delta\varepsilon \sum_{n'=0}^{n}\sum_{j'=0}^{\infty} dp(j'\Delta\varepsilon) \\ \times[S_\lambda(j\Delta\varepsilon, j'\Delta\varepsilon, (n-n')\Delta t)f(j'\Delta\varepsilon, n'\Delta t) \\ - S_\lambda(j'\Delta\varepsilon, j\Delta\varepsilon, (n-n')\Delta t)f(j\Delta\varepsilon, n'\Delta t)] \end{aligned}$$

with S_λ given as in (2.3)(c). The Fermi golden rule operator Q_0 is discretized by

$$Q_0[f](j\Delta\varepsilon,t) = \sum_{\nu=\pm 1} \pi a_\nu[dp((j+\nu K)\Delta\varepsilon)f((j+\nu K)\Delta\varepsilon,t) - dp((j-\nu K)\Delta\varepsilon)f(j\Delta\varepsilon,t)], \tag{4.2}$$

where, for notational simplicity, we simply set $dp(\varepsilon) = 0$ for $\varepsilon < 0$. The first order collision operator Q_1 in (3.7)(c) is given in its weak formulation by

$$\begin{aligned} \int dp(\varepsilon)[\phi(\varepsilon)Q_1[f](\varepsilon,t)]d\varepsilon = 16\pi^2 \sum_{\nu=\pm 1} a_\nu \int d\varepsilon \int d\varepsilon' \\ \times \ln(|\varepsilon' - \varepsilon + \nu|)\partial_\varepsilon\partial_{\varepsilon'}\left[\sqrt{4\varepsilon\varepsilon'}\partial_t f(\varepsilon,t)(\phi(\varepsilon') - \phi(\varepsilon))\right]. \end{aligned}$$

In this weak formulation the integrals are guaranteed to converge. It is therefore allowed to truncate the logarithmic singularity in the integral kernel. We define

$$\ln^0(j\Delta\varepsilon) = \begin{cases} \ln(j\Delta\varepsilon), & j > 0, \\ \ln(\Delta\varepsilon), & j = 0, \end{cases}$$

and discretize Q_1 in a weak finite difference sense by requiring that

$$\begin{aligned}\Delta\varepsilon \sum_{j=0}^{\infty} dp(j\Delta\varepsilon)[\phi(j\Delta\varepsilon)Q_1[f](j\Delta\varepsilon, n\Delta t)] \\ = 16\pi^2(\Delta\varepsilon)^2 \sum_{\nu=\pm 1} a_\nu \sum_{j=0}^{\infty}\sum_{j'=0}^{\infty} \\ \ln^0(|j'-j+\nu K|\Delta\varepsilon) D_j^+ D_{j'}^+ \left[\sqrt{4jj'}\Delta\varepsilon D_n^+ f(j\Delta\varepsilon, n\Delta t)(\phi(j'\Delta\varepsilon)-\phi(j\Delta\varepsilon))\right]\end{aligned}$$

hold for all grid-test-functions ϕ. Here D^+ denotes the usual forward difference operators acting on the respective indices; i.e.,

$$\begin{aligned}&\text{(a)} \quad D_j^+ f(j\Delta\varepsilon, t) = \frac{f((j+1)\Delta\varepsilon, t) - f(j\Delta\varepsilon, t)}{\Delta\varepsilon}, \\ &\text{(b)} \quad D_n^+ f(\varepsilon, n\Delta t) = \frac{f(\varepsilon, (n+1)\Delta t) - f(\varepsilon, n\Delta t)}{\Delta t},\end{aligned} \tag{4.3}$$

holds. Expressing the first order collision operator in a strong form on the discrete level, i.e., choosing a discrete δ-function for the test function ϕ, gives

(4.4)
$$\begin{aligned}&Q_1[f](j\Delta\varepsilon, n\Delta t) \\ &= \Delta\varepsilon \sum_{j'=0}^{\infty} dp(j\Delta\varepsilon)[S_1(j\Delta\varepsilon, j'\Delta\varepsilon) D_n^+ f(j'\Delta\varepsilon, n\Delta t) - S_1(j'\Delta\varepsilon, j\Delta\varepsilon) D_n^+ f(j\Delta\varepsilon, n\Delta t)],\end{aligned}$$

with the discrete scattering cross section S_1 given by

$$S_1(j\Delta\varepsilon, j'\Delta\varepsilon) = \sum_{\nu=\pm 1} a_\nu D_j^- D_{j'}^- \ln^0 |(j-j'+\nu K)\Delta\varepsilon|,$$

which is the appropriate approximation to the singular integral kernel (4.1). Here D^- denotes the backward differencing operator, analogously to the definition of D^+ in (4.3).

We now proceed to verify Theorem 3.2 numerically. Besides the verification of the asymptotic analysis of the previous section, the purpose of this exercise is also to gain some confidence in the weak difference discretization before computing asymptotic solutions to the Levinson equation. More precisely, we will verify the consequence of Theorem 3.2 given in (3.9), namely that the smoothed version of the full collision operator Q_λ applied to a highly oscillatory function is approximated by the zero- and first order terms Q_0 and Q_1 applied to the smoothed function. Figure 1 shows the signal chosen for this verification, which consists of the function $f(\varepsilon, t) = (1 + \cos(20t))/(1 + 3\varepsilon^2)$, i.e., a smooth function of ε modulated by a rapid oscillation in time. Figure 2 shows the filtered signal $f^F(\varepsilon, t)$, obtained by convoluting f with a Gaussian in time. We now compute $Q_\lambda[f]$ and the corresponding smoothed version $Q_\lambda^F[f]$ according to (3.8) and compare the result to $(Q_0 + \lambda Q_1)[f^F]$. As a measure for the approximation we chose the energy given by the formula

$$\langle \varepsilon Q \rangle(t) = \Delta\varepsilon \sum_{j=0}^{\infty} dp(j\Delta\varepsilon) j\Delta\varepsilon Q(j\Delta\varepsilon, t).$$

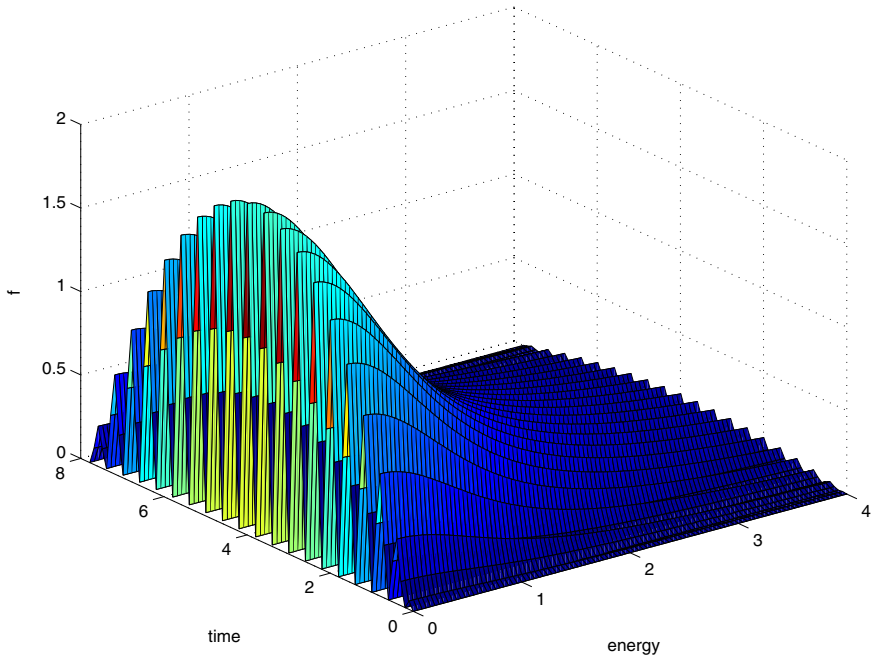


FIG. 1. *Unfiltered signal.*

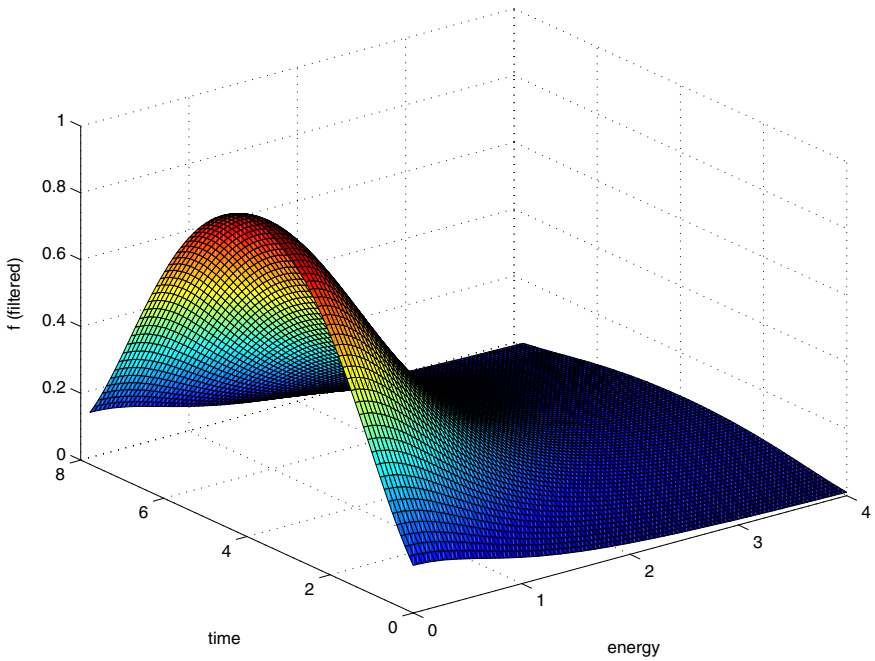


FIG. 2. *Filtered signal.*

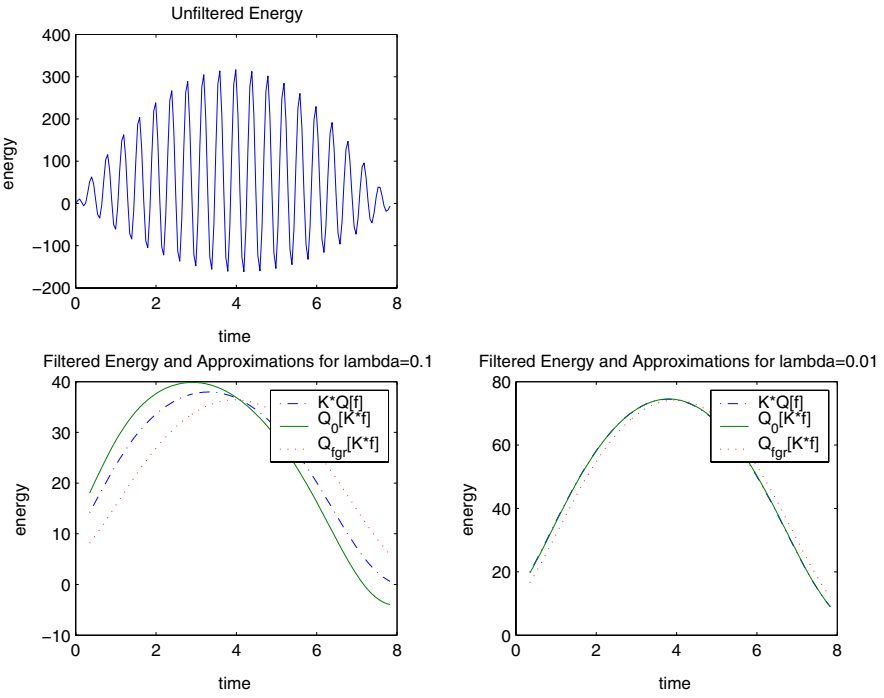


FIG. 3. *Comparison of energies.*

Figure 3 shows the highly oscillatory energy $\langle \varepsilon Q_\lambda[f] \rangle$ and compares $\langle \varepsilon Q_\lambda^F[f] \rangle$ to the values of $\langle \varepsilon Q_0[f^F] \rangle$ and $\langle \varepsilon (Q_0[f^F] + \lambda Q_1[f^F]) \rangle$ for $\lambda = 0.1$ and $\lambda = 0.01$. Figure 3 first confirms that the smoothed collision operator converges to the Fermi golden rule applied to the smoothed signal pointwise in time and that the approximation is improved by adding the first order correction, which is a direct consequence of the weak convergence given in Theorem 3.2.

We now turn to the more interesting question of whether, and in what sense, the solution of the zero field Levinson equation (1.2) is approximated by the solution of the asymptotic equation

$$\partial_t f = Q_0[f] + \lambda Q_1[f]. \tag{4.5}$$

To this end, we will compute with more realistic parameters. F in (1.2)(b) denotes the frequency of a particular lattice state and is given by the formula

$$F(\xi) = \sqrt{\frac{q^2 \hbar \omega}{2V|\xi|^2 \varepsilon_0} \left(\frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_s} \right)}, \tag{4.6}$$

where ξ is the momentum vector corresponding to the lattice state, ε_0 is the dielectricity constant (for vacuum), and ε_∞ and ε_s are the usual corrections to ε_0, taking into account the property of the crystal. The values for the physical parameters in

section 1 are summarized in the table below:

Symbol	Value	Unit	Meaning
q	$1.602 * 10^{-19}$	C	Electron charge
$\hbar$	$1.054 * 10^{-34}$	$\mathrm{kgm}^2/\mathrm{sec}$	Planck constant
h	$6.626196 * 10^{-34}$	$\mathrm{kgm}^2/\mathrm{sec}$	$h = 2\pi\hbar$
m_*	$0.063 * 0.109 * 10^{-31}$	kg	Effective electron mass
$\hbar\omega$	0.036	eV	Emission/absorption energy
ε_0	$8.85 * 10^{-12}$	$\frac{C}{V\mathrm{m}}$	Dielectricity constant (vacuum)
ε_∞	10.92	1	
ε_s	12.9	1	

We are considering the system at room temperature; i.e., the inverse temperature β in section 2 has a value of $\beta = 40(eV)^{-1}$, which gives a value of $n = 0.3105$ for the occupation number n. We consider only a single lattice state corresponding to the lattice being in equilibrium; i.e., we choose $|\xi|^2 = \frac{m_*}{\beta}$. Using these values, one computes a value of $\lambda = 0.0113$ for the dimensionless parameter λ, which suggests that we are in the appropriate asymptotic regime.

The asymptotic solution of the Levinson equation. The solution of the asymptotic equation (4.5) is complicated by the following facts. First, the equation is implicit in time, since the first order perturbation operator Q_1 acts on the time derivative of the solution f. Second, the implicit term is nonlocal in the energy variable, and third, this nonlocal implicit integral term contains a singular kernel. These factors make the actual numerical solution of (4.5) highly nontrivial. One could, for instance, be tempted to replace the time derivative of the density function in Q_1 in first order by $Q_0[f]$ and solve the explicit equation

$$\partial_t f = Q_0[f] + \lambda \int dp[S_1(p,p')Q_0[f](p',t) - S_1(p',p)Q_0[f](p,t)] \tag{4.7}$$

instead. It is, however, relatively easy to see (and has been verified numerically) that (4.7) is ill posed. At the level of computational complexity considered in this paper it would be feasible to directly discretize (4.5) using an implicit time discretization. We would then have to consider the artificial numerical diffusion generated by implicit methods, which is a major factor since we want to compare asymptotic solutions to the highly oscillatory solution of the Levinson equation. It should be pointed out here that we have made life particularly simple by considering solutions which are functions of energy only. As soon as we would introduce a field term, or consider spatially inhomogeneous problems, we would have to resort to some form of particle-based discretization, and the solution of implicit equations would become a major issue. With an eye to the future particle-based solution of inhomogeneous problems, the easiest way out of this dilemma is to actually solve for the asymptotic expansion of f given by (4.5). That is, to write $f = f_0 + \lambda f_1$ and to solve the system

$$\text{(a)} \quad \partial_t f_0 = Q_0[f_0], \qquad \text{(b)} \quad \partial_t f_1 = Q_0[f_1] + Q_1[f_0]. \tag{4.8}$$

Now the time derivative of the zero order term f_0, which appears in (4.8)(b), can be replaced by (4.8)(a), and we actually solve

$$\begin{aligned} &\text{(a)} \quad \partial_t f_0 = Q_0[f_0], \\ &\text{(b)} \quad \partial_t f_1 = Q_0[f_1] + \int dp[S_1(p,p')Q_0[f_0](p',t) - S_1(p',p)Q_0[f_0](p,t)]. \end{aligned} \tag{4.9}$$

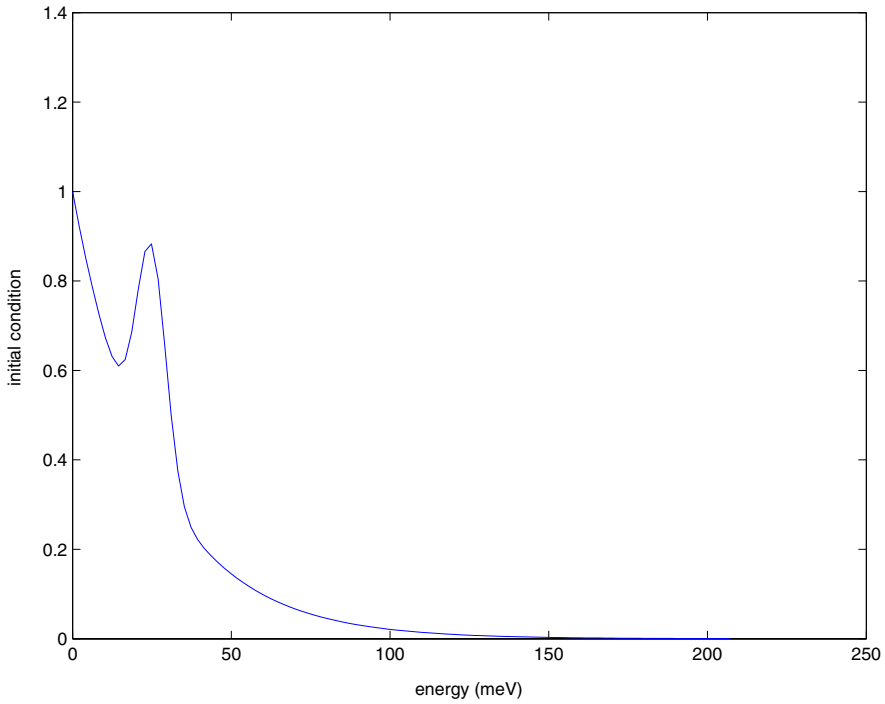


FIG. 4. *Initial condition for the Levinson equation.*

Equation (4.9)(a) is just the Boltzmann equation with the Fermi golden rule operator, and (4.9)(b) is also the same Boltzmann equation with an additional source term, which could be discretized using weighted particles. It is in this form that the numerical experiments below have been carried out, using the discretizations (4.2) and (4.4) for the operators Q_1 and Q_2.

We choose as initial condition (shown in Figure 4) the equilibrium Maxwellian with a second peak added. Thus we expect the second peak to be eliminated by the evolution of the Levinson equation as time advances. Figure 5 shows the solution f_λ of the Levinson equation (1.2) as a function of energy and time, and it exhibits the expected oscillations in time, albeit not to the same extent as the test example. To compare this solution with the asymptotic solution of (4.9), we smooth it in the same way as in the test example, i.e., by convoluting it with a Gaussian in time, shown in Figure 6. Figure 7 shows the solution of (4.9) for the same parameters. Figure 8 compares the full solution f_λ of the Levinson equation to the solution f_0 of the Fermi golden rule and $f_0 + \lambda f_1$ of (4.9) at different points in time. We observe that the solution f_0 of the Fermi golden rule has essentially reached steady state, while the full solution f_λ still evolves, i.e., the quantum effect causes a significantly longer relaxation time. This behavior is captured more or less by the asymptotic solution $f_0 + \lambda f_1$.

The structure of the equilibrium solution for the Fermi golden rule is determined by the fact that we have chosen a simple constant value of the lattice state frequency F in (4.6), corresponding to a δ-function collision potential [10]. This implies that the kernel of the Fermi golden rule operator Q_0 contains not only Maxwellians, but Maxwellians multiplied by arbitrary $\hbar\omega$-periodic functions of energy [7], [8], [12]. The steps in the equilibrium solution in the lower-right panel of Figure 8 represent the

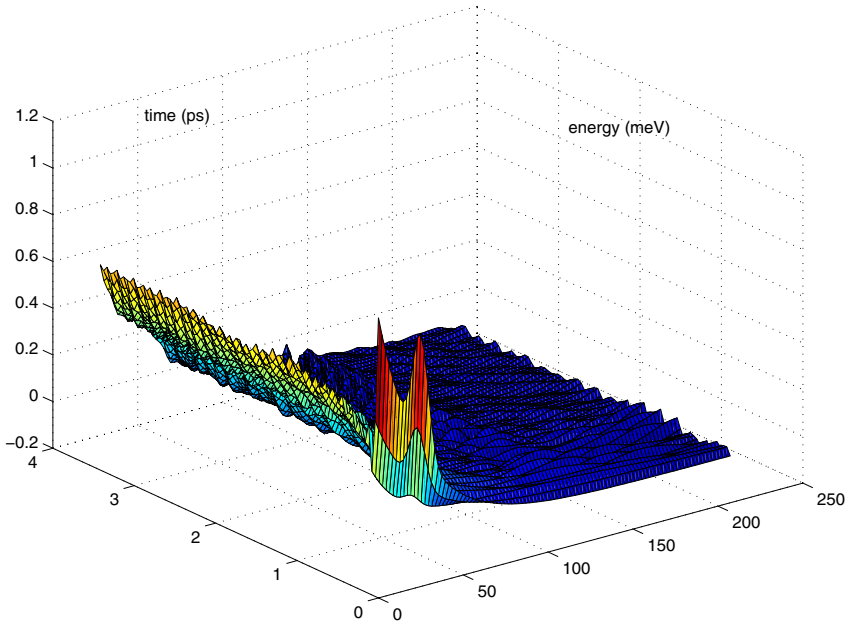


FIG. 5. *Solution of the Levinson equation.*

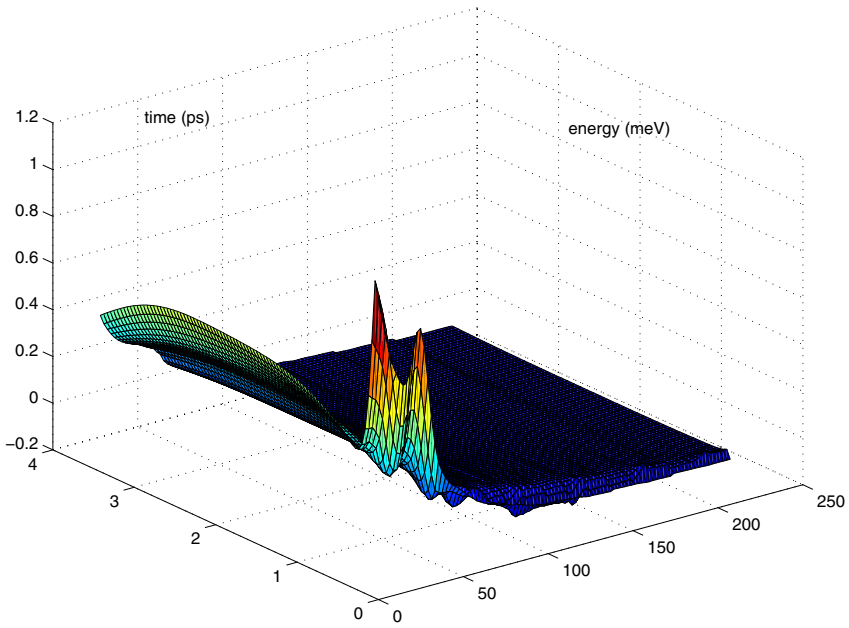


FIG. 6. *Filtered solution of the Levinson equation.*

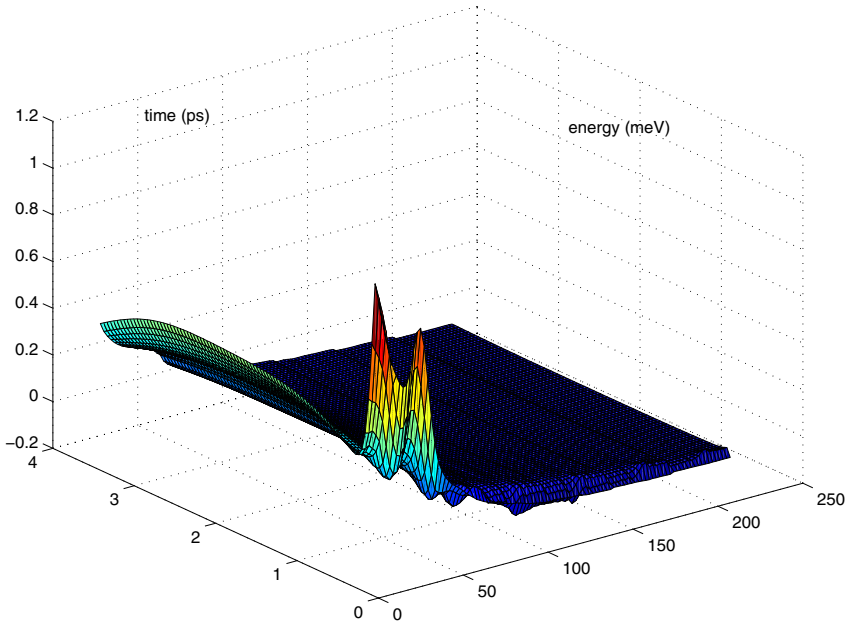


FIG. 7. *Approximate solution of the Levinson equation.*

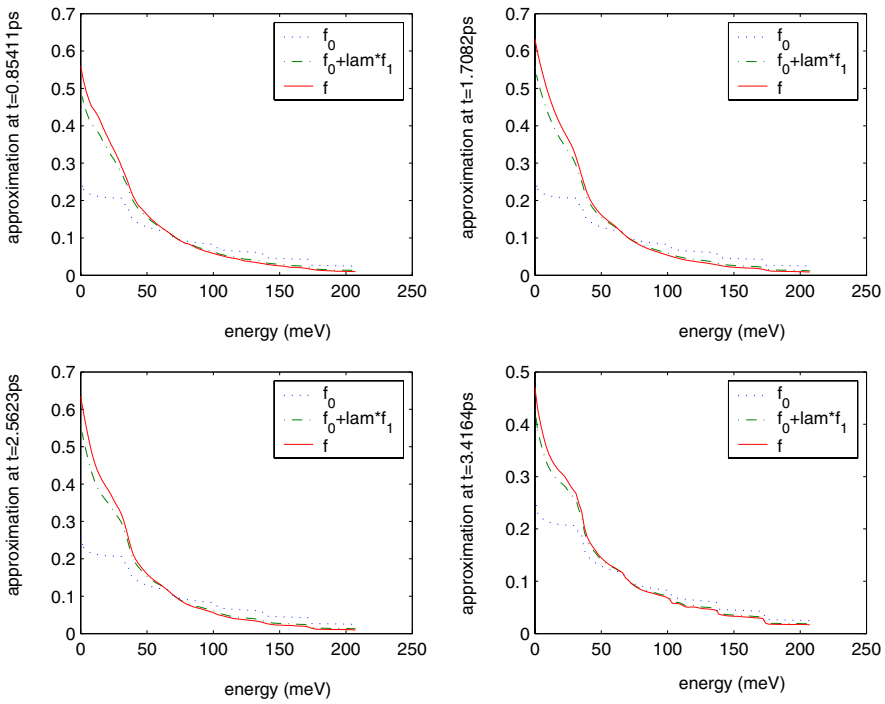


FIG. 8. *Snaphots of the solution and its approximations.*

projection of the initial solution into this space.

We were unable to continue the comparison beyond the given point in time due to memory constraints, since the solution of the Levinson equation requires time steps much smaller than λ to resolve the oscillations, and the storage of all previous time steps because of its nonlocality in time. It should be pointed out that the solution of the asymptotic system (4.9) does not suffer from these constraints, and (4.9) could be solved with much larger time steps on much longer timescales. The asymptotic solution $f_0 + \lambda f_1$ will eventually, however, converge to the same equilibrium solution, since the system (4.9) clearly has the same steady states as the original Fermi golden rule equation (1.2). Thus the quantum corrections give, at least in the absence of an electric field term, a purely transient effect.

5. Conclusions. Based on the Levinson equation, which in turn is derived from a weak interaction limit for the many body Schrödinger equation, we have derived a corrective term to the semiclassical Fermi golden rule collision operator. This corrective term is only mildly nonlocal in time in the sense that it is a local operator acting on the time derivative of the density function. It therefore renders itself much more easily to simulations on long timescales than did the original Levinson operator. We have shown the weak convergence of the corrected operator to the Levinson operator; i.e., we have proven the oscillatory limit for large times. Furthermore, we have demonstrated numerically the convergence of the solution to the Levinson equation towards the system resulting from the corresponding asymptotic expansion. From a numerical standpoint the complexity of this system is equivalent to that of solving a standard Boltzmann equation with additional source terms.

REFERENCES

[1] P. Argyres, *Quantum kinetic equations for electrons in high electric and phonon fields*, Phys. Lett. A, 171 (1992), pp. 43–61.

[2] N. Ashcroft and M. Mermin, *Solid State Physics*, Holt-Saunders, New York, 1976.

[3] J. Barker and D. Ferry, *Self-scattering path-variable formulation of high-field, time-dependent, quantum kinetic equations for semiconductor transport in the finite collision-duration regime*, Phys. Rev. Lett., 42 (1979), pp. 1779–1781.

[4] A. Bertoni, P. Bordone, R. Brunetti, and C. Jacoboni, *The Wigner function for electron transport in mesoscopic systems,* J. Phys. Cond. Matter, 11 (1999), pp. 5999–6012.

[5] F. Fromlet, P. Markowich, and C. Ringhofer, *A Wigner function approach to phonon scattering*, VLSI Design, 9 (1999), pp. 339–350.

[6] I. Levinson, *Translational invariance in uniform fields and the equation for the density matrix in the Wigner representation,* Sov. Phys. JETP, 30 (1970), pp. 362–367.

[7] P. Markowich, F. Poupaud, and C. Schmeiser, *Diffusion approximation of nonlinear electron-phonon collision mechanisms,* Math. Model. Numer. Anal., 29 (1995), pp. 857–869.

[8] P. Markowich and C. Schmeiser, *The drift-diffusion limit for electron-phonon interaction in semiconductors,* Math. Models Methods Appl. Sci., 7 (1997), pp. 707–729.

[9] M. Nedjalkov, R. Kosik, H. Kosina, and S. Selberherr, *A Wigner equation for nanometer and femtosecond transport regime,* in Proceedings of the First IEEE Conference on Nanotechnology, Maui, Hawaii, IEEE, Piscataway, NJ, 2001, pp. 277–281.

[10] F. Rossi and T. Kuhn, *Theory of ultrafast phenomena in photoexcited semiconductors*, Rev. Mod. Phys., 74 (2002), pp. 895–950.

[11] J. Schilp, T. Kuhn, and G. Mahler, *Electron-phonon quantum kinetics in pulse-excited semiconductors: Memory and renormalization effects,* Phys. Rev. B, 50 (1994), pp. 5435–5447.

[12] C. Schmeiser and A. Zwirchmayr, *Elastic and drift-diffusion limits of electron-phonon interaction in semiconductors,* Math. Models Methods Appl. Sci., 8 (1998), pp. 37–53.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 1954–1986

RETROFOCUSING OF ACOUSTIC WAVE FIELDS BY ITERATED TIME REVERSAL*

B. LARS G. JONSSON†, MATS GUSTAFSSON‡, VAUGHAN H. WESTON§, AND MAARTEN V. DE HOOP¶

Abstract. In the present paper an iterative time-reversal algorithm that retrofocuses an acoustic wave field to its controllable part is established. For a fixed temporal support, i.e., transducer excitation time, the algorithm generates an optimal retrofocusing in the least-squares sense. Thus the iterative time-reversal algorithm reduces the temporal support of the excitation from the requirement of negligible remaining energy to the requirement of controllability. The time-reversal retrofocusing is analyzed from a boundary-control perspective where time reversal is used to steer the acoustic wave field towards a desired state. The wave field is controlled by transducers located at subsets of the boundary, i.e., the controllable part of the boundary. The time-reversal cavity and time-reversal mirror cases are analyzed. In the cavity case, the transducers generate a locally plane wave in the fundamental mode through a set of ducts. Numerical examples are given to illustrate the convergence of the iterative time-reversal algorithm. In the mirror case, a homogeneous half space is considered. For this case the analytic expression for the retrofocused wave field is given for finite temporal support. It is shown that the mirror case does not have the same degree of steering as the cavity case. It is also shown that the pressure can be perfectly retrofocused for infinite temporal support. Two examples are given that indicate that the influence of the evanescent part of the wave field is small.

Key words. time reversal, retrofocusing, wave splitting, acoustic

AMS subject classifications. 74J20, 35L45, 35L50, 93B15, 49K20

DOI. 10.1137/S0036139903426964

1. Introduction.

Time-reversal acoustics is based on recording the wave field using a set of transducers, time-reversing the recorded signal, and retransmitting the result. The retransmitted wave field propagates back in the medium towards its source of origin [5, 9, 10, 12, 13]. In this paper, the time-reversal approach is analyzed from a boundary-control perspective [2, 3] in which time reversal is used to steer the acoustic wave field towards a desired state, corresponding to the original state. The boundary is divided into a controllable and an uncontrollable part. On the controllable part of the boundary, transducers are used to record or generate the acoustic wave field. The uncontrollable part of the boundary is assumed to be acoustically hard [26].

Both the time-reversal cavity and the time-reversal mirror have been extensively studied by Fink and coworkers; see, e.g., [5, 9, 10, 12, 13]. The cavity and mirror cases describe measurement situations with transducers surrounding the original source and only occupying a limited angular area, respectively. Applications of time-reversal

*Received by the editors May 1, 2003; accepted for publication (in revised form) January 26, 2004; published electronically August 19, 2004. This work was supported by the Swedish Research Council for Engineering Sciences and by the Wenner–Gren Foundation.
http://www.siam.org/journals/siap/64-6/42696.html

†The Fields Institute and the Department of Mathematics, University of Toronto, Toronto, Ontario, M5S 3G3 Canada, and Division for Electromagnetic Theory, The Alfvén Laboratory, Royal Institute of Technology, SE-100 44 Stockholm, Sweden (ljonsson@math.toronto.edu).

‡Department of Electroscience, Lund Institute of Technology, Lund University, Box 118, SE-221 00 Lund, Sweden (mats@es.lth.se).

§Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395. Current address: 2222 Carberry Dr., West Lafayette, IN 47906 (vhw2222crby@worldnet.att.net).

¶Center for Wave Phenomena, Colorado School of Mines, Golden, CO 80401-1887 (mdehoop@mines.edu).

algorithms include lithotripsy, pulse focusing, medical imaging, inverse scattering [10, 13], and optimal distinguishability measurements [6, 7]. The time-reversal approach gives a perfect retrofocusing if the transducers surround the original source, i.e., the time-reversal cavity, and the wave field is recorded until the wave field is quiescent; see, e.g., [2]. If the conditions for local energy decay are satisfied [1, 23], the retrofocusing error can be made arbitrarily small as observation or measurement time approaches infinity. An analysis of the super-resolving property of the time-reversal mirror is presented in [4]. Applications to communications are given in [19].

From boundary control theory, it is known that transducers can steer the wave field towards an arbitrary field distribution if the region is controllable [3, 22]. If the configuration is not controllable, an optimal control produces an optimal retrofocused wave field. The present paper establishes an iterative time-reversal algorithm that retrofocuses an acoustic wave field to its controllable part. The obtained iterative time-reversal algorithm reduces the temporal support of the transducer excitation from the requirement of negligible remaining energy to the requirement of controllability. In particular, for a fixed temporal support of the excitation, the algorithm generates an optimal retrofocusing in the least-squares sense. The characteristics of the transducers are included.

The considered cavity is a bounded domain with a perforated acoustically hard boundary. Transducers induce the wave field through the fundamental (or plane wave) mode in a set of ducts [26]. Numerical examples are given to illustrate the convergence of the iterative time-reversal algorithm. In the mirror case, a homogeneous half space is considered. For this case, it is shown that the pressure can be perfectly retrofocused for infinite temporal support. For finite temporal support the algorithm gives an optimal control for the propagating part of the wave field. A closed-form representation for the retrofocused wave field, with finite temporally supported excitation, from an initial Dirac-pressure source is given. Its behavior in the long time limit is calculated. Two examples of retrofocusing of a pressure source are given, and the influence of the evanescent part of the wave is discussed. In both the cavity and the mirror the iterated time-reversal procedure does not make use of the medium properties of the interior.

The optimal measurements [6, 7] are discussed in the sense of maximal measured least-squares distinguishability of a scattering operator relative to a reference scattering operator. The obtained algorithm is the limit of a renormalized series of iterated time reversals. It is noted in [6] that the limit of this renormalized series is a time-harmonic, frequency tuned wave form that corresponds to the frequency such that the largest eigenvalue of the squared time-reversed reflection operator attains its maximum in a given frequency interval. The algorithm in the present paper uses *the sum* of the iterated squared time-reversed response operators to obtain the retrofocusing of an initially prescribed field. Thus the resulting control for retrofocusing has a multifrequency content, as opposed to the algorithms proposed in [6, 7].

The present paper begins with a short discussion of the boundary control of acoustic wave fields. Transducers are introduced, and their restrictions on the boundary conditions are analyzed. In section 3, the sufficient conditions for optimal boundary control with respect to a least-squares measure is given. As the first example we consider the case where the material parameters are unknown; see section 4. In section 5, a cavity with acoustically hard walls and attached ducts is analyzed. Numerical examples are given. The time-reversal algorithm is shown to efficiently retrofocus the field of an initial pulse source. In section 6, the mirror case is analyzed for a homogeneous material. The optimal control for measurements with negligible evanescent

part is derived. For the nonnegligible evanescent part of the measured response, the closed-form expression for the time reversal of an initial pressure source is derived. Here we utilize a constant wavespeed to obtain an analytical expression. It is shown that, for an initial pressure distribution, the evanescent part of the control has a marginal influence on the resulting field in the long excitation time limit. Section 7 is a discussion of the results.

2. The acoustic wave field and boundary control.

2.1. The control state. In this section we state the acoustic equations, the boundary conditions, and initial values assumed in the further analysis. The goal is to focus the acoustic wave field to a desired state at a given finite time. To quantify the focusing, relative to the desired state, we use a weighted L^2 measure, in the form of an energy functional. We also introduce the boundary control, the admittance operator, a representation of the characteristics of a transducer, and the response operator due to a boundary control.

The control state is an acoustic wave field in the domain $\Omega \in \mathbb{R}^3$ and time interval $[0, T]$. The boundary of the domain, $\partial\Omega$, is assumed to be piecewise C^1, and thus the normal to the boundary is well defined almost everywhere. At the open subset $\Gamma_{\mathrm{t}} \subset \partial\Omega$, *the controllable part of the boundary*, we have a set of transducers that are used to generate an acoustic wave field in the domain. The control state at time t is represented by the pressure, $p = p(\boldsymbol{x}, t)$, and the particle velocity, $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t)$. Given a desired final state, $\{p_T, \boldsymbol{v}_T\}$, we quantify the degree of focusing between the control state at time T, $\{p(\cdot, T), \boldsymbol{v}(\cdot, T)\}$, and the desired state, with the weighted L^2 functional

$$\mathcal{J} = \frac{1}{2} \int_{\Omega} (\kappa(\boldsymbol{x}) |p(\boldsymbol{x}, T) - p_T(\boldsymbol{x})|^2 + \rho(\boldsymbol{x}) |\boldsymbol{v}(\boldsymbol{x}, T) - \boldsymbol{v}_T(\boldsymbol{x})|^2) \, \mathrm{dV}(\boldsymbol{x}). \tag{2.1}$$

The control state $\{p, \boldsymbol{v}\}$ satisfies the source-free acoustic equations

$$\begin{cases} \kappa(\boldsymbol{x}) \partial_t p + \nabla \cdot \boldsymbol{v} = 0 \\ \rho(\boldsymbol{x}) \partial_t \boldsymbol{v} + \nabla p = \boldsymbol{0} \end{cases} \quad \text{for } \boldsymbol{x} \in \Omega \text{ and } t \in [0, T], \tag{2.2}$$

where the compressibility $\kappa(\boldsymbol{x})$ and the density $\rho(\boldsymbol{x})$ model the interaction between the acoustic wave field and the material. To ensure the existence of solutions, it is assumed that material parameters κ and ρ are positive and belong to $\mathrm{L}^\infty(\Omega)$; i.e., the parameters are bounded and measurable. No explicit knowledge of the parameters is assumed, except that they belong to the mentioned class. In the process of retrofocusing, we assume the initial conditions

$$p(\boldsymbol{x}, 0) = 0 \quad \text{and} \quad \boldsymbol{v}(\boldsymbol{x}, 0) = \boldsymbol{0} \quad \text{for } \boldsymbol{x} \in \Omega. \tag{2.3}$$

In our model the transducers are supported on the controllable part Γ_{t} of the boundary; see Figure 5.1. Given a boundary control [22], $p_+ = p_+(\boldsymbol{x}, t)$, $\boldsymbol{x} \in \Gamma_{\mathrm{t}}$, the characteristics of the transducers determine how the field is induced in the domain. Here, we model the transducer characteristic with the transducer's admittance operator $\mathcal{Y}$ that maps its domain $\mathcal{D} \subset \mathrm{L}^2$ to L^2 and is invertible. Thus the boundary condition at the controllable part of the boundary, Γ_{t} (see Figure 2.1) takes the form,

$$\frac{(\mathcal{Y}p)(\boldsymbol{x}, t) + v_{\mathrm{n}}(\boldsymbol{x}, t)}{2} = p_+(\boldsymbol{x}, t) \quad \text{for } \boldsymbol{x} \in \Gamma_{\mathrm{t}} \text{ and } t \in [0, T], \tag{2.4}$$

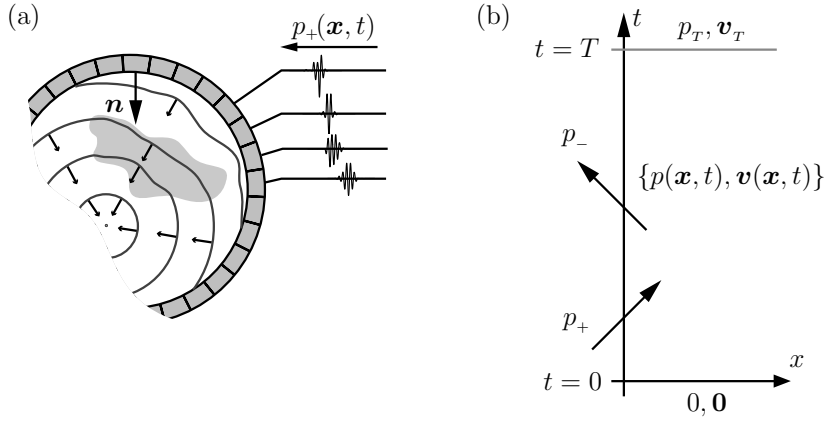


FIG. 2.1. *The control state*: *The boundary control p_+ is prescribed at the boundary, and given quiescent initial conditions, we measure the response p_-. The desired final state $\{p_T, \boldsymbol{v}_T\}$ is shown as the state at time $t = T$ in panel* (b).

where v_n is the normal component of the particle velocity, i.e., $v_\mathrm{n} = \boldsymbol{v} \cdot \boldsymbol{n}$ and $\boldsymbol{n}$ is the inward unit normal vector to the boundary; see Figure 2.1(a). The boundary condition above is said to be in the velocity normalization [16, 17, 18]. The case with $\mathcal{Y} = 1$ in (2.4) represents boundary conditions in the form of a locally plane wave propagating inward into the domain, i.e., in the $\boldsymbol{n}$-direction [26], and this case is considered in section 5. The time-reversal mirror case with the transducer modeled by the wave splitting operator [18] is considered in section 6.

The remaining, noncontrollable, enclosed part of the boundary, the wall that we denote as *the uncontrollable part* of the boundary, $\Gamma_\mathrm{w} = \partial\Omega \setminus \Gamma_\mathrm{t}$ (see, e.g., Figure 5.1), is made of a particular material with characteristics $\mathcal{U}$, giving the boundary condition

$$(\mathcal{U}p)(\boldsymbol{x},t) + v_\mathrm{n}(\boldsymbol{x},t) = 0 \quad \text{for } \boldsymbol{x} \in \Gamma_\mathrm{w} \text{ and } t \in [0,T]. \tag{2.5}$$

Here, $\mathcal{U}$ is a continuous mapping from L^2 to L^2. Observe that $\mathcal{U} = 0$ in (2.5) corresponds to an acoustically hard (uncontrollable) boundary. We use this as an example.

The operators $\mathcal{Y}$ and $\mathcal{U}$ are chosen such that the acoustic wave equation is well posed with this boundary condition [20, section 8.2]. The Kreiss–Lorenz class of boundary conditions is of the type $p_\mathrm{in} = \mathcal{C}p_\mathrm{out}+$ source terms, where p_in and p_out are defined as the eigenvectors to the matrix of the normal to the boundary-derivative, in our case $p \pm v_\mathrm{n}$. For this class of boundary conditions, there is a requirement on the principal part size of the coupling term $\mathcal{C}$. In our case, $\mathcal{C} = (\mathcal{Y}^{-1} - 1)(\mathcal{Y}^{-1} + 1)^{-1}$. In the examples that we consider, $\mathcal{Y} = 1$ or $p_\mathrm{out} = 0$, and hence the coupling term vanishes. For a more detailed discussion and more general boundary conditions, see [20]. The above discussion holds also for the uncontrollable part of the boundary, i.e., for $\mathcal{U}$.

We assume that the transducers can also be used as receivers, and we measure the outgoing field component, p_-, at the boundary; see Figure 2.1(b). This component is given by

$$p_-(\boldsymbol{x},t) = \frac{(\mathcal{Y}p)(\boldsymbol{x},t) - v_\mathrm{n}(\boldsymbol{x},t)}{2} \quad \text{for } \boldsymbol{x} \in \Gamma_\mathrm{t} \text{ and } t \in [0,T]. \tag{2.6}$$

The relation between the boundary control p_+ and the measured outgoing field component p_-, for zero initial conditions, is the response operator $\mathcal{R}$, which is also called the scattering operator or reflection operator [6, 7],

$$p_-(\boldsymbol{x},t) = (\mathcal{R}p_+)(\boldsymbol{x},t) \quad \text{for } \boldsymbol{x} \in \Gamma_{\mathrm{t}} \text{ and } t \in [0,T]. \tag{2.7}$$

Given a boundary control, we assume that we can obtain its response by measurements on the domain, i.e., using the transducers to measure the response to the given boundary control.

In our formulation, the boundary control is not uniquely determined. There are many acoustic wave field configurations where it is clear that several boundary controls minimize $\mathcal{J}$. One such example is a homogeneous slab where we consider long enough excitation times such that the wave field can pass through the slab. The final internal field in the slab does not depend on fields that have left the slab, and hence the control is not unique.

2.2. The observation states. In this section, we define the notion of initial observation, an observation, and an observation state. The observation of an observation state is used in constructing the boundary control for the control state.

To distinguish between the control states and the field used for observation, we introduce the notation $\{q(\boldsymbol{x},t), \boldsymbol{u}(\boldsymbol{x},t)\}$ for the observation state at times $t \in [-T,0]$. The observation state $\{q, \boldsymbol{u}\}$ solves (2.2), and for convenience we let the observation take place in the time interval $[-T,0]$.

We define the initial observation state through its initial conditions

$$q(\boldsymbol{x},-T) = p_T(\boldsymbol{x}) \quad \text{and} \quad \boldsymbol{u}(\boldsymbol{x},-T) = -\boldsymbol{v}_T(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \Omega. \tag{2.8}$$

The observation of the initial observation state is carried out with the receivers coinciding with the transducers for the control state; i.e., the measurement in terms of the field at the boundary is

$$q_-(\boldsymbol{x},t) = \frac{(\mathcal{Y}q)(\boldsymbol{x},t) - u_{\mathrm{n}}(\boldsymbol{x},t)}{2} \quad \text{for } \boldsymbol{x} \in \Gamma_{\mathrm{t}} \text{ and } t \in [-T,0]. \tag{2.9}$$

Consequently, the boundary condition for the controllable part of the boundary takes the form

$$\frac{(\mathcal{Y}q)(\boldsymbol{x},t) + u_{\mathrm{n}}(\boldsymbol{x},t)}{2} = q_+ \quad \text{for } \boldsymbol{x} \in \Gamma_{\mathrm{t}} \text{ and } t \in [-T,0]. \tag{2.10}$$

The initial observation is thus given by $q_-^{(0)} = q_-$, setting $q_+ = 0$ in (2.10). On the uncontrollable part of the boundary, the field satisfies the boundary condition (cf. (2.5))

$$(\mathcal{U}q)(\boldsymbol{x},t) + u_{\mathrm{n}}(\boldsymbol{x},t) = 0 \quad \text{for } \boldsymbol{x} \in \Gamma_{\mathrm{w}} \text{ and } t \in [-T,0]. \tag{2.11}$$

The relation between q_- and q_+ and the initial conditions are, by the superposition principle,

$$q_-(\boldsymbol{x},t) = (\mathcal{R}q_+)(\boldsymbol{x},t) + q_-^{(0)}(\boldsymbol{x},t) \quad \text{for } \boldsymbol{x} \in \Gamma_{\mathrm{t}} \text{ and } t \in [-T,0], \tag{2.12}$$

where $\mathcal{R}$ coincides with the $\mathcal{R}$ in (2.7) since q_- and p_- are both the measured response with the same receivers from the acoustic wave equations with identical domains and type of boundary conditions.

With the above given information about the response operator and the initial observation, we will search for an optimal boundary control, i.e., the control applied in (2.4) such that the resulting control state at time T minimizes the least-squares functional $\mathcal{J}$ in (2.1), for a given $\{p_T, \boldsymbol{v}_T\}$.

3. Retrofocusing by time reversal. The time-reversal operator $\mathcal{T}$ is defined as

$$(\mathcal{T}p)(\boldsymbol{x},t) = p(\boldsymbol{x},-t). \tag{3.1}$$

If $\{p,\boldsymbol{v}\}$ solves the acoustic wave equation (2.2), then so does $\{\mathcal{T}p, -\mathcal{T}\boldsymbol{v}\}$.

3.1. Energies. In this section, we define the energy corresponding to the acoustic wave field, and reformulate the least-squares functional $\mathcal{J}$ on the interior of the domain into a functional on the boundary.

The energy at time t of the observation state $\{q(\boldsymbol{x},t), \boldsymbol{u}(\boldsymbol{x},t)\}$ is defined by

$$E[q,\boldsymbol{u}](t) = \frac{1}{2}\int_\Omega (\kappa(\boldsymbol{x})|q(\boldsymbol{x},t)|^2 + \rho(\boldsymbol{x})|\boldsymbol{u}(\boldsymbol{x},t)|^2)\,\mathrm{dV}(\boldsymbol{x}). \tag{3.2}$$

Energy conservation is given as

$$E[q,\boldsymbol{u}](0) - E[q,\boldsymbol{u}](-T) = \int_{-T}^0 \int_{\partial\Omega} q(\boldsymbol{x},t)u_{\mathrm{n}}(\boldsymbol{x},t)\mathrm{dS}(\boldsymbol{x})\,\mathrm{d}t, \tag{3.3}$$

by the Gauss theorem with the direction of a normal unit vector as in Figure 2.1(a).

To rewrite $\mathcal{J}$ in a form more suitable for our analysis, consider the summation of the equations for p, $\boldsymbol{v}$ and $\mathcal{T}q$, $\mathcal{T}\boldsymbol{u}$:

$$\begin{cases} \kappa(\boldsymbol{x})\partial_t(p-\mathcal{T}q) + \nabla\cdot(\boldsymbol{v}+\mathcal{T}\boldsymbol{u}) = 0 \\ \rho(\boldsymbol{x})\partial_t(\boldsymbol{v}+\mathcal{T}\boldsymbol{u}) + \nabla(p-\mathcal{T}q) = \boldsymbol{0} \end{cases} \quad \text{for } \boldsymbol{x}\in\Omega \text{ and } t\in[0,T]. \tag{3.4}$$

Multiplication of the equations with $p-\mathcal{T}q$ and $v+\mathcal{T}u$, respectively, and integration over time and space together with the Gauss theorem gives

$$\begin{aligned} &\int_\Omega \kappa(\boldsymbol{x})|p(\boldsymbol{x},t)-(\mathcal{T}q)(\boldsymbol{x},t)|^2 + \rho(\boldsymbol{x})|\boldsymbol{v}(\boldsymbol{x},t)+(\mathcal{T}\boldsymbol{u})(\boldsymbol{x},t)|^2\,\mathrm{dV}(\boldsymbol{x})\big|_{t=0}^T \\ &\quad = \int_0^T\int_{\partial\Omega} (p(\boldsymbol{x},t)-(\mathcal{T}q)(\boldsymbol{x},t))(v_{\mathrm{n}}(\boldsymbol{x},t)+(\mathcal{T}u_{\mathrm{n}})(\boldsymbol{x},t))\mathrm{dS}(\boldsymbol{x})\,\mathrm{d}t. \end{aligned} \tag{3.5}$$

Let $E_0 = E[q,\boldsymbol{u}](0)$, and recall that at $t=0$, $p=0$ and $\boldsymbol{v}=0$. Now, using the initial condition for q, $\boldsymbol{u}$ from (2.8), the left-hand side at $t=T$ is $\mathcal{J}$, and at $t=0$ is $-E_0$. Thus

$$\mathcal{J} = E_0 + \int_0^T\int_{\partial\Omega} (p(\boldsymbol{x},t)-(\mathcal{T}q)(\boldsymbol{x},t))(v_{\mathrm{n}}(\boldsymbol{x},t)+(\mathcal{T}u_{\mathrm{n}})(\boldsymbol{x},t))\mathrm{dS}(\boldsymbol{x})\,\mathrm{d}t. \tag{3.6}$$

3.2. Controllability. In this section, we show that the uncontrollable subspace of the wave-field solutions is orthogonal to the controllable subspace of wave-field solutions. We introduce the concept of "equal fields on the boundary."

In general, it is not possible to retrofocus the wave field to the desired wave field $\{p_T, \boldsymbol{v}_T\}$. At best the wave field retrofocuses to its controllable part, i.e., to the part of $\{p_T, \boldsymbol{v}_T\}$ that is possible to reach from the boundary. This projection to the

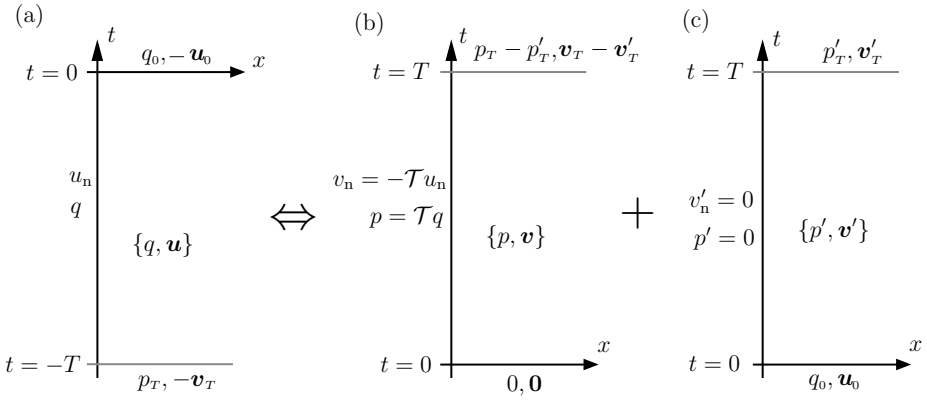


FIG. 3.1. *Controllable and uncontrollable subspaces for the equal fields on the boundary. Boundary fields such that* $p = \mathcal{T} q$ *and* $v_n = -\mathcal{T} u_n$ *decompose the original state into its controllable and uncontrollable parts.* (a) *In the observation state, the wave field is decomposed into the observed boundary field* $\{q, u_n\}$ *and the nonobserved remaining field* $\{q_0, -\boldsymbol{v}_0\}$. (b) *In the control state, the boundary field* (2.6) *is time-reversed and retransmitted into the region.* (c) *The error term corresponds to the nonobservable and noncontrollable part of the wave field.*

controllable part is achieved when the boundary fields are identical in the control and observation state, viz.,

$$p = \mathcal{T} q \quad \text{and} \quad v_n = -\mathcal{T} u_n \quad \text{when } \boldsymbol{x} \in \Gamma_t \text{ and } t \in [0, T]. \tag{3.7}$$

This goal of retrofocusing cannot always be achieved. For an example of imperfect retrofocus, see section 6. We denote the condition (3.7) as the "equal fields on the boundary" condition. If the "equal fields on the boundary" condition is achieved, it gives a constructive description of the controllable, uncontrollable, observable, and unobservable parts of the wave field. Observe that it is not possible to enforce (3.7) for the acoustic wave equation together with arbitrary initial data. In general, one set of boundary condition (2.4), (2.5), uniquely determines both the pressure and the particle velocity in the region. In this section, we examine what the "equal fields at the boundary" condition implies. In section 3.3 it is shown that the time reversal approach can be used to achieve the "equal fields on the boundary" condition from well posed initial boundary value problems if the admittance operator commutes with the time-reversal operator and the uncontrollable part of the boundary is acoustically hard or soft.

The observation (measurement) of the response of a boundary control p_+ and an initial state $\{p_0(\boldsymbol{x}), \boldsymbol{v}_0(\boldsymbol{x})\}$ is expressed as

$$p_- = \mathcal{R} p_+ + \mathcal{O}_{\Gamma_t}(p_0, \boldsymbol{v}_0) \quad \text{for } \boldsymbol{x} \in \Gamma_t \text{ and } t \in [0, T], \tag{3.8}$$

where $\mathcal{O}_{\Gamma_t}$ is a linear map from $L^2(\Omega)$ to $L^2(\Gamma_t \times [0, T])$. The unobservable initial states, N_Ω, are defined as

$$N_\Omega = \{\{p_0, \boldsymbol{v}_0\} \in L^2(\Omega) : \mathcal{O}_{\Gamma_t}(p_0, \boldsymbol{v}_0) = 0 \in L^2(\Gamma_t \times [0, T])\}, \tag{3.9}$$

i.e., they form the null space of $\mathcal{O}_{\Gamma_t}$ and thus a closed linear subspace of L^2_Ω. We define the observable initial states as the orthogonal complement to N_Ω, and $N^\perp_\Omega$ is a closed linear subspace of L^2_Ω; hence $L^2_\Omega = N_\Omega \oplus N^\perp_\Omega$. Solving the control problem (2.2)–(2.4)

for some p_+, and its corresponding observation problem (2.2), (2.8), (2.10), and (2.11) for some q_+, gives the corresponding field $\{p, v_\mathrm{n}\}$ and $\{q, u_\mathrm{n}\}$ at the controllable part of the boundary Γ_t. If this field for some given p_+, q_+ satisfies (3.7), then from the superposition principle (see Figure 3.1) we note that the controllable and uncontrollable parts of the wave field coincide with the observable and unobservable parts. In the next section, we derive a sufficient condition on the transducer admittance $\mathcal{Y}$ so that the "equal fields on the boundary" condition is achievable.

Now, since L^2_Ω is a Hilbert space with the κ, ρ-weighted standard inner product, the least-squares functional (2.1) is minimized by projecting the desired state on the controllable subspace, $\mathrm{N}^\perp_\Omega$. This follows from the orthogonal projection theorem in Hilbert spaces; see, e.g., [25]. The final state $\{p_T - p'_T, \boldsymbol{v}_T - \boldsymbol{v}'_T\}$ is the controllable part of the original state $\{p_T, \boldsymbol{v}_T\}$. The error $\{p'_T, \boldsymbol{v}'_T\}$ is the uncontrollable (and unobserved) part of the original state. In other words with p_T, $\boldsymbol{v}_T$ and p'_T, $\boldsymbol{v}'_T$ defined as above, then

$$\int_\Omega (\kappa(\boldsymbol{x})(p_T(\boldsymbol{x}) - p'_T(\boldsymbol{x}))p'_T(\boldsymbol{x}) + \rho(\boldsymbol{x})(\boldsymbol{v}_T(\boldsymbol{x}) - \boldsymbol{v}'_T(\boldsymbol{x})) \cdot \boldsymbol{v}'_T(\boldsymbol{x}))\,\mathrm{dV}(\boldsymbol{x}) = 0. \tag{3.10}$$

To derive this, we use energy estimates of the three problems depicted in Figure 3.1. For an acoustically hard or soft uncontrollable part of the boundary, the energy balance of Figure 3.1(a) takes the form

$$E[p_T, \boldsymbol{v}_T] + \int_{-T}^{0} \int_{\Gamma_\mathrm{t}} q u_\mathrm{n} \mathrm{dS}\,\mathrm{d}t = E[q_0, \boldsymbol{u}_0], \tag{3.11}$$

while for Figure 3.1(b)

$$\int_0^T \int_{\Gamma_\mathrm{t}} p v_\mathrm{n} \mathrm{dS}\,\mathrm{d}t = E[p_T - p'_T, \boldsymbol{v}_T - \boldsymbol{v}'_T], \tag{3.12}$$

and for Figure 3.1(c), $E[q_0, \boldsymbol{u}_0] = E[p'_T, \boldsymbol{v}'_T]$. The "equal field on the boundary" condition gives $p v_\mathrm{n} = -\mathcal{T} q \mathcal{T} u_\mathrm{n}$. Combining the three energy balance equations gives the identity

$$E[p_T, \boldsymbol{v}_T] = E[p'_T, \boldsymbol{v}'_T] + E[p_T - p'_T, \boldsymbol{v}_T - \boldsymbol{v}'_T]. \tag{3.13}$$

Expansion of the right-hand side gives

$$\begin{aligned} &E[p'_T, \boldsymbol{v}'_T] + E[p_T - p'_T, \boldsymbol{v}_T - \boldsymbol{v}'_T] \\ &= E[p_T, \boldsymbol{v}_T] + 2\int_\Omega (\kappa(\boldsymbol{x})(p_T(\boldsymbol{x}) - p'_T(\boldsymbol{x}))p'_T(\boldsymbol{x}) + \rho(\boldsymbol{x})(\boldsymbol{v}_T(\boldsymbol{x}) - \boldsymbol{v}'_T(\boldsymbol{x})) \cdot \boldsymbol{v}'_T(\boldsymbol{x}))\,\mathrm{dV}(\boldsymbol{x}). \end{aligned} \tag{3.14}$$

Substitution of (3.14) into (3.13) and canceling $E[p_T, \boldsymbol{v}_T]$ gives the orthogonality condition (3.10).

Thus, if it is possible to generate "equal fields on the boundary," we are able to reconstruct the controllable part of the final state $\{p_T - p'_T, \boldsymbol{v}_T - \boldsymbol{v}'_T\}$.

3.3. Iterated time-reversal retrofocusing. In this section, a well-posed iterative algorithm is introduced to solve the condition (3.7) of "equal fields on the boundary" for a class of transducer admittances.

Using a standard energy argument in Figure 3.1(b), with either $p = \mathcal{T} q$ or $v_3 = -\mathcal{T} u_3$ for $x \in \Gamma_\mathrm{t}$ and either acoustically soft, $p = 0$, or acoustically hard, $v_3 = 0$, for $x \in \Gamma_\mathrm{w}$, the interior field is uniquely determined.

The "equal boundary fields condition" (3.7) can be rewritten in the observed quantities $\{p_-, q_-\}$ and the boundary controls $\{p_+, q_+\}$ using (2.4), (2.6), (2.9), and (2.10). We find

$$\begin{cases} \mathcal{Y}^{-1}p_+ + \mathcal{Y}^{-1}p_- = \mathcal{T}\mathcal{Y}^{-1}q_+ + \mathcal{T}\mathcal{Y}^{-1}q_-, \\ p_+ - p_- = -\mathcal{T}q_+ + \mathcal{T}q_-, \end{cases} \qquad \boldsymbol{x} \in \Gamma_{\mathrm{t}},\ t \in [0,T]. \tag{3.15}$$

In terms of the response operator (2.12) we note that, for a given boundary control q_+ and given initial observation $q_-^{(0)}$, we have $q_- = \mathcal{R}q_+ + q_-^{(0)}$.

If (3.15) admits a solution, elimination of q_- and p_- in (3.15) gives that p_+ and q_+ satisfy the system of equations

$$\begin{cases} p_+ = \frac{1}{2}(\mathcal{Y}\mathcal{T}\mathcal{Y}^{-1} + \mathcal{T})\mathcal{R}q_+ + \frac{1}{2}(\mathcal{Y}\mathcal{T}\mathcal{Y}^{-1} - \mathcal{T})q_+ + \frac{1}{2}(\mathcal{Y}\mathcal{T}\mathcal{Y}^{-1} + \mathcal{T})q_-^{(0)}, \\ q_+ = \frac{1}{2}(\mathcal{Y}\mathcal{T}\mathcal{Y}^{-1} + \mathcal{T})\mathcal{R}p_+ + \frac{1}{2}(\mathcal{Y}\mathcal{T}\mathcal{Y}^{-1} - \mathcal{T})p_+, \end{cases} \tag{3.16}$$

for $\boldsymbol{x} \in \Gamma_{\mathrm{t}}$ and $t \in [0,T]$. If the admittance commutes with the time reversal, i.e.,

$$\mathcal{Y}\mathcal{T} = \mathcal{T}\mathcal{Y}, \tag{3.17}$$

then the requirement (3.16) simplifies to the linear system

$$\begin{cases} p_+ = \mathcal{T}\mathcal{R}q_+ + \mathcal{T}q_-^{(0)}, \\ q_+ = \mathcal{T}\mathcal{R}p_+, \end{cases} \quad \text{or} \quad \begin{pmatrix} 1 & -\mathcal{T}\mathcal{R} \\ -\mathcal{T}\mathcal{R} & 1 \end{pmatrix} \begin{pmatrix} p_+ \\ q_+ \end{pmatrix} = \begin{pmatrix} \mathcal{T}q_-^{(0)} \\ 0 \end{pmatrix}. \tag{3.18}$$

This system can be solved in a variety of ways if the response operator $\mathcal{R}$ is known. In the case where only the action of $\mathcal{R}$ on an incident field is known, as in our case, the system is preferably solved by iterative methods. If $\mathcal{R}$ is sufficiently small, i.e., the spectral radius of $\mathcal{R}$ in L^2 is smaller than 1, an iterative scheme of the Jacobi type [29] converges. This gives the iterated time-reversal algorithm

$$\begin{cases} p_+^{(n)} = \mathcal{T}\mathcal{R}q_+^{(n-1)} = \mathcal{T}q_-^{(n-1)} \\ q_+^{(n)} = \mathcal{T}\mathcal{R}p_+^{(n)} = \mathcal{T}p_-^{(n)} \end{cases} \qquad \text{for } n = 1, 2, \ldots, \tag{3.19}$$

where $q_-^{(0)}$ is the initial measurement; see Figure 3.2 and section 2.2. If $\|\mathcal{R}q_+^{(n)}\|_{\mathrm{L}^2} = \|q_+^{(n)}\|_{\mathrm{L}^2}$ for any n, say n^*, then the algorithm has converged and can be terminated. In this case the input field does not give any contribution to the final field and would just repeat itself. The boundary control and the final state are given by

$$p_+ = \sum_{n=1}^{N} p_+^{(n)} = \sum_{n=1}^{N} (\mathcal{T}\mathcal{R})^{2n}\mathcal{T}q_-^{(0)} \quad \text{and} \quad \{p_T, \boldsymbol{v}_T\} = \sum_{n=1}^{N} \{p_T^{(n)}, \boldsymbol{v}_T^{(n)}\}, \tag{3.20}$$

respectively. Here $N = \min(n^*, \infty)$. The iterated time-reversal algorithm (3.19) is initiated by recording the output field, $q_-^{(0)}$, generated by the original state $\{p_T, -\boldsymbol{v}_T\}$. This recorded output field is time-reversed and re-emitted into the domain. Recording and time reversal of the corresponding output field iterates the algorithm.

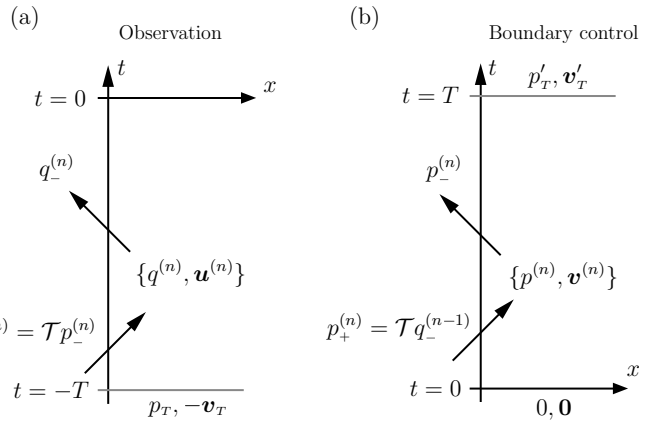


FIG. 3.2. *The iterated time-reversal algorithm.* (b) *an observation* $q_-^{(n-1)}$ *is time-reversed and used as a boundary control, to produce the output fields and the final state* $\{p_T^{(n)}, \boldsymbol{v}_T^{(n)}\}$ *and the output* $p_-^{(n)}$. (a) *the output,* $p_-^{(n)}$, *is recorded, time-reversed, and used as boundary control for the observation states, to produce the observation* $q_-^{(n)}$, *that, once again, is used to improve the final state.*

4. Example: Time-reversal retrofocusing. In this section, the theory of section 3 is applied to the problem of retrofocusing a wave field towards its initial state when both the initial state and the material parameters of the object are unknown. This problem has been thoroughly analyzed by Fink and coworkers; see, e.g., [10, 11, 12, 13]. In the time-reversal retrofocusing it is assumed that an initial state $\{q_T, \boldsymbol{u}_T\}$ exists in the object at time $t = -T$. This initial wave field is generated either by sources inside the object or by a field on the surface of the object. Since the material parameters and the original field distribution, in general, are unknown, the distribution of the retrofocused field is not known. However, it is known that the field retrofocuses towards its original field distribution.

The output $q_-^{(0)}$ is recorded at the boundary Γ_t for times $-T < t < 0$; see Figure 4.1(a). This initial observed field is time-reversed and re-emitted into the

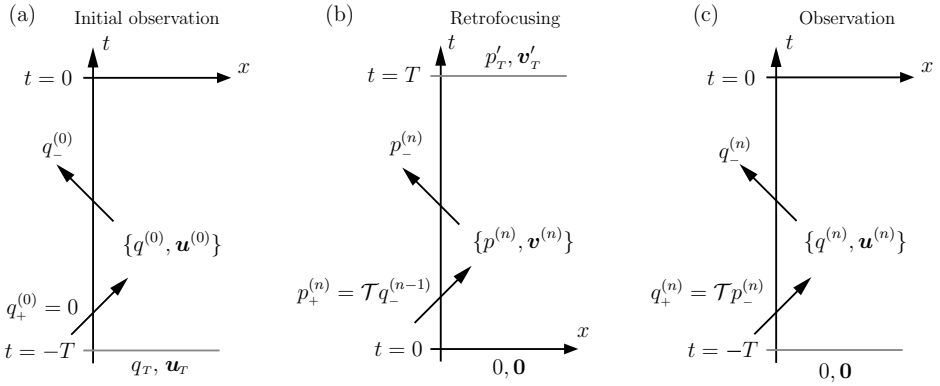


FIG. 4.1. *Iterated time-reversal retrofocusing.* (a) *The output field is recorded from the original field as the initial observation or the initial step of the algorithm* $n = 0$. (b) *in the retrofocusing, the recorded output* $q_-^{(n-1)}$ *is time-reversed and re-emitted into the domain to produce the final state* $\{p_T^{(n)}, -\boldsymbol{v}_T^{(n)}\}$ *and the output* $p_-^{(n)}$. (c) *in the observation, the output* $p_-^{(n)}$ *is time-reversed and re-emitted into the domain to give the output* $q_-^{(n)}$.

domain; see Figure 4.1(b). The retrofocusing is carried out with the iterative time-reversal algorithm (3.19). From section 3, it is concluded that the final state $\{p'_T, -\boldsymbol{v}'_T\}$ coincides with the controllable part of the initial state, $\{q_T, \boldsymbol{u}_T\}$. Observe that the retrofocusing does not require knowledge of the initial state nor of the material parameters. Moreover, if the initial state and the material parameters are unknown, the distribution of the retrofocused wave field is also unknown.

5. Example: Time-reversal cavity.

5.1. Acoustically hard boundary with ducts. In this section, the acoustically hard boundary cavity is considered, with transducers and receivers in the form of narrow ducts. The particular form of the transducers and receivers corresponds to a simple admittance operator. We present a numerical simulation in three dimensions of the retrofocusing and the resulting fields.

The cavity is a bounded region with a perforated acoustically hard boundary. The perforations are located in the end of a set of ducts. The wave field is induced through the perforations Γ_{t}; see Figure 5.1. If the ducts are sufficiently narrow and long, it is only the fundamental mode that propagates, i.e., a locally plane wave, propagating in the $\boldsymbol{n}$-direction; see [24, 26]. In this case the admittance operator reduces to a scalar constant,

$$\mathcal{Y} = \mathcal{Y}_0 = \sqrt{\kappa_0/\rho_0}. \tag{5.1}$$

The sufficient condition (3.17) of commutation between the admittance operator and the time-reversal operator is trivially satisfied. Moreover, the energy balance (3.3) gives $\|\mathcal{R}q_+^{(n)}\|_{\mathrm{L}^2} \leq \|q_+^{(n)}\|_{\mathrm{L}^2}$ for all n, and hence the algorithm (3.19) converges. (Here, the norm, $\|\cdot\|_{\mathrm{L}^2}$, is the L^2-norm over space and time.) The uncontrollable part of the boundary is acoustically hard, i.e., $\mathcal{U} = 0$. For this case the iterated time-reversal focusing algorithm (3.19) reduces to $p_+^{(n)} = \mathcal{T}q_-^{(n-1)}$ and $q_+^{(n)} = \mathcal{T}p_-^{(n)}$.

5.2. Numerical results. To illustrate the iterated time-reversal algorithm, a numerical example is presented. A cubic cavity with side length $L = 1$ and four horns

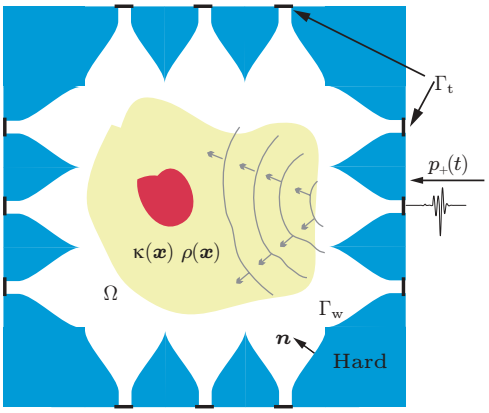


FIG. 5.1. *The cavity geometry. The boundary of the cavity is divided into the transducer surface, Γ_{t}, at the wave guide openings, and the acoustically hard wall, Γ_{w}. The transducers induce the boundary control, $p_+(\boldsymbol{x}, t)$, as a locally plane wave propagating in the $\boldsymbol{n}$-direction, where $\boldsymbol{n}$ is the inward unit normal.*

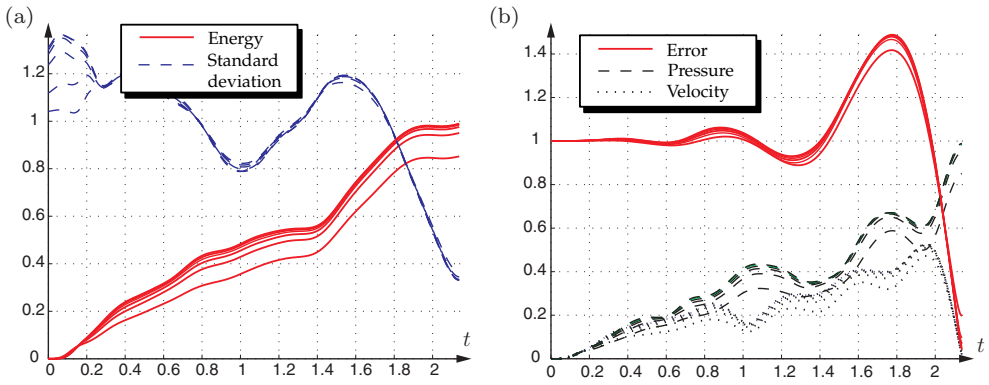


FIG. 5.2. *Example of a focused pressure in the cavity. Each improvement of the respective family of lines above is obtained by including one more term in the sum* (3.20). *The vertical axes show nondimensional values where the energies, the standard deviation, and the error are normalized with respect to the original energy, a uniform field distribution, and the error of quiescent fields, respectively.* (a) *The field energy is concentrated at* $T = 2.2\,\mathrm{s}$. *The standard deviation of the energy around its center point is minimized at the focusing time* T. (b) *The error term is* $\sqrt{\mathcal{J}}$, *where* $\mathcal{J}$ *is the least-square functional. The pressure part of the energy dominates the velocity part at the focusing time* T.

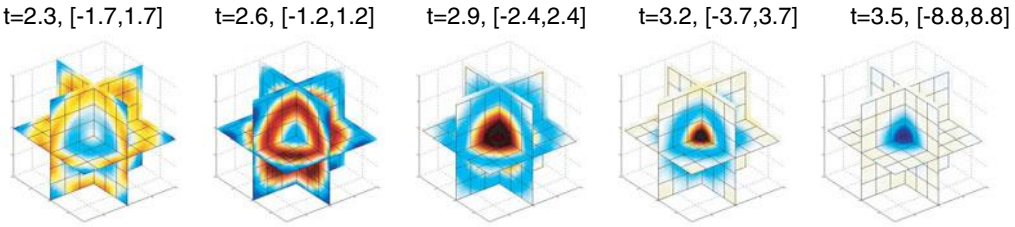


FIG. 5.3. *Example of a retrofocused pressure in the cavity; the range is given in brackets. The focusing time is* $T = 3.5$. *Note that the pressure amplitude range at* $t = 3.2$ *is considerably smaller than at* $t = 3.5$, (3.7 *versus* 8.8)*; hence the concentration of the field is highest at* $t = 3.5$. *Also note the extended region with "gray" at* $t = 3.2$.

attached to each side is considered. The acoustic wave equation is solved with a standard leap-frog scheme, where the cavity, the horns, and the ducts are discretized on an equidistant grid with discretization $L/82$. The fundamental mode is induced with a Huygens surface in the ducts, and the ducts are terminated with a perfectly matched layer. The temporal step is chosen according to the CFL condition [31] to minimize the numerical dispersion.

In Figure **??**(a), the first two moments of the energy distribution are depicted. It is obvious that the field energy is concentrated around the focusing time $T = 2.2$. At these times the energy is centered around the focusing point. The concentration of the energy is measured with the variation of the energy. The variation is scaled such that an energy with uniform distribution has unit variation. From the variation curve, it is clear that the wave field is concentrated around the focusing point at the focusing time but that the wave field is not concentrated for other times. The field energy of each field component is shown in Figure **??**(b). The retrofocused field for a case with retrofocusing at $T = 3.5$ is illustrated in Figure 5.3.

6. Example: Time-reversal mirror. For the time-reversal mirror, we consider the half space $\Omega = \{\boldsymbol{x} : x_3 \geq 0\}$, with homogeneous material parameters $\kappa = \kappa_0$ and $\rho = \rho_0$. The transducers are located at the plane $\Gamma_{\mathrm{t}} = \{\boldsymbol{x} \in \mathbb{R}^3 : \ x_3 = 0\}$. Thus we assume that the transducers can prescribe a boundary condition on the entire plane; cf. (2.4).

6.1. The impedance operator. A nonreflective admittance operator [16, 17, 32, 33] is given here as an explicit integral operator, as well as its adjoint with respect to the standard L^2-inner product.

The characteristics of the transducers and receivers are modeled by the wave splitting admittance operator [18], $\mathcal{Y}$, with symbol [30, 8]

$$\boldsymbol{y}(\tilde{\boldsymbol{\xi}}, s) = s^{-1}\sqrt{s^2\kappa_0\rho_0 + \tilde{\xi}^2}, \tag{6.1}$$

where s is the Laplace transform variable corresponding to time and $\tilde{\boldsymbol{\xi}} = (\xi_1, \xi_2)$ is the transverse Fourier variable corresponding to $\tilde{\boldsymbol{x}} = (x_1, x_2)$. We use the notation $\tilde{x} = |\tilde{\boldsymbol{x}}|$ and $\tilde{\xi} = |\tilde{\boldsymbol{\xi}}|$ to denote the norms of $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{\xi}}$, respectively. Let $\mathcal{L}$ denote the temporal Laplace transform and $\mathcal{F}$ denote the spatial transverse Fourier transform; the $\mathcal{Y}$ acting of the pressure can be expressed in terms of the symbol as

$$(\mathcal{Y}p)(x, v) = \mathcal{L}^{-1}\mathcal{F}^{-1}(y(\tilde{\boldsymbol{\xi}}, s)p(\tilde{\boldsymbol{\xi}}, x_3, s)), \tag{6.2}$$

where we have used $p(\tilde{\boldsymbol{\xi}}, x_3, s)$ to denote the Laplace and Fourier transforms of $p(x, t)$.

The above square root is taken with the branch-cut at the negative real axis, i.e., $\sqrt{s^2} = s$, when $\operatorname{Re} s > 0$. An energy renormalization of the field removes the constant material parameters in the acoustic wave equation (2.2) [14, p. 37] and consequently in (6.1). The inverse of the admittance, the impedance, satisfies the transform relation

$$\frac{s}{\sqrt{s^2 + \tilde{\xi}^2}} = \mathcal{F}\mathcal{L}\frac{\delta'(t - |\tilde{\boldsymbol{x}}|)}{2\pi|\tilde{\boldsymbol{x}}|} \tag{6.3}$$

for $t \geq 0$, which follows from a straightforward calculation. Thus the time-space representation of the impedance operator action is

$$(\mathcal{Y}^{-1}p_+)(\tilde{\boldsymbol{x}}, t) = \int_{\mathbb{R}^2}\int_0^t \frac{\delta'(t - t' - |\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}'|)}{2\pi|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}'|} p_+(\tilde{\boldsymbol{x}}', t')\,\mathrm{d}t'\,\mathrm{d}x_1'\,\mathrm{d}x_2' \tag{6.4}$$

for sufficiently smooth controls, p_+. Here, $\boldsymbol{x} \in \Gamma_{\mathrm{t}}$, $t \in [0, T]$.

In section 6 and the appendices, we use Fourier and Laplace transforms. To utilize their properties, we consider the case that all fields, p, $\boldsymbol{v}$ and q, $\boldsymbol{u}$, have temporal support contained in $[0, T]$. Consequently, the time-reversal operator is redefined as

$$(\mathcal{T}p_+)(\boldsymbol{x}, t) = p_+(\boldsymbol{x}, T - t), \quad \boldsymbol{x} \in \Omega, \ t \in [0, T]. \tag{6.5}$$

For section 6 and forward we use (6.5) instead of (3.1). As the system is linear and independent of starting time, the change of definition of $\mathcal{T}$ is only a matter of shifting the solution with respect to time.

$\mathcal{T}$ does not commute with an admittance $\mathcal{Y}$ of the form (6.1). Indeed, observe that the adjoint $\mathcal{Y}^*$ of $\mathcal{Y}$ with respect to the standard L^2-inner product over time and space at the boundary is

$$((\mathcal{Y}^*)^{-1}p_+)(\tilde{\boldsymbol{x}}, t) = \int_{\mathbb{R}^2}\int_t^T \frac{\delta'(t' - t - |\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}'|)}{2\pi|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}'|} p_+(\tilde{\boldsymbol{x}}', t')\,\mathrm{d}t'\,\mathrm{d}x_1'\,\mathrm{d}x_2'. \tag{6.6}$$

It follows that $\mathcal{Y}$ is not self-adjoint. By the variable substitution $\tau = T - t'$, we obtain

$$\begin{aligned}((\mathcal{Y}^*)^{-1}p_+)(\tilde{\boldsymbol{x}},t) &= \mathcal{T}\int_{\mathbb{R}^2}\int_0^t \frac{\delta'(t-\tau-|\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{x}}'|)}{2\pi|\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{x}}'|}(\mathcal{T}p_+)(\tilde{\boldsymbol{x}}',\tau)\,\mathrm{d}\tau\,\mathrm{d}x_1'\,\mathrm{d}x_2' \\ &= (\mathcal{T}\mathcal{Y}^{-1}\mathcal{T}p_+)(\tilde{\boldsymbol{x}},t).\end{aligned} \tag{6.7}$$

6.2. Nonsolvability of "equal fields on the boundary." In this section, we show that for the time-reversal mirror configuration with boundary condition (2.4) and admittance operator (6.4), the given algorithm does not have an optimal boundary control. We derive an approximate boundary control that agrees with the equation for optimal boundary control in the nonevanescent part of the measured field.

For the homogeneous half space, with transducer characteristics $\mathcal{Y}^{-1}$ as in (6.4), the boundary condition (2.4) is equivalent to a splitting of the field into an in- and an outgoing constituent at the boundary [14, 15, 16, 17]. Thus a transducer with the (6.4) characteristics is perfectly matched to the domain and does not introduce any transducer mismatch reflections. Hence, as the medium is homogeneous, the response operator vanishes; i.e., $\mathcal{R}p_+ = 0$, likewise $\mathcal{R}q_+ = 0$.

If we consider the solvability of the requirement of "equal fields on the boundary," with (6.7) the equations (3.15) reduce to

$$(\mathcal{Y}^{-1} + (\mathcal{Y}^*)^{-1})p_+ = 2(\mathcal{Y}^*)^{-1}\mathcal{T}q_-^{(0)}. \tag{6.8}$$

The operator $(\mathcal{Y}^{-1} + (\mathcal{Y}^*)^{-1})$ is not invertible everywhere on the range of $(\mathcal{Y}^*)^{-1}$. Thus the requirement of "equal fields on the boundary" is too strong a condition for this admittance. To analyze this situation we rewrite the least-squares functional, $\mathcal{J}$, with the definitions of the boundary control and its observations, as

$$\mathcal{J} = E_0 + \int_0^T \int_{\Gamma_t} (\mathcal{Y}^{-1}p_+ - \mathcal{T}\mathcal{Y}^{-1}(q_+ + q_-^{(0)}))(p_+ + \mathcal{T}(q_+ - q_-^{(0)}))\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}t. \tag{6.9}$$

The least-squares functional $\mathcal{J}$ does not have any critical points for this choice of admittance. To see this, we denote the above integral over time and boundary by the inner product $\langle\cdot,\cdot\rangle_{\Gamma_t\times[0,T]}$, and we observe that the field is real-valued and that the operator $\mathcal{Y}^{-1}$ maps real-valued functions into real-valued functions. Upon determining the variation with respect to p_+, we find the requirement for critical points to be

$$\begin{aligned}\langle \mathcal{D}\mathcal{J}, \delta p_+\rangle_{\Gamma_t\times[0,T]} &= 2\,\mathrm{Re}\langle(\mathcal{Y}^{-1} + (\mathcal{Y}^*)^{-1})p_+ - (\mathcal{T}\mathcal{Y}^{-1} + (\mathcal{Y}^*)^{-1}\mathcal{T})q_-^{(0)} \\ &\quad + ((\mathcal{Y}^*)^{-1}\mathcal{T} - \mathcal{T}\mathcal{Y}^{-1})q_+, \delta p_+\rangle_{\Gamma_t\times[0,T]} = 0\end{aligned} \tag{6.10}$$

for all $\delta p_+ \in \mathrm{L}^2$. Using property (6.7), we find that (6.10) simplifies to

$$\langle(\mathcal{Y}^{-1} + (\mathcal{Y}^*)^{-1})p_+, \delta p_+\rangle_{\Gamma_t\times[0,T]} = \langle 2(\mathcal{Y}^*)^{-1}\mathcal{T}q_-^{(0)}, \delta p_+\rangle_{\Gamma_t\times[0,T]} \tag{6.11}$$

for all $\delta p_+ \in \mathrm{L}^2$. This equation is equivalent to (6.8), and to confirm its nonsolvability for general $q_-^{(0)}$, we apply Parseval's formula to (6.11) and obtain

$$\mathrm{H}(\omega^2 - \tilde{\xi}^2)\frac{|\omega|}{\sqrt{\omega^2-\tilde{\xi}^2}}p_+(\tilde{\boldsymbol{\xi}},\omega) = \lim_{\eta\to 0^+}\frac{(\eta - \mathrm{i}\omega)}{\sqrt{(\eta-\mathrm{i}\omega)^2+\tilde{\xi}^2}}\mathrm{e}^{-\mathrm{i}\omega T}q_-^{(0)}(\tilde{\xi},-\omega), \tag{6.12}$$

where $\mathrm{H}(\cdot)$ is the Heaviside (step) function. It is obvious that (6.12) does not have a solution for all $q_-^{(0)}$, in particular for $|\omega| < \tilde{\xi}$, since the left-hand side is zero while the right-hand side can be nonzero, depending on $q_-^{(0)}$. In the Fourier domain for the corresponding Green's function the frequency region $|\omega| < \tilde{\xi}$ is the nonpropagating part of the field; hence we denote the part where $q_-^{(0)} \neq 0$ for $|\omega| < \tilde{\xi}$ as *the evanescent part* of $q_-^{(0)}$ (cf., e.g., [26]). Thus the "equal fields on the boundary condition" is not applicable in the time-reversal mirror, and thus the algorithm does not yield an optimum for this case.

We conclude that for general $q_-^{(0)} \in \mathrm{L}^2$ the least-squares functional, $\mathcal{J}$, does not have a critical point in terms of the field at the boundary. However, if $q_-^{(0)}$ does not have an evanescent part, then $\mathcal{J}$ has a critical point and the corresponding optimal control is

$$p_+ = \mathcal{T} q_-^{(0)}\big|_{|\omega| \geq \tilde{\xi}}. \tag{6.13}$$

6.3. Approximate boundary controls. As shown in section 6.2 it is only possible to satisfy the equal fields on the boundary condition for the propagating part of the wave field. Hence, it is not clear how to choose the control in the nonpropagating or evanescent part of the wave field. Here, we consider three different controls, labeled $p_+^{(1)}$, $p_+^{(2)}$, and $p_+^{(3)}$, that satisfy (6.11) in the propagating regime.

Case one. $p_+^{(1)}$ is the particle-velocity normalized control; i.e., observation of the initial state $q_-^{(0)}$ is measured in particle-velocity normalization, time-reversed, and retransmitted, viz.,

$$p_+^{(1)} = \mathcal{T} q_-^{(0)}. \tag{6.14}$$

Case two. To construct $p_+^{(2)}$, we begin with a consideration of the control problem in pressure normalization. The boundary condition and the measurement take the form

$$\frac{p + \mathcal{Y}^{-1} v_3}{2} = p_{+,\mathrm{N_p}}^{(2)} \quad \text{and} \quad m_{(2)} = \frac{p - \mathcal{Y}^{-1} v_3}{2}. \tag{6.15}$$

Now consider a measurement of the initial pressure pulse in this normalization $m_{(2)}$, time-reversed and retransmitted, viz., $p_{+,\mathrm{N_p}}^{(2)} = \mathcal{T} m_{(2)}$. The result from this control is the second case. To express this for particle-velocity normalization we utilize the linearity of the problem to obtain that the control $p_+^{(2)}$ in the form

$$p_+^{(2)} = \mathcal{Y} \mathcal{T} \mathcal{Y}^{-1} q_-^{(0)} \tag{6.16}$$

yields the same internal field as $p_{+,\mathrm{N_p}}^{(2)}$, but with the particle-velocity boundary conditions.[1] Note that (6.7) gives the relations $\mathcal{Y}\mathcal{T}\mathcal{Y}^{-1} = \mathcal{T}\mathcal{Y}^{*}\mathcal{Y}^{-1} = \mathcal{Y}(\mathcal{Y}^{-1})^{*}\mathcal{T}$.

Case three. The control $p_+^{(3)}$ is the linear combination of the first two cases, i.e.,

$$p_+^{(3)} = p_+^{(1)} + p_+^{(2)} = (1 + \mathcal{Y}(\mathcal{Y}^{-1})^{*})\mathcal{T} q_-^{(0)}. \tag{6.17}$$

Note from (6.12) that $p_+^{(3)}$ cuts out the nonpropagating part for the $q_-^{(0)}$. The results of Case three are discussed in section 6.7.

[1] The fields inside the domain from the controls $p_+^{(2)}$ and $p_{+,\mathrm{N_p}}^{(2)}$ with the respective normalized boundary conditions are the same. This is utilized to explicitly calculate the response of $p_+^{(2)}$; see Appendix C.

6.4. Control operators. In this section, we derive the control operators for the respective normalizations of p_+, for boundary data with temporal duration T, i.e., an operator that takes the boundary control to the respective final states at $t = T$.

The control operator in particle velocity-normalization is defined as

$$\{p^{(1)}(\cdot, T),\ \boldsymbol{v}^{(1)}(\cdot, T)\} = \mathcal{W}_{\mathrm{v}} p_+^{(1)}. \tag{6.18}$$

To derive the explicit form of $\mathcal{W}_{\mathrm{v}}$, we solve the acoustic wave equation, (2.2), and (2.3), together with the transducer boundary condition (2.4) and quiescent initial conditions. We obtain (see Appendix A.1)

$$p^{(1)}(\boldsymbol{x}, T) = \int_0^T \int_{\mathbb{R}^2} \frac{\delta(T - t' - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{2\pi|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|} \partial_{t'}(\mathrm{H}(t') p_+^{(1)}(\tilde{\boldsymbol{x}}', t'))\, \mathrm{d}x_1'\, \mathrm{d}x_2'\, \mathrm{d}t' \tag{6.19}$$

and

$$\boldsymbol{v}^{(1)}(\boldsymbol{x}, T) = -\nabla \int_0^T \int_{\mathbb{R}^2} \frac{\delta(T - t' - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{2\pi|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|} p_+^{(1)}(\tilde{\boldsymbol{x}}', t')\, \mathrm{d}x_1'\, \mathrm{d}x_2'\, \mathrm{d}t'. \tag{6.20}$$

The control operator in the pressure normalization, $\mathcal{W}_{\mathrm{p}}$, is obtained by solving the acoustic wave equation with transducer boundary condition (6.15) and quiescent initial conditions. We obtain (see Appendix C)

$$p^{(2)}(\boldsymbol{x}, T) = -\partial_3 \int_0^T \int_{\mathbb{R}^2} \frac{\delta(T - t' - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{2\pi|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|} p_{+,\mathrm{N_p}}^{(2)}(\tilde{\boldsymbol{x}}', t')\, \mathrm{d}x_1'\, \mathrm{d}x_2'\, \mathrm{d}t' \tag{6.21}$$

and

$$\boldsymbol{v}^{(2)}(\boldsymbol{x}, T) = \nabla \partial_3 \int_0^T \int_{\mathbb{R}^2} \frac{\delta(T - t' - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{2\pi|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|} \int_0^{t'} p_{+,\mathrm{N_p}}^{(2)}(\tilde{\boldsymbol{x}}', t'')\, \mathrm{d}t''\, \mathrm{d}x_1'\, \mathrm{d}x_2'\, \mathrm{d}t'. \tag{6.22}$$

Thus we have an explicit expression for the control operator in both normalizations, expressed in terms of a common Green's function. Note that the form of the control operators has the typical retarded time dependence that is associated with hyperbolic systems.

6.5. The solution operator. In this section, we use the solution operator with initial conditions corresponding to a pressure pulse to derive the field at the boundary. We also construct the boundary controls for the two cases (6.14) and (6.16).

To construct the boundary controls, $p_+^{(1)}$, $p_{+,\mathrm{N_p}}^{(2)}$, corresponding to an initial pressure pulse, the time-reversed output field component is needed as data. It is obtained by solving the acoustic equations (2.2) with the initial value $\{p_T, -\boldsymbol{v}_T\}$ at $t = 0$. The solution is (see [21])

$$\begin{pmatrix} q(\boldsymbol{x}, t) \\ \boldsymbol{u}(\boldsymbol{x}, t) \end{pmatrix} = \begin{pmatrix} \partial_t & -\nabla\cdot \\ -\nabla & I\partial_t \end{pmatrix} \int_{\mathbb{R}^3} \frac{\delta(t - |\boldsymbol{x} - \boldsymbol{x}'|)}{4\pi|\boldsymbol{x} - \boldsymbol{x}'|} \begin{pmatrix} p_T(\boldsymbol{x}') \\ -\boldsymbol{v}_T(\boldsymbol{x}') \end{pmatrix} \mathrm{d}x_1'\, \mathrm{d}x_2'\, \mathrm{d}x_3', \tag{6.23}$$

for an irrotational initial velocity, i.e., $\nabla \times \boldsymbol{v}_T = 0$. Now, to generate the output field component at the boundary $\Gamma_{\mathrm{t}} = \{\boldsymbol{x} \in \mathbb{R}^3 : x_3 = 0\}$, we assume that $\operatorname{supp} p_T$ and $\operatorname{supp} \boldsymbol{v}_T$ are bounded and contained in the half space $x_3 > 0$. Furthermore, we impose a transparent boundary condition at Γ_{t}, i.e., no reflection at the boundary $x_3 = 0$.

For the measurement of the outgoing wave (2.9) this is equivalent to measurement with perfectly matched receivers of the solution to (2.2) at the boundary.

To obtain an explicit representation of the field, we let the initial field be a pulse in the pressure, with source point $\tilde{\boldsymbol{x}} = 0$, $x_3 = z_0 > 0$, that is,

$$p_T(\boldsymbol{x}) = \delta(\tilde{\boldsymbol{x}})\,\delta(x_3 - z_0) \quad \text{and} \quad \boldsymbol{v}_T = 0. \tag{6.24}$$

The choice of pulse (6.24) substituted into (6.23) makes the field $\{q, \boldsymbol{u}\}$ into the components of the pressure Green's function. At the boundary, its $\{q, u_3\}$-component becomes

$$\begin{aligned}\begin{pmatrix} q(\tilde{\boldsymbol{x}},0,t) \\ u_3(\tilde{\boldsymbol{x}},0,t)\end{pmatrix} &= \begin{pmatrix} \partial_t \\ -\partial_3\end{pmatrix} \int_{\mathbb{R}^3} \frac{\delta(t-|\boldsymbol{x}-\boldsymbol{x}'|)}{4\pi|\boldsymbol{x}-\boldsymbol{x}'|}\,\delta(\tilde{\boldsymbol{x}}')\,\delta(x_3'-z_0)\,\mathrm{d}x_1'\,\mathrm{d}x_2'\,\mathrm{d}x_3'\Big|_{x_3=0} \\ &= \begin{pmatrix} \partial_t \\ \partial_{z_0}\end{pmatrix} \frac{\delta(t-\sqrt{\tilde{x}^2+z_0^2})}{4\pi\sqrt{\tilde{x}^2+z_0^2}}.\end{aligned} \tag{6.25}$$

This field is measured by transducers yielding $q_-^{(0)}$ and $m_{(2)}$, respectively; see (2.6) and (6.15).

The transducer is perfectly matched to the domain, equivalent to a transparent boundary condition, and measures the velocity normalized out-going component of the wave; cf. [18]. As the receiver and transducer characteristics are identical, the measured response $\{q_-^{(0)}, m_{(2)}\}$ in the respective normalizations, Cases one and two, is (cf. (2.6), (6.15))

$$q_-^{(0)}(\tilde{\boldsymbol{x}},t) = \frac{1}{2}((\mathcal{Y}q)(\tilde{\boldsymbol{x}},0,t) - u_3(\tilde{\boldsymbol{x}},0,t)), \tag{6.26}$$

$$m_{(2)}(\tilde{\boldsymbol{x}},t) = \frac{1}{2}(q(\tilde{\boldsymbol{x}},0,t) - (\mathcal{Y}^{-1}u_3)(\tilde{\boldsymbol{x}},0,t)). \tag{6.27}$$

Substituting the expressions for $\{q, u_3\}$ (cf. (6.25)) into (6.26) and (6.27) gives (see Appendix A.2 and Appendix C)

$$q_-^{(0)}(\tilde{\boldsymbol{x}},t) = -\partial_{z_0}\frac{\delta(t-\sqrt{\tilde{x}^2+z_0^2})}{4\pi\sqrt{\tilde{x}^2+z_0^2}}, \tag{6.28}$$

$$m_{(2)}(\tilde{\boldsymbol{x}},t) = \partial_t\frac{\delta(t-\sqrt{\tilde{x}^2+z_0^2})}{4\pi\sqrt{\tilde{x}^2+z_0^2}}. \tag{6.29}$$

The measurements, $\{q_-^{(0)}, m_{(2)}\}$, of the field at the surface start at time $t = 0$. We notice the expected delayed arrival in the measurement, because the initial pulse is at depth $x_3 = z_0$. The "measurement" ends at $t = T$ and, in general, the field at this time is nonzero. Hence, to describe the measured field, we have to introduce a step function that removes the field after $t = T$. Now, time reversal in accordance with (6.5) of $q_-^{(0)}$ and $m_{(2)}$ are the controls that we search for,

$$p_+^{(1)}(\tilde{\boldsymbol{x}},t) = \mathrm{H}(t)q_-^{(0)}(\tilde{\boldsymbol{x}},T-t) = -\mathrm{H}(t)\partial_{z_0}\frac{\delta(T-t-\sqrt{\tilde{x}^2+z_0^2})}{4\pi\sqrt{\tilde{x}^2+z_0^2}}, \tag{6.30}$$

$$p_{+,\mathrm{N_p}}^{(2)}(\tilde{\boldsymbol{x}},t) = \mathrm{H}(t)m_{(2)}(\tilde{\boldsymbol{x}},T-t) = -\mathrm{H}(t)\partial_t\frac{\delta(T-t-\sqrt{\tilde{x}^2+z_0^2})}{4\pi\sqrt{\tilde{x}^2+z_0^2}}. \tag{6.31}$$

6.6. Retrofocused fields and their properties. Here, the explicit form of the retrofocused field in Case one is used to analyze the long time limit of the retrofocused field. In the long time limit, Case one retrofocuses the pressure perfectly, modulo a numerical factor, whereas the particle velocity shows a nonzero remainder and is hence not perfectly retrofocused. We also give the explicit form of the pressure for Case two. We also show a number of graphs describing the degree of retrofocusing versus time duration of the measured data.

For the boundary control (6.30) we can explicitly obtain the final state as distributions, denoted by $\{p^{(1)}(\cdot,T),\boldsymbol{v}^{(1)}(\cdot,T)\}$, for any finite time T. To this end, substitute the control $p_+^{(1)}$ of (6.30) into (6.19) and (6.20). Integration with respect to time gives the response,

(6.32)
$$\begin{aligned} p^{(1)}(\boldsymbol{x},T) &= -\partial_{z_0}\partial_t \int_{\mathbb{R}^2} \mathrm{H}(t-|\boldsymbol{x}-\tilde{\boldsymbol{x}}'|)\frac{\delta(T-t+|\boldsymbol{x}-\tilde{\boldsymbol{x}}'|-\sqrt{(\tilde{x}')^2+z_0^2})}{8\pi^2|\boldsymbol{x}-\tilde{\boldsymbol{x}}'|\sqrt{(\tilde{x}')^2+z_0^2}}\,\mathrm{d}x_1'\,\mathrm{d}x_2'\Bigg|_{t=T} \\ &= \frac{-z_0}{4\pi^2 x_3\tilde{x}^2}\partial_3 \frac{\tilde{x}^2-x_3^2+z_0^2}{([T^2-x_3^2-(\tilde{x}-\sqrt{T^2-z_0^2})^2][(\tilde{x}+\sqrt{T^2-z_0^2})^2+x_3^2-T^2])_+^{1/2}} \end{aligned}$$

and

(6.33)
$$\begin{aligned} \boldsymbol{v}^{(1)}(\boldsymbol{x},T) &= \nabla\partial_{z_0}\int_{\mathbb{R}^2} \mathrm{H}(t-|\boldsymbol{x}-\tilde{\boldsymbol{x}}'|)\frac{\delta(T-t+|\boldsymbol{x}-\tilde{\boldsymbol{x}}'|-\sqrt{(\tilde{x}')^2+z_0^2})}{8\pi^2|\boldsymbol{x}-\tilde{\boldsymbol{x}}'|\sqrt{(\tilde{x}')^2+z_0^2}}\,\mathrm{d}x_1'\,\mathrm{d}x_2'\Bigg|_{t=T} \\ &= \nabla\Bigg[\frac{-z_0T}{\pi^2(4\tilde{x}^2z_0^2+(\tilde{x}^2+x_3^2-z_0^2)^2)} \\ &\qquad \frac{\tilde{x}^2-x_3^2+z_0^2}{([T^2-x_3^2-(\tilde{x}-\sqrt{T^2-z_0^2})^2][(\tilde{x}+\sqrt{T^2-z_0^2})^2+x_3^2-T^2])_+^{1/2}}\Bigg]. \end{aligned}$$

For details we refer to Appendix B.

The analogous derivation for Case two follows the substitution of the boundary control $p_{+,\mathrm{N_p}}^{(2)}$ of (6.31) into (6.21) and (6.22). Upon integration (see Appendix C), we find that

(6.34)
$$\begin{aligned} p^{(2)}(\boldsymbol{x},T) &= \frac{-1}{4\pi^2\tilde{x}^2}\partial_3 \frac{\tilde{x}^2-x_3^2+z_0^2}{([T^2-x_3^2-(\tilde{x}-\sqrt{T^2-z_0^2})^2][(\tilde{x}+\sqrt{T^2-z_0^2})^2+x_3^2-T^2])_+^{1/2}} \end{aligned}$$

and

(6.35)
$$\begin{aligned} \boldsymbol{v}^{(2)}(\boldsymbol{x},T) = -\nabla\partial_3\int_{\mathbb{R}^2}\Bigg(&\mathrm{H}\left(T-\sqrt{(\tilde{x}')^2+z_0^2}\right)\frac{\delta(\sqrt{(\tilde{x}')^2+z_0^2}-|\boldsymbol{x}-\tilde{\boldsymbol{x}}'|)}{8\pi^2|\boldsymbol{x}-\tilde{\boldsymbol{x}}'|\sqrt{(\tilde{x}')^2+z_0^2}} \\ &-\delta\left(T-\sqrt{(\tilde{x}')^2+z_0^2}\right)\frac{\mathrm{H}(\sqrt{(\tilde{x}')^2+z_0^2}-|\boldsymbol{x}-\tilde{\boldsymbol{x}}'|)}{8\pi^2|\boldsymbol{x}-\tilde{\boldsymbol{x}}'|\sqrt{(\tilde{x}')^2+z_0^2}}\Bigg)\mathrm{d}x_1'\,\mathrm{d}x_2'. \end{aligned}$$

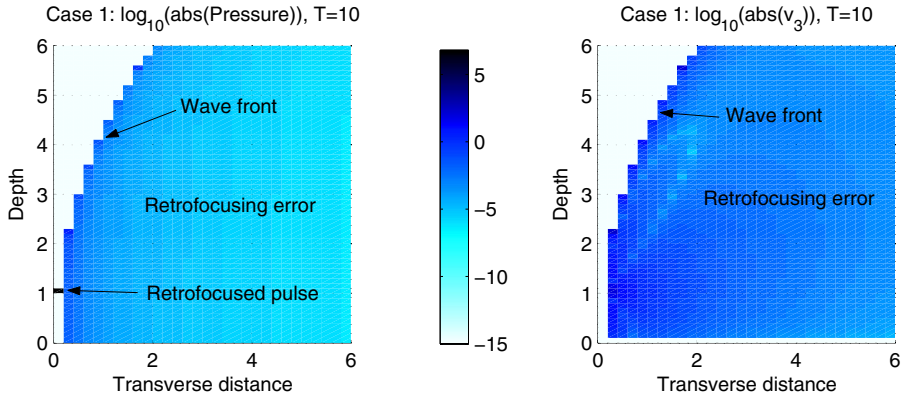


FIG. 6.1. *The retrofocused $p^{(1)}$ and $v_3^{(1)}$, for $T = 10$, with initial state a Dirac pulse at $z_0 = 1$, $\tilde{\boldsymbol{x}} = 0$, and its boundary control $p_+^{(1)}$. The pressure to high degree, concentrated to $x_3 = 0$, $z_0 = 1$, with a small retrofocus error, and the velocity have no corresponding concentration; cf.* (6.36). *The graphs show that the boundary control, $p_+^{(1)}$, is imposed at depth $x_3 = 0$ and has a support at $x_3 = 0$ and $\tilde{x} \leq \sqrt{T^2 - z_0^2}$. In the graph, notice the out-going wave front with radius $T = 10$ with center located at all $\tilde{\boldsymbol{x}}$ such that $|\tilde{\boldsymbol{x}}| = \sqrt{T^2 - z_0^2}$, here at $\{\tilde{x}, x_3\} = \{\sqrt{99}, 0\}$, due to the finite time cut-off of the measurement. Recall that the graphs are distributions; hence the graphs are smoothed around the wave front set.*

The difference in pressure between the pressure normalization and the particle-velocity normalization is the factor z_0/x_3. Furthermore, observe that v_3 is zero on the x_3-axis for $x_3 \neq z_0$. The integration of the first term in (6.35) is analogous to the normal particle-velocity normalization; cf. (B.21).

The first observation on the above final state is that $p^{(1)}$, $p^{(2)}$, and $v_3^{(1)}$ depend only on $\tilde{x}$ and not on $\tilde{\boldsymbol{x}}$; i.e., they are independent of polar angle—the angle between x_1 and x_2. The denominator $(\dots)_+^{1/2}$ describes wave fronts induced from the nonzero field at the end of a finite measurement time. These wave fronts are centered on the circle $x_3 = 0$, $\{\tilde{\boldsymbol{x}} : \tilde{x} = \sqrt{T^2 - z_0^2}\}$. The cut-off in $(\dots)_+^{1/2}$, with the polar angle symmetry, results in a field having a domain shape that resembles a donut, cut horizontally just below the middle. A cross section of the field is shown in Figure 6.1 for $T = 10$; i.e., the excitation time equals ten times the time it takes for the initial pulse to reach the surface.

With the retrofocusing of this pulse, note that for an expansion as $T \to \infty$,

$$p^{(1)} = \frac{z_0}{4\pi^2}(T^{-1}\tilde{x}^{-3} + (2T)^{-3}(\tilde{x}^{-1} + 2\tilde{x}^{-3}(x_3^2 + z_0^2) + 3\tilde{x}^{-5}(x_3^2 - z_0^2))) + \mathrm{O}(T^{-5})$$

and

$$v_3^{(1)} = \frac{x_3 z_0((x_3^2 - z_0^2)^2 - 3\tilde{x}^4 - 2\tilde{x}^2(x_3^2 + z_0^2))}{\pi^2 \tilde{x}(\tilde{x}^4 + (x_3^2 - z_0^2)^2 + 2\tilde{x}^2(x_3^2 + z_0^2))^2} + \mathrm{O}(T^{-2}), \tag{6.36}$$

together with

$$\begin{pmatrix} v_1^{(1)} \\ v_2^{(1)} \end{pmatrix} = \left(\frac{z_0(3\tilde{x}^2 + 4z_0^2)}{2\pi^2 \tilde{x}^2(\tilde{x}^2 + 4z_0^2)} \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} + \mathrm{O}((x_3 - z_0)^2) \right) + \mathrm{O}(T^{-2}), \tag{6.37}$$

where θ is the polar angle. Furthermore, at the axis $\tilde{x} = 0$ the retrofocused field is supported only at $x_3 = z_0$. In Figure 6.2 we plot the pressure away from the singular

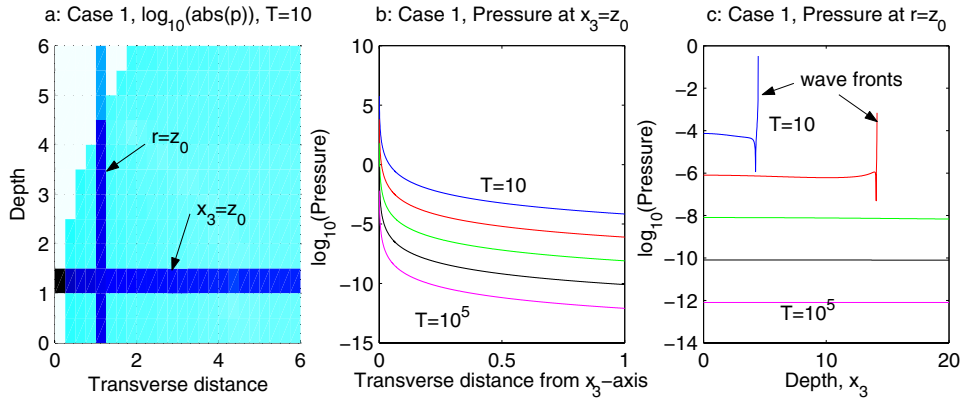


FIG. 6.2. *Plots of the pressure for different excitation times T. In graph* (b) *and* (c) *each lower line corresponds to the T values* 10, 10^2, 10^3, 10^4, 10^5. (a) *The black lines indicates where the values in graphs* (b) *and* (c) *are taken.* (b) *At $x_3 = z_0$ we see the damping to the Dirac pulse, for the area around the singular point.* (c) *At $r := |\tilde{\boldsymbol{x}}| = z_0$ we see that the fields damp as T increases; the endings of the two top lines result from reaching the wave front; cf. Figure* 6.2(a).

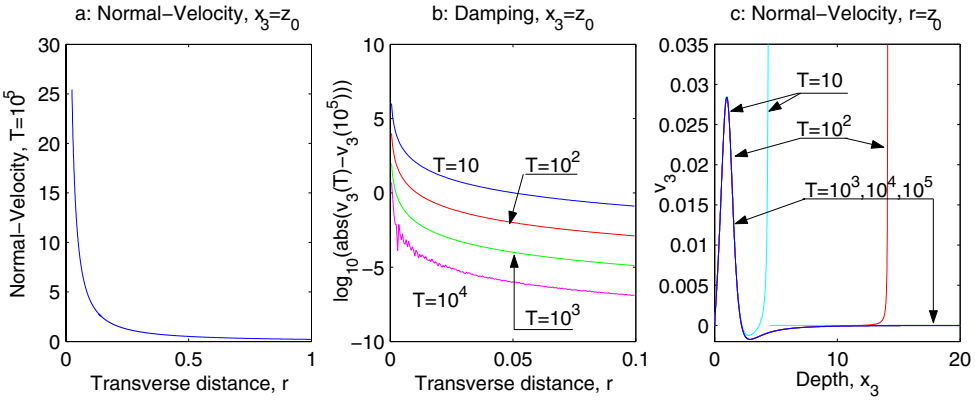


FIG. 6.3. *Plots of $v_3^{(1)}$ for a source at depth $z_0 = 1$ and its boundary control $p_+^{(1)}$.* (a) *The "remaining" field as T becomes large; cf.* (6.36). (b) *Plots $\log_{10}(|v_3^{(1)}(T) - v_3^{(1)}(t = 10^5)|)$ for $T = 10$, 10^2, 10^3, 10^4, along $x_3 = z_0$. In the lowest line the unevenness is due to numerical inaccuracies.* (c) *Plot of $v_3^{(1)}$ along $r := |\tilde{\boldsymbol{x}}| = z_0$ for times $T = 10$ to -20^5. All lines start along the same "remaining" velocity at $x_3 = 0$. The first to deviate is $v_3^{(1)}(T = 10)$, which encounters its wave front at $x_3 \approx 4$ (cf. Figure* 6.1(b)*); the next to go off to infinity is $T = 10^2$, which encounters its wave front at $x_3 \approx 15$.*

point for different excitation times T. It is apparent that the field is rapidly damped with respect to T and $\tilde{x}$. The analogous plots for v_3 are shown in Figure 6.3, where it is apparent that v_3 rapidly approaches its remaining, nonzero, distribution.

In the limit $T \to \infty$ we obtain for the pressure (see Appendix D)

$$\lim_{T\to\infty} \int p^{(1)}(\tilde{\boldsymbol{x}}, x_3, T)\phi(\tilde{x}, x_3)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 = \frac{1}{2}\phi(0, z_0), \tag{6.38}$$

for a compactly supported test function ϕ. This result agrees with the result presented in [14], obtained by an argument that utilized symmetries of the cavity case.

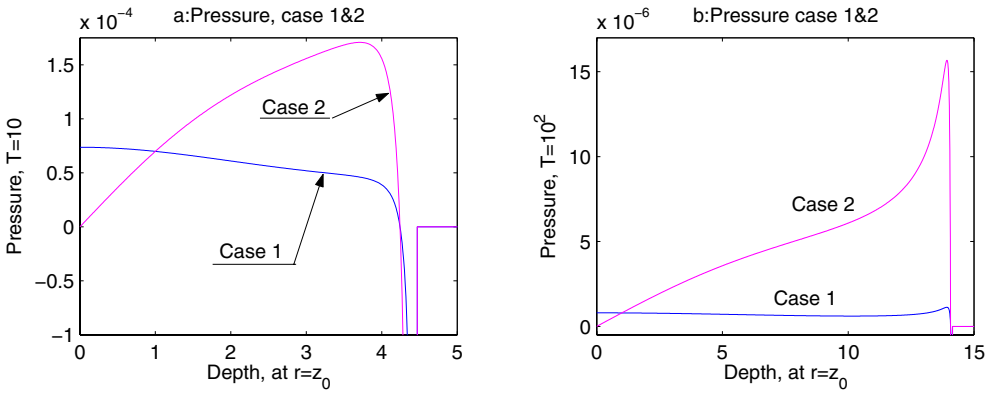


FIG. 6.4. *Comparison between Cases one and two. The plots are along* $r := |\tilde{\boldsymbol{x}}| = z_0$. (a) *The pressure, for* $T = 10$. *The wave front appears at* $r \approx 4$. (b) *The same plot as in Figure* 6.4(a) *but with* $T = 10^2$.

Thus we obtain a perfect retrofocusing in the pressure component, modulo an amplitude factor of 2, in the presence of an evanescent component in the control. The factor of 2 appears since the "controllable part" of the wave field is essentially half of the original wave field, the up-going part of the original pressure pulse that reaches the surface.

We cannot expect a perfect retrofocusing of the pure pressure pulse state for controls on only the half plane, as such controls can only generate "down-going" waves, whereas the pure pressure pulse contains fields that radiate in all directions. The residual of the velocity component ensures that the generated, retrofocused field remains down-going, and hence the velocity component remains nonzero, i.e., not perfectly "retrofitted" to zero.

6.7. Influence of the evanescent part of the boundary control. Here, we discuss the influence of the evanescent part of the boundary control on the resulting fields.

As we noted in section 6.3, the two given boundary controls differ only in the evanescent part, and thus by comparing the responding fields of the respective cases, we compare controls that differ only in the evanescent region. As we derived only the pressure component of the pressure normalized fields explicitly, let us study the difference in pressure. The pressure field differs only by a factor x_3/z_0, and hence the difference is independent of time. To understand the difference between the retrofocused fields we plot the pressure for $\tilde{x} = z_0$; see Figure 6.4.

For Case one, we proved that, as $T \to \infty$, the pressure concentrates at the source point $x_3 = z_0$, $\tilde{x} = 0$. The appearance of the factor x_3/z_0 for Case two does not change the conclusion for $T \to \infty$, as $x_3/z_0 = 1$ at the source point. From the plots in Figure 6.4 we notice the apparent difference between the two cases, but upon observation of the amplitudes involved we conclude that, for sufficiently large T, the evanescent part of the control has a negligible influence. As both cases of controls have a pressure component that converges to a pulse, so must their sum, the response to $p_+^{(3)}$, as the acoustic wave equation is linear.

As we have already shown that the sum of the two controls does not have an evanescent part, and the sum still converges to the pressure Dirac pulse, the influence

of the evanescent part in the long time limit on the pressure component of the final state is small. As the initial pressure is a Dirac pulse, this implies that for any initial state with quiescent velocity and nonzero pressure distribution, the evanescent part of the field has a marginal effect for sufficiently long excitation times.

7. Discussion. The use of time reversal in experiments and theory, both for linear acoustics and electromagnetics, has increased rapidly in the last years. Here, we have developed an iterative time-reversal algorithm for the purpose of retrofocusing that differs from the iterative time-reversal algorithms described by Fink and colleagues [2, 5, 9, 10, 11, 12, 13, 27] and Cheney and coworkers [6, 7]. The algorithms are identical in the first step where they reduce to classical time reversal. The present algorithm retrofocuses the wave field towards the controllable part of its originating distribution; i.e., it uses the transducers in an optimal way to recreate the original (initial) wave field. This is achieved by the construction of identical fields at the transducers in the original and retrofocused states. In contrast to this, the iterative time-reversal algorithm described in [6, 7, 13] retransmit the wave field such that the reflection is maximized. This produces a focusing on the largest scatterer and largest eigenfunction in [13] and [6, 7], respectively.

The iterative time-reversal algorithm is especially useful in strongly multiple scattering cases such as the cavity described in section 5. For this type of geometry, a few iterations improve the retrofocusing, as illustrated by the numerical examples. For the time-reversal mirror examined in this paper, the iterative time-reversal algorithm reduces to the classical time reversal since the homogeneous half space is nonreflecting. But also in this case, the boundary control analysis shows that the iterative time-reversal algorithm is optimally retrofocusing, in the least-squares sense, when the evanescent part of the measured wave field is negligible.

In the half space geometry, both a direct time reversal of the recorded wave field and a weighted time reversal are considered. The two controls differ only in the evanescent part of the wave field. The analytic representation of the pressure field is given for the two controls when the initial field is a pressure Dirac pulse. As expected, the retrofocusing is not perfect; i.e., only the controllable part of the wave field is retrieved. In this case the controllable part is essentially half of the wave field since only the up-going part [18] of the original pressure pulse reaches the surface and is retransmitted as a down-going wave field. The retrofocused field concentrates around the initial pulse point, and as the excitation time approaches infinity, the pressure pulse retrofocuses to half the initial pulse. However, the velocity component does not vanish in the large time limit since the resulting field has to remain down-going.

Appendix A. Calculations on the half space.

A.1. Derivation of the control operator. Here, we give the explicit derivations to obtain the control operator for the particle-velocity normalization.

To derive the control operator for the homogeneous half space, we solve the system of equations

$$\begin{cases} \partial_t p + \nabla \cdot \boldsymbol{v} = 0, & \boldsymbol{x} \in \Omega, \quad t \in (0, T], \\ \partial_t \boldsymbol{v} + \nabla p = 0, & \boldsymbol{x} \in \Omega, \quad t \in (0, T], \\ p = 0,\ \boldsymbol{v} = 0, & \boldsymbol{x} \in \Omega, \quad t = 0, \\ \dfrac{1}{2}(\mathcal{Y} p + v_3) = p_+^{(1)}, & \boldsymbol{x} \in \partial\Omega, \quad t \in [0, T], \end{cases} \tag{A.1}$$

where the boundary condition is in the particle-velocity normalization.

We Laplace transform the field in time, and with the use of $p(\boldsymbol{x},0)=0$ and $\boldsymbol{v}(\boldsymbol{x},0)=0$, together with the enforced causality, the resulting field is analytic for $\operatorname{Re} s \geq 0$. Furthermore, the correspondence $\partial_t \to s$ holds. We Fourier transform the transverse coordinates, $\tilde{\boldsymbol{x}} \to \tilde{\boldsymbol{\xi}}$, and upon eliminating the transverse particle-velocities, we obtain the "two-way" equation for linear acoustic waves

$$\begin{cases} (\partial_3 + \boldsymbol{a}) f(\tilde{\boldsymbol{\xi}}, x_3, s) = 0, & x_3 > 0, \\ \boldsymbol{y} p + v_3 = p_+^{(1)}(\tilde{\boldsymbol{\xi}}, s), & x_3 = 0, \\ s\boldsymbol{v}_\perp(\tilde{\boldsymbol{\xi}}, x_3, s) + \mathrm{i}\tilde{\boldsymbol{\xi}} p(\tilde{\boldsymbol{\xi}}, x_3, s) = 0, & x_3 > 0, \end{cases} \tag{A.2}$$

$f = (p, v_3)$ and where the admittance operator symbol, $\boldsymbol{y}$, is defined in (6.1). Here, the acoustic system's matrix, $\boldsymbol{a}$, has the form

$$\boldsymbol{a} = \begin{pmatrix} 0 & s \\ s + s^{-1}\tilde{\xi}^2 & 0 \end{pmatrix}. \tag{A.3}$$

The formal solution to (A.2) is derived through wave splitting; see, e.g., [14, 15, 16, 17]. Also notice the freedom of "normalization," pointed out in [16, 17, 18]; the normalization is arbitrary and related to the transducer characteristics, where we have chosen the boundary condition in the particle-velocity normalization; i.e., $p_+^{(1)}$ is of "dimension" particle-velocity. The formal solution to (A.2) is

$$\begin{cases} f(\tilde{\boldsymbol{\xi}}, x_3, s) = \mathrm{e}^{-x_3\sqrt{s^2+\tilde{\xi}^2}} \boldsymbol{\eta}^+ p_+^{(1)}(\tilde{\boldsymbol{\xi}}, s), \\ \boldsymbol{v}_\perp(\tilde{\boldsymbol{\xi}}, x_3, s) = -\dfrac{\mathrm{i}\tilde{\boldsymbol{\xi}}}{s} \mathrm{e}^{-x_3\sqrt{s^2+\tilde{\xi}^2}} (\boldsymbol{\eta}^+ p_+^{(1)}(\tilde{\boldsymbol{\xi}}, s))_1, \end{cases} \tag{A.4}$$

where $\boldsymbol{\eta}^+$, $(\boldsymbol{\eta}^-)$ is the eigenvector corresponding to the positive (negative) eigenvalue, $\pm\sqrt{s^2+\tilde{\xi}^2}$ of $\boldsymbol{a}$, and has the form, in particle-velocity normalization,

$$\boldsymbol{\eta}^\pm = \begin{pmatrix} \dfrac{s}{\sqrt{s^2+\tilde{\xi}^2}} \\ \pm 1 \end{pmatrix}. \tag{A.5}$$

With the above reformulations, the symbol of the control operator, $\mathcal{W}_{\mathrm{v}}$, becomes

$$\boldsymbol{w}^+ = \mathrm{e}^{-x_3\sqrt{s^2+\tilde{\xi}^2}} \begin{pmatrix} \dfrac{s}{\sqrt{s^2+\tilde{\xi}^2}} \\ -\dfrac{\mathrm{i}\tilde{\boldsymbol{\xi}}}{\sqrt{s^2+\tilde{\xi}^2}} \\ 1 \end{pmatrix}. \tag{A.6}$$

To transform this operator back into space and time, we observe the identity

$$\mathcal{F}^{-1} \frac{\mathrm{e}^{-x_3\sqrt{s^2+\tilde{\xi}^2}}}{\sqrt{s^2+\tilde{\xi}^2}} = \int_0^\infty \frac{\tilde{\xi}\, \mathrm{J}_0(\tilde{\xi}\tilde{x}) \mathrm{e}^{-x_3\sqrt{s^2+\tilde{\xi}^2}}}{2\pi\sqrt{s^2+\tilde{\xi}^2}}\, \mathrm{d}\tilde{\xi} = \frac{\mathrm{e}^{-s|\boldsymbol{x}|}}{2\pi|\boldsymbol{x}|}, \tag{A.7}$$

by using Purdnikov, Brychkov and Marichev's relation 2.12.10.10 [28]. Thus

$$\mathcal{L}^{-1}\mathcal{F}^{-1} \frac{\mathrm{e}^{-x_3\sqrt{s^2+\tilde{\xi}^2}}}{\sqrt{s^2+\tilde{\xi}^2}} = \frac{\delta(t-|\boldsymbol{x}|)}{2\pi|\boldsymbol{x}|}, \tag{A.8}$$

and as the inverse transform of this integral kernel is known, we express the symbol of the control operator, $\boldsymbol{w}^+$, in terms of the above kernel as

$$\begin{pmatrix} p \\ \tilde{\boldsymbol{v}} \\ v_3 \end{pmatrix} = (\boldsymbol{w}^+ p_+^{(1)})(\tilde{\boldsymbol{\xi}}, x_3, s) = \begin{pmatrix} s \\ -\mathrm{i}\tilde{\boldsymbol{\xi}} \\ -\partial_3 \end{pmatrix} \frac{\mathrm{e}^{-x_3\sqrt{s^2+\tilde{\xi}^2}}}{\sqrt{s^2+\tilde{\xi}^2}} p_+^{(1)}(\tilde{\boldsymbol{\xi}}, s), \tag{A.9}$$

where $\tilde{\boldsymbol{v}} = \{v_1, v_2\}$. Using (A.8) on (A.9), we find that

$$p(\boldsymbol{x}, T) = \int_0^T \int_{\mathbb{R}^2} \frac{\delta(T - t' - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{2\pi|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|} \partial_{t'}(\mathrm{H}(t') p_+^{(1)}(\tilde{\boldsymbol{x}}', t'))\, \mathrm{d}x_1'\, \mathrm{d}x_2'\, \mathrm{d}t' \tag{A.10}$$

and

$$\boldsymbol{v}(\boldsymbol{x}, T) = -\nabla \int_0^T \int_{\mathbb{R}^2} \frac{\delta(T - t' - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{2\pi|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|} p_+^{(1)}(\tilde{\boldsymbol{x}}', t')\, \mathrm{d}x_1'\, \mathrm{d}x_2'\, \mathrm{d}t'. \tag{A.11}$$

Thus we have an explicit expression for the control operator, when the boundary condition is in the particle-velocity normalization. The control operator gives the field inside the domain, once the control at the boundary is known.

A.2. The measured data. In this section a calculation to obtain the explicit form of the measured field at the boundary for Case one is presented.

From (6.26) we have that the measured data has the form

$$q_-^{(0)}(\tilde{\boldsymbol{x}}, t) = \frac{1}{2}((\mathcal{Y}q)(\tilde{\boldsymbol{x}}, 0, t) - u_3(\tilde{\boldsymbol{x}}, 0, t)). \tag{A.12}$$

In the Laplace–Fourier domain the equivalent field has the representation

$$q_-^{(0)}(\tilde{\boldsymbol{\xi}}, s) = \frac{1}{2}\left(s^{-1} q(\tilde{\boldsymbol{\xi}}, 0, s)\sqrt{s^2+\tilde{\xi}^2} - u_3(\tilde{\boldsymbol{\xi}}, 0, s)\right). \tag{A.13}$$

With the relation (A.8) we Laplace–Fourier transform the field $\{q(\tilde{\boldsymbol{x}}, 0, t), u_3(\tilde{\boldsymbol{x}}, 0, t)\}$ at the boundary; cf. (6.25). Substituting the result into (A.13) gives

$$\begin{pmatrix} q(\tilde{\boldsymbol{\xi}}, 0, s) \\ u_3(\tilde{\boldsymbol{\xi}}, 0, s) \end{pmatrix} = \frac{1}{2}\mathrm{e}^{-z_0\sqrt{s^2+\tilde{\xi}^2}} \boldsymbol{\eta}^-(\tilde{\boldsymbol{\xi}}, s) \Rightarrow q_-^{(0)}(\tilde{\boldsymbol{\xi}}, s) = \frac{1}{2}\mathrm{e}^{-z_0\sqrt{s^2+\tilde{\xi}^2}}, \tag{A.14}$$

where $\boldsymbol{\eta}^-$ is an eigenvector of $\boldsymbol{a}$ in the particle-velocity normalization; see (A.5). In the time-space domain, using (A.8), we find that

$$q_-^{(0)}(\tilde{\boldsymbol{x}}, t) = -\partial_{z_0} \frac{\delta(t - \sqrt{\tilde{x}^2 + z_0^2})}{4\pi\sqrt{\tilde{x}^2 + z_0^2}}. \tag{A.15}$$

Appendix B. The field in the domain for control $p_+^{(1)}$. Given the control $p_+^{(1)}$ in (6.30), we substitute it into (6.19) and (6.20). Below we explicitly calculate the two resulting distributions.

B.1. The pressure component. In this section we give a detailed derivation of the pressure distribution for Case one.

Let us introduce the help quantities,

$$R_1 = \sqrt{(\tilde{x}')^2 + z_0^2}, \qquad R_2 = \sqrt{|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}'|^2 + x_3^2}, \quad \text{and} \quad \tau = T - t. \tag{B.1}$$

The pressure, $p^{(1)}$, with the use of (B.1) is represented as

$$\begin{aligned} p^{(1)}(\boldsymbol{x},t) &= -\partial_{z_0}\partial_t \int_{\mathbb{R}^2} \mathrm{H}(t-R_2)\frac{\delta(\tau+(R_2-R_1))}{8\pi^2 R_1 R_2}\,\mathrm{d}x_1'\,\mathrm{d}x_2' \\ &= \partial_{z_0}x_3^{-1}\partial_3 \int_{\mathbb{R}^2} \mathrm{H}(t-R_2)\frac{\delta(\tau+(R_2-R_1))}{8\pi^2 R_1}\,\mathrm{d}x_1'\,\mathrm{d}x_2'. \end{aligned} \tag{B.2}$$

The delta Dirac traces out a curve for $\tilde{\boldsymbol{x}}'$; to find the curve consider the $\tilde{\boldsymbol{x}}'$ such that $R_1 - R_2 - \tau = 0$:

$$\begin{gathered} R_1^2 = (\tau+R_2)^2 = \tau^2 + R_2^2 + 2\tau R_2 \Leftrightarrow \\ -\tau\sqrt{x_3^2+|\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{x}}'|^2} = (\tilde{x}^2 + x_3^2 - z_0^2 + \tau^2)/2 - \tilde{\boldsymbol{x}}\cdot\tilde{\boldsymbol{x}}' \Leftrightarrow \\ \tau^2(x_3^2+\tilde{x}^2+(\tilde{x}')^2 - 2\tilde{\boldsymbol{x}}\cdot\tilde{\boldsymbol{x}}') = ((\tilde{x}^2+x_3^2-z_0^2+\tau^2)/2 - \tilde{\boldsymbol{x}}\cdot\tilde{\boldsymbol{x}}')^2 \Leftrightarrow \\ \tau^2(x_3^2+\tilde{x}^2) - A^2/4 = -\tau^2(\tilde{x}')^2 + (\tilde{\boldsymbol{x}}\cdot\tilde{\boldsymbol{x}}')^2 - A(\tilde{\boldsymbol{x}}\cdot\tilde{\boldsymbol{x}}'); \end{gathered} \tag{B.3}$$

hence $\delta(R_1 - R_2 - \tau)$ traces out a conical section. Here,

$$A = \tilde{x}^2 + x_3^2 - z_0^2 + \tau^2. \tag{B.4}$$

We observe the freedom of choice in the coordinates $\tilde{\boldsymbol{x}}'$, and hence we choose the particular coordinate system for $\tilde{\boldsymbol{x}}'$ such that $\tilde{\boldsymbol{x}} = (\tilde{x}, 0)$. This is equivalent to rotating the coordinate system of $\tilde{\boldsymbol{x}}'$. The compatibility condition imposed on the solution associated with the square roots is

$$-\operatorname{sgn}(\tau) = \operatorname{sgn}(\tilde{x}^2 + x_3^2 - z_0^2 + \tau^2 - 2\tilde{x}x_1'), \tag{B.5}$$

as is observed from the second line of (B.3). Now we rewrite (B.3) into a more standard form for conical sections,

$$c = (\tilde{x}^2 - \tau^2)(x_1' - x_1^0)^2 - \tau^2(x_2')^2, \tag{B.6}$$

and hence the set of $\{x_1', x_2'\}$, which fulfills (B.6), traces out a curve in space. In this case

$$x_1^0 = \tilde{x}\frac{A/2-\tau^2}{\tilde{x}^2-\tau^2} \quad \text{and} \quad c = \tau^2(x_3^2+\tilde{x}^2) - \frac{A^2}{4} + (\tilde{x}^2-\tau^2)(x_1^0)^2. \tag{B.7}$$

We notice that if $\tilde{x} > \tau$, the curve traced out is a hyperbola; when $\tilde{x} < \tau$, it is an ellipse; and when $\tilde{x} = \tau$, it is a line parallel to x_1'-axis.

In evaluating the integral (B.2) we consider only the limiting case when $t = T$, i.e., $\tau = 0$. Thus the integral reduces to

$$p^{(1)}(\boldsymbol{x},T) = \frac{1}{x_3}\partial_{z_0}\partial_3 \int_{\mathbb{R}^2} \mathrm{H}(T-R_2)\frac{\delta(R_2-R_1)}{8\pi^2 R_1}\,\mathrm{d}x_1'\,\mathrm{d}x_2'. \tag{B.8}$$

We find that

$$x_1^0 = \frac{A}{(2\tilde{x})} = \frac{\tilde{x}}{2}\left(1 + \frac{x_3^2-z_0^2}{\tilde{x}^2}\right), \quad c = 0,\ A = \tilde{x}^2 + x_3^2 - z_0^2, \tag{B.9}$$

and the conical curve (B.6) collapses into the line

$$x_1' = x_1^0 \Rightarrow x_1 = \frac{\tilde{x}}{2}\left(1 + \frac{x_3^2-z_0^2}{2\tilde{x}^2}\right). \tag{B.10}$$

This line fulfills the compatibility condition (B.5), and hence it is a solution.

To evaluate the integral we introduce a change of coordinates, $\Psi = R_1 - R_2$, and the arc length, s, along the line (B.10). The integral in those coordinates collapses into an integral with integrand

$$(R_2|\nabla\Psi|)^{-1}\,\mathrm{d}s = \tilde{x}^{-1}\,\mathrm{d}x_2'. \tag{B.11}$$

To see this, first note that

$$\mathrm{d}s = \sqrt{1+\left(\frac{\partial x_1'}{\partial x_2'}\right)^2}\,\mathrm{d}x_2' = \mathrm{d}x_2' \tag{B.12}$$

and

$$(R_1R_2)^2|\nabla'(R_1-R_2)|^2_{\Psi=0} = |(R_2-R_1)\tilde{\boldsymbol{x}}' + R_1\tilde{\boldsymbol{x}}|^2_{\Psi=0} = R_1^2\tilde{x}^2|_{\Psi=0}, \tag{B.13}$$

where we have used $R_1 = R_2$, or equivalently $x_1' = x_1^0$. Hence, the pressure integral becomes

$$p^{(1)}(\boldsymbol{x},T) = \frac{1}{8\pi^2\tilde{x}x_3}\partial_{z_0}\partial_3\int_{\mathbb{R}}\mathrm{H}(T-R_1)|_{R_1=R_2}\,\mathrm{d}x_2', \tag{B.14}$$

i.e., the length of the line, $x_1' = x_1^0$, inside the circle described by $\mathrm{H}(T-R_1)$. The height, x_2', where $x_1' = x_1^0$ crosses the circle, is

$$\begin{aligned} &T^2 = z_0^2 + (x_1')^2 + (x_2')^2|_{x_1'=x_1^0} = z_0^2 + (x_1^0)^2 + (x_2')^2 \Rightarrow \\ &x_2' = \pm\sqrt{T^2 - z_0^2 - (x_1^0)^2}, \end{aligned} \tag{B.15}$$

and hence

$$p^{(1)}(\boldsymbol{x},T) = \frac{1}{4\pi^2\tilde{x}x_3}\partial_{z_0}\partial_3\left(\sqrt{T^2-(z_0^2+(x_1^0)^2)}\mathrm{H}\left(T-\sqrt{z_0^2+(x_1^0)^2}\right)\right). \tag{B.16}$$

We now let the derivative with respect to the parameter z_0 act on the distribution. We first observe that

$$\partial_{z_0}\sqrt{T^2-(z_0^2+(x_1^0)^2)} = \frac{-z_0(x-x_1^0)}{\tilde{x}\sqrt{T^2-z_0^2-(x_1^0)^2}}, \tag{B.17}$$

so that

$$\begin{aligned} p^{(1)}(\boldsymbol{x},T) &= \frac{-z_0}{4\pi^2x_3\tilde{x}^2}\partial_3\frac{\tilde{x}-x_1^0}{(T^2-z_0^2-(x_1^0)^2)_+^{1/2}} \\ &= \frac{-z_0}{4\pi^2x_3\tilde{x}^2}\partial_3\frac{\tilde{x}^2-x_3^2+z_0^2}{(4\tilde{x}^2(T^2-z_0^2)-A^2)_+^{1/2}}. \end{aligned} \tag{B.18}$$

We rewrite the denominator in the form

$$\left(\left[T^2-x_3^2-\left(\tilde{x}-\sqrt{T^2-z_0^2}\right)^2\right]\left[\left(\tilde{x}+\sqrt{T^2-z_0^2}\right)^2+x_3^2-T^2\right]\right)_+^{1/2}, \tag{B.19}$$

where the plus sign indicates that we consider it as a generalized function, and the value within the outer parentheses must be positive. With (B.18) and (B.19) we have obtained the pressure inside the domain corresponding to the control $p_+^{(1)}$. It is a distribution, and is considered as acting on smooth test functions.

B.2. The velocity component. With the same notation as in the previous section, we write the particle velocity inside the domain for the boundary control $p_+^{(1)}$, in the limit $t = T$, as

$$\boldsymbol{v}^{(1)}(\boldsymbol{x}, T) = \nabla\partial_{z_0} \int_{\mathbb{R}^2} \mathrm{H}(T - R_2) \frac{\delta(R_2 - R_1)}{8\pi^2 R_1 R_2} \,\mathrm{d}x_1' \,\mathrm{d}x_2'. \tag{B.20}$$

With the change to the arc-length coordinates, we have (cf. (B.11))

$$\boldsymbol{v}^{(1)}(\boldsymbol{x}, T) = \nabla \frac{1}{8\pi^2\tilde{x}} \partial_{z_0} \int_{\mathbb{R}} \frac{\mathrm{H}(T - R_2)}{R_1}\bigg|_{R_1=R_2} \mathrm{d}x_2'. \tag{B.21}$$

Analogous to the derivation leading up to (B.10), the condition $R_1 = R_2$ is equivalent to $x_1' = x_1^0$, and hence

$$R_1|_{x_1'=x_1^0} = ((x_1^0)^2 + z_0^2 + (x_2')^2)^{1/2}. \tag{B.22}$$

The integral is straightforward once we notice that the step function imposes the boundary value of x_2' (see (B.15))

$$\begin{aligned} &\boldsymbol{v}^{(1)}(\boldsymbol{x}, T) \\ &= \nabla \left[\frac{1}{8\pi^2\tilde{x}} \partial_{z_0} \ln \left(\left| x_2' + \sqrt{(x_2')^2 + z_0^2 + (x_1^0)^2} \right| \right) \right] \Big|_{x_2'=\{x_2': R_1|_{x_1'=x_1^0}=T\}} \\ &= \nabla \left[\frac{1}{8\pi^2\tilde{x}} \partial_{z_0} \left(\ln \left(\frac{T + \sqrt{T^2 - z_0^2 - (x_1^0)^2}}{T - \sqrt{T^2 - z_0^2 - (x_1^0)^2}} \right) \mathrm{H}\left(T - c^{-1}\sqrt{z_0^2 + (x_1^0)^2} \right) \right) \right]. \end{aligned} \tag{B.23}$$

Upon evaluating the derivative with respect to z_0, we obtain

$$\boldsymbol{v}^{(1)}(\boldsymbol{x}, T) = \nabla \left(\frac{z_0}{4\pi^2\tilde{x}^2} \frac{T(x_1^0 - \tilde{x})}{(T^2 - z_0^2 - (x_1^0)^2)_+^{1/2} (z_0^2 + (x_1^0)^2)} \right), \tag{B.24}$$

where the generalized function in the denominator is the same as for the pressure (B.17) and can be rewritten as (B.19).

Appendix C. The response field for Case two. To obtain the response for the control in Case two, we start to examine the characteristics of our transducers. The derivation in Appendices A and B is for the particle-velocity normalization of the boundary condition, both for the control and for the measurement. If we use the pressure normalization instead of velocity normalization, then the boundary condition takes the form (cf. (6.15))

$$\frac{p + \mathcal{Y}^{-1} v_3}{2} = p_{+,\mathrm{N_p}}^{(2)} \quad \text{and} \quad m_{(2)} = \frac{q - \mathcal{Y}^{-1} u_3}{2}. \tag{C.1}$$

The control operator is then expressed in terms of the eigenvector in the pressure normalization; i.e., $\boldsymbol{\eta}^+$ in (A.4) is replaced by $\boldsymbol{\eta}_{(2)}^+$, where

$$\boldsymbol{\eta}_{(2)}^{\pm} = \begin{pmatrix} 1 \\ \pm s^{-1}\sqrt{s^2 + \tilde{\xi}^2} \end{pmatrix}. \tag{C.2}$$

Thus the symbol of the control operator, $\boldsymbol{w}^{+}_{(2)}$, in this normalization is

$$\begin{pmatrix} p \\ \tilde{\boldsymbol{v}} \\ v_3 \end{pmatrix} = \left(\boldsymbol{w}^{+}_{(2)} p^{(2)}_{+,\mathrm{N_p}}\right)(\tilde{\boldsymbol{\xi}}, x_3, s) = \begin{pmatrix} -\partial_3 \\ s^{-1}\mathrm{i}\tilde{\boldsymbol{\xi}}\partial_3 \\ s^{-1}\partial_3^2 \end{pmatrix} \frac{\mathrm{e}^{-x_3\sqrt{s^2+\tilde{\xi}^2}}}{\sqrt{s^2+\tilde{\xi}^2}} p^{(2)}_{+,\mathrm{N_p}}(\tilde{\boldsymbol{\xi}}, s). \tag{C.3}$$

Using (A.8), we obtain the control operator as

$$p(\boldsymbol{x}, T) = -\partial_3 \int_0^T \int_{\mathbb{R}^2} \frac{\delta(T - t' - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{2\pi|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|} p^{(2)}_{+,\mathrm{N_p}}(\tilde{\boldsymbol{x}}', t')\, \mathrm{d}x_1'\, \mathrm{d}x_2'\, \mathrm{d}t' \tag{C.4}$$

and

$$\boldsymbol{v}(\boldsymbol{x}, T) = \nabla\partial_3 \int_0^T \int_{\mathbb{R}^2} \frac{\delta(T - t' - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{2\pi|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|} \int_0^{t'} p^{(2)}_{+,\mathrm{N_p}}(\tilde{\boldsymbol{x}}', t'')\, \mathrm{d}t''\, \mathrm{d}x_1'\, \mathrm{d}x_2'\, \mathrm{d}t'. \tag{C.5}$$

Hence, the normalization of the boundary condition changes the field, as expected; cf. (6.19) and (6.20). The change related to the different transducer normalizations can be compared to solving a partial differential equation with the Neumann and Dirichlet boundary conditions, respectively.

The field from the pressure pulse at the surface in (6.25), $\{q, u_3\}$, is independent of normalization, but we measure particle-velocity (see (C.1)); thus in the pressure normalization the measured signal becomes

$$\begin{pmatrix} q(\tilde{\boldsymbol{\xi}}, 0, s) \\ u_3(\tilde{\boldsymbol{\xi}}, 0, s) \end{pmatrix} = \frac{s\mathrm{e}^{-z_0\sqrt{s^2+\tilde{\xi}^2}}}{2\sqrt{s^2+\tilde{\xi}^2}} \boldsymbol{\eta}^{-}_{(2)}(\tilde{\boldsymbol{\xi}}, s) \Rightarrow m_{(2)}(\tilde{\boldsymbol{\xi}}, s) = \frac{s\mathrm{e}^{-z_0\sqrt{s^2+\tilde{\xi}^2}}}{2\sqrt{s^2+\tilde{\xi}^2}} \tag{C.6}$$

in the transform domain, and hence

$$m_{(2)}(\tilde{\boldsymbol{x}}, t) = \partial_t \frac{\delta(t - \sqrt{\tilde{x}^2 + z_0^2})}{4\pi\sqrt{\tilde{x}^2 + z_0^2}}. \tag{C.7}$$

Thus the control corresponding to the $m_{(2)}$ measurement is given by

$$p^{(2)}_{+,\mathrm{N_p}} = \mathrm{H}(t) m_{(2)}(\tilde{\boldsymbol{x}}, T - t) = -\mathrm{H}(t)\partial_t \frac{\delta(T - t - \sqrt{\tilde{x}^2 + z_0^2})}{4\pi\sqrt{\tilde{x}^2 + z_0^2}}. \tag{C.8}$$

We substitute the control (C.8) into (C.4) and (C.5) to obtain

$$p^{(2)}(\boldsymbol{x}, T) = \partial_3 z_0^{-1} \partial_{z_0} \int_{\mathbb{R}^2} \mathrm{H}(T - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|) \frac{\delta(|\boldsymbol{x} - \tilde{\boldsymbol{x}}'| - \sqrt{(\tilde{x}')^2 + z_0^2})}{8\pi^2|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|}\, \mathrm{d}x_1'\, \mathrm{d}x_2' \tag{C.9}$$

and

$$\boldsymbol{v}^{(2)}(\boldsymbol{x}, T) = -\nabla\partial_3 z_0^{-1} \partial_{z_0} \int_{\mathbb{R}^2} \mathrm{H}\left(T - \sqrt{(\tilde{x}')^2 + z_0^2}\right) \frac{\mathrm{H}(\sqrt{(\tilde{x}')^2 + z_0^2} - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{8\pi^2|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|}\, \mathrm{d}x_1'\, \mathrm{d}x_2'. \tag{C.10}$$

We rewrite (C.10) by evaluating ∂_{z_0}; thus

$$\begin{aligned} \boldsymbol{v}^{(2)}(\boldsymbol{x}, T) = -\nabla\partial_3 \int_{\mathbb{R}^2} & \mathrm{H}\left(T - \sqrt{(\tilde{x}')^2 + z_0^2}\right) \frac{\delta(\sqrt{(\tilde{x}')^2 + z_0^2} - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{8\pi^2|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|\sqrt{(\tilde{x}')^2 + z_0^2}} \\ & - \delta\left(T - \sqrt{(\tilde{x}')^2 + z_0^2}\right) \frac{\mathrm{H}(\sqrt{(\tilde{x}')^2 + z_0^2} - |\boldsymbol{x} - \tilde{\boldsymbol{x}}'|)}{8\pi^2|\boldsymbol{x} - \tilde{\boldsymbol{x}}'|\sqrt{(\tilde{x}')^2 + z_0^2}}\, \mathrm{d}x_1'\, \mathrm{d}x_2'. \end{aligned} \tag{C.11}$$

The evaluation of (C.9) is analogous to the corresponding case in the particle velocity normalization (see (B.8)), and thus we obtain

(C.12)
$$p^{(2)}(\boldsymbol{x},T) = \frac{-1}{4\pi^2\tilde{x}^2}\partial_3\left(\frac{\tilde{x}^2 - x_3^2 + z_0^2}{([T^2 - x_3^2 - (\tilde{x} - \sqrt{T^2 - z_0^2})^2][(\tilde{x} + \sqrt{T^2 - z_0^2})^2 + x_3^2 - T^2])_+^{1/2}}\right).$$

Appendix D. The pressure response distribution for $T \to \infty$. In this section, we give the details for the calculation of the limit $T \to \infty$ of the pressure response (6.32) in Case one.

We rewrite the denominator of (6.32) in the more convenient form

(D.1)
$$p^{(1)}(\boldsymbol{x},T) = \frac{-z_0}{4\pi^2 x_3\tilde{x}^2}\partial_3\left(\frac{\tilde{x}^2 - x_3^2 + z_0^2}{([(\sqrt{T^2 - z_0^2} + \sqrt{T^2 - x_3^2})^2 - \tilde{x}^2][\tilde{x}^2 - (\sqrt{T^2 - z_0^2} - \sqrt{T^2 - x_3^2})^2])_+^{1/2}}\right).$$

In the limit $T \to \infty$, let $\phi = \phi(\tilde{\boldsymbol{x}}, x_3)$ be a compactly supported test function, i.e., smooth and bounded. We require that

$$\operatorname{supp}\phi \subset \{\boldsymbol{x} \in \mathbb{R}^3 : x_3 > 0\} \quad \text{and} \quad \operatorname{diam}\phi \le D_\phi, \tag{D.2}$$

for some fixed number $D_\phi > 0$. As we are interested only in the limit $T \to \infty$, we require that $T \gg z_0$. Let us define the pressure functional $p_f = p_f(T)$ as

$$p_f = \int_{\mathbb{R}^3_+ \cap \operatorname{supp}\phi} p^{(1)}(\boldsymbol{x},T)\phi(\tilde{\boldsymbol{x}}, x_3)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3. \tag{D.3}$$

By partial integration we push the derivative to the test function to obtain

$$\begin{aligned} p_f = &\int_{\mathbb{R}^3_+ \cap \operatorname{supp}\phi} \frac{z_0}{4\pi^2}\left(\partial_3\frac{\phi(\tilde{\boldsymbol{x}}, x_3)}{x_3}\right) \\ &\times \frac{(\tilde{x}^2 - x_3^2 + z_0^2)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3}{\tilde{x}^2([(\sqrt{T^2 - z_0^2} + \sqrt{T^2 - x_3^2})^2 - \tilde{x}^2][\tilde{x}^2 - (\sqrt{T^2 - z_0^2} - \sqrt{T^2 - x_3^2})^2])_+^{1/2}}. \end{aligned} \tag{D.4}$$

We eliminate the appearance of T in the integrand with the change of variables

$$\check{z}_0 = \frac{z_0}{\sqrt{T^2 - z_0^2}}, \qquad \check{x}_3 = \frac{x_3}{\sqrt{T^2 - z_0^2}}, \qquad \check{r} = \frac{\tilde{x}}{\sqrt{T^2 - z_0^2}}, \tag{D.5}$$

giving $\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 = (T^2 - z_0^2)^{3/2}\check{r}\,\mathrm{d}\check{r}\,\mathrm{d}\check{x}_3\,\mathrm{d}\theta$. With this change of variables the denominator takes the form

$$(\dots)_+^{1/2} = (T^2 - z_0^2)\left(\left[\left(1 + \sqrt{1 - \check{x}_3^2 + \check{z}_0^2}\right)^2 - \check{r}^2\right]\left[\check{r}^2 - \left(1 - \sqrt{1 - \check{x}_3^2 + \check{z}_0^2}\right)^2\right]\right)_+^{1/2}.$$

For notational convenience let

(D.6)
$$\psi(\check{r}\cos\theta, \check{r}\sin\theta, \check{x}_3, T) = \frac{\check{z}_0}{2\pi}\frac{\partial}{\partial\check{x}_3}\frac{\phi(\sqrt{T^2 - z_0^2}\check{r}\cos\theta, \sqrt{T^2 - z_0^2}\check{r}\sin\theta, \sqrt{T^2 - z_0^2}\check{x}_3)}{\check{x}_3}.$$

The limit of the integration, $\mathbb{R}^3_+ \cap \operatorname{supp}\phi$, in the new variables is included in the set

$$(\check{x}_3 - \check{z}_0)^2 + \check{r}^2 \le \frac{D_\phi^2}{T^2 - z_0^2} \equiv \check{D}_\phi^2, \tag{D.7}$$

for sufficiently large T. Hence, we only have to consider small $\check{r}$ and $|\check{x}_3 - \check{z}_0|$ as we make T arbitrary large. In particular, we can expand the test function ψ around $\check{r} = 0$: $\psi(\check{\boldsymbol{x}}, x_3) = \psi(0, 0, x_3, T) + \check{r}\psi_{\check{x}_1}\cos\theta + \check{r}\psi_{\check{x}_2}\sin\theta + \mathrm{O}(\check{r}^2)$. As the integrand, apart from the test function, is independent of θ, we obtain

$$\begin{aligned} p_f = &\int_{\sqrt{(\check{x}_3-\check{z}_0)^2+\check{r}^2}\le\check{D}_\phi} (\psi(0,0,\check{x}_3,T) + \mathrm{O}(T^{-2})) \\ &\times \frac{(\check{r}^2 - \check{x}_3^2 + \check{z}_0^2)\,\mathrm{d}\check{r}\,\mathrm{d}\check{x}_3}{\check{r}([(1+\sqrt{1-\check{x}_3^2+\check{z}_0^2})^2 - \check{r}^2][\check{r}^2 - (1-\sqrt{1-\check{x}_3^2+\check{z}_0^2})^2])_+^{1/2}}. \end{aligned} \tag{D.8}$$

The $\check{r}$-integral can be integrated exactly, but as we are only interested in the limit $T \to \infty$, we simplify the above expression as both $|\check{x}_3 - \check{z}_0|$ and $\check{r}$ are bounded above by $\check{D}_\phi = \mathrm{O}(T^{-1}) \ll 1$ for sufficiently large T. Applying Taylor expansion gives

$$1 + \sqrt{1-\check{x}_3^2+\check{z}_0^2} = 2 + \mathrm{O}(T^{-1}), \qquad 1 - \sqrt{1-\check{x}_3^2+\check{z}_0^2} = \frac{\check{x}_3^2 - \check{z}_0^2}{2} + \mathrm{O}(T^{-2}), \tag{D.9}$$

and hence

$$(\dots)_+^{1/2} = ([4\check{r}^2 - (\check{x}_3^2 - \check{z}_0^2)^2])_+^{1/2} + \mathrm{O}(T^{-1}). \tag{D.10}$$

Thus the pressure functional becomes

$$p_f = \int_{\sqrt{(\check{x}_3-\check{z}_0)^2+\check{r}^2}\le\check{D}_\phi} \frac{\check{r}^2 - \check{x}_3^2 + \check{z}_0^2}{\check{r}[4\check{r}^2 - (\check{x}_3^2 - \check{z}_0^2)^2]_+^{1/2}} \psi(0,0,\check{x}_3,T)\,\mathrm{d}\check{r}\,\mathrm{d}\check{x}_3 + \cdots. \tag{D.11}$$

The disk $\sqrt{(\check{x}_3 - \check{z}_0)^2 + \check{r}^2} \le \check{D}_\phi$, together with the step function indicated by the plus sign on $(\dots)_+^{1/2}$, is depicted in Figure D.1. With the change of variables

$$\check{r} = \frac{|\check{x}_3^2 - \check{z}_0^2|u}{2} = \frac{\zeta u}{2}, \tag{D.12}$$

we find

$$p_f = \int_{|\check{x}_3-\check{z}_0|f(\check{x}_3)\le 2\check{D}_\phi} \int_1^{2\zeta^{-1}\sqrt{\check{D}_\phi^2-(\check{x}_3-\check{z}_0)^2}} \psi(0,0,\check{x}_3,T) \frac{\zeta^2u^2/4 - \check{x}_3^2 + \check{z}_0^2}{\zeta u(u^2-1)^{1/2}}\,\mathrm{d}u\,\mathrm{d}\check{x}_3 + \cdots,$$

where $f(\check{x}_3) = \sqrt{4 + (\check{x}_3 + \check{z}_0)^2}$. Upon integrating, we have

$$\begin{aligned} p_f = -\int_{|\check{x}_3-\check{z}_0|f(\check{x}_3)\le 2\check{D}_\phi} \psi(0,0,\check{x}_3,T) \Bigg(&\frac{(\check{x}_3^2 - \check{z}_0^2)\pi}{2|\check{x}_3^2 - \check{z}_0^2|} - \frac{1}{4}\sqrt{4\check{D}_\phi^2 - 4(\check{x}_3 - \check{z}_0)^2 - \zeta^2} \\ &- \frac{\check{x}_3^2 - \check{z}_0^2}{\zeta}\arctan(\zeta(4\check{D}_\phi^2 - (\check{x}_3 - \check{z}_0)^2 - \zeta^2)^{-1/2})\Bigg)\mathrm{d}\check{x}_3 + \cdots. \end{aligned}$$

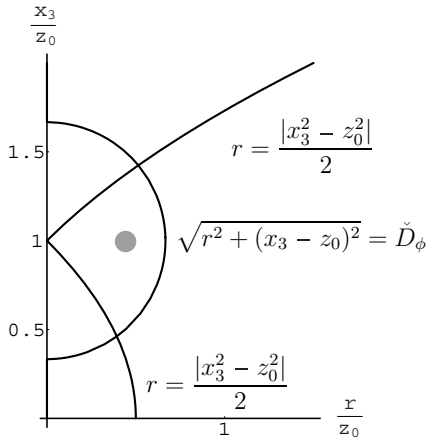


FIG. D.1. *The triangle with the gray dot is the area of integration. The outer circle bounds the domain of ψ, and the cut-off, $(\dots)_+^{1/2}$, is shown as the two lines entering the half-circle.*

Observe that on the given interval we have the upper bound

$$\frac{1}{4}\sqrt{4\check{D}_\phi^2 - 4(\check{x}_3 - \check{z}_0)^2 - \zeta^2} \le \frac{\check{D}_\phi}{2}; \tag{D.13}$$

this, together with the facts that the square root is a continuous function and that the test function is bounded above, gives that the integral of this term vanishes as $T \to \infty$. For the arctan term, we observe that it is continuous at $\check{x}_3 = \check{z}_0$, since $\arctan\epsilon = \epsilon + \cdots$, and furthermore, on the given interval

$$|\operatorname{sgn}(\check{x}_3^2 - \check{z}_0^2)\arctan(\zeta(4\check{D}_\phi^2 - (\check{x}_3 - \check{z}_0)^2 - \zeta^2)^{-1/2})| \le \frac{\pi}{2}; \tag{D.14}$$

thus, as the test function is bounded, this term also gives a vanishing contribution to the integral.

With the above considerations the pressure functional becomes

$$p_f = -\frac{\pi}{2}\int_{|\check{x}_3 - \check{z}_0| f(\check{x}_3) \le 2\check{D}_\phi} \psi(0,0,\check{x}_3,T)\frac{\check{x}_3^2 - \check{z}_0^2}{|\check{x}_3^2 - \check{z}_0^2|}\,\mathrm{d}\check{x}_3 + \mathrm{O}(T^{-1}). \tag{D.15}$$

If we now substitute the expression for ψ, (D.6), we find that

$$p_f = -\frac{\check{z}_0}{4}\int_{|\check{x}_3 - \check{z}_0| f(\check{x}_3) \le 2\check{D}_\phi} \left(\partial_{\check{x}_3}\frac{\phi(0,\check{x}_3\sqrt{T^2 - z_0^2})}{\check{x}_3}\right)\operatorname{sgn}(\check{x}_3^2 - \check{z}_0^2)\,\mathrm{d}\check{x}_3 + \mathrm{O}(T^{-1}). \tag{D.16}$$

This integral is evaluated as

$$\begin{aligned} p_f &= \frac{\check{z}_0}{4}\int_{|\check{x}_3 - \check{z}_0| f(\check{x}_3) \le 2\check{D}_\phi} \phi\left(0,\check{x}_3\sqrt{T^2 - z_0^2}\right)(\check{x}_3^{-1}\partial_{\check{x}_3}\operatorname{sgn}(\check{x}_3^2 - \check{z}_0^2))\,\mathrm{d}\check{x}_3 + \mathrm{O}(T^{-1}) \\ &= \frac{1}{2}\phi(0,z_0) + \mathrm{O}(T^{-1}). \end{aligned} \tag{D.17}$$

In the integration we used that z_0 is always in the domain $|\check{x}_3 - \check{z}_0| f(\check{x}_3) \leq 2\check{D}_\phi$, for sufficiently small $\check{D}_\phi$ or correspondingly for large enough T. In the limit $T \to \infty$ we find that the distribution reduces to a delta Dirac at $\tilde{\boldsymbol{x}} = 0$ and $x_3 = z_0$. Hence, in the limit we get back half the original pressure pulse in the pressure component.

REFERENCES

[1] G. S. S. Ávila and D. G. Costa, *Asymptotic properties of general symmetric hyperbolic systems*, J. Funct. Phys., 35 (1980), pp. 49–63.

[2] C. Bardos and M. Fink, *Deterministic Mathematical Analysis of the Time Reversal Mirror*, http://www.msri.org/publications/video/index03.html, 2001.

[3] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[4] P. Blomgren, G. Papanicolaou, and H. Zhao, *Super-resolution in time-reversal acoustics*, J. Acoust. Soc. Am., 111 (2002), pp. 230–248.

[5] D. Cassereau and M. Fink, *Time-reversal of ultrasonic fields—Part* III*: Theory of the closed time-reversal cavity*, IEEE Trans. Ultrason. Ferroelec. Freq. Contr., 39 (1992), pp. 579–592.

[6] M. Cheney, D. Isaacson, and M. Lassas, *Optimal acoustical measurements*, SIAM J. Appl. Math., 61 (2001), pp. 1628–1647.

[7] M. Cheney and G. Kristensson, *Optimal electromagnetic measurements*, J. Electromagn. Waves Appl., 15 (2001), pp. 1323–1336.

[8] H. O. Cordes, *The Technique of Pseudodifferential Operators*, London Mathematical Society Lecture Notes Series 202, Cambridge University Press, Cambridge, UK, 1995.

[9] A. Derode, P. Roux, and M. Fink, *Robust acoustic time reversal with high-order multiple scattering*, Phys. Rev. Lett., 75 (1995), pp. 4206–4209.

[10] M. Fink, *Time-reversal mirrors*, J. Phys. D: Appl. Phys., 26 (1993), pp. 1333–1350.

[11] M. Fink, *Time reversed acoustics*, Physics Today, 3 (1997), pp. 34–40.

[12] M. Fink, *Time-reversed acoustics*, Scientific American, 281 (1999), pp. 91–97.

[13] M. Fink and C. Prada, *Acoustic time-reversal mirrors*, Inverse Problems, 17 (2001), pp. R1–R38.

[14] M. Gustafsson, *Wave Splitting in Direct and Inverse Scattering Problems*, Ph.D. thesis, Lund University, Lund, Sweden, 2000; available online at http://www.es.lth.se/home/mats.

[15] S. He, S. Ström, and V. H. Weston, *Time Domain Wave-Splitting and Inverse Problems*, Oxford University Press, Oxford, UK, 1998.

[16] M. V. de Hoop, *Generalization of the Bremmer coupling series*, J. Math. Phys., 37 (1996), pp. 3246–3282.

[17] B. L. G. Jonsson, *Directional Decomposition in Anisotropic Heterogeneous Media for Acoustic and Electromagnetic Fields*, Ph.D. thesis, Royal Institute of Technology, Stockholm, Sweden, 2001; available online at http://media.lib.kth.se/dissengref.asp?dissnr=3099.

[18] B. L. G. Jonsson and M. V. de Hoop, *Wave field decomposition in anisotropic fluids*, Acta Appl. Math., 67 (2001), pp. 117–171.

[19] A. Kim, P. Blomgren, and G. Papanicolaou, *Spatial focusing and intersymbol interference in time reversal communications*, submitted, 2003; available online at http://georgep.stanford.edu/~papanico/pubs.html.

[20] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier-Stokes Equations*, Academic Press, San Diego, CA, 1989.

[21] R. Leis, *Initial Boundary Value Problems in Mathematical Physics*, B. G. Teubner, Stuttgart, 1986.

[22] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Die Grundlehren Math. Wiss. 170, Springer-Verlag, Berlin, 1971.

[23] C. S. Morawetz, *Notes on Time Decay and Scattering for Some Hyperbolic Problems*, CBMS-NSF Regional Co. Ser. in Appl. Math. 19, SIAM, Philadelphia, 1975.

[24] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Vol. II, MacGraw-Hill, New York, 1953.

[25] A. W. Naylor and G. R. Sell, *Linear Operator Theory in Engineering and Science*, 2nd ed., Springer-Verlag, New York, 1982.

[26] A. D. Pierce, *Acoustics: An Introduction to its Physical Principles and Applications*, Acoustical Society of America, New York, 1989.

[27] C. Prada, J.-L. Thomas, and M. Fink, *The iterative time reversal process: Analysis of the convergence*, J. Acoust. Soc. Am., 97 (1995), pp. 62–71.

[28] A. P. Purdnikov, Y. A. Brychkov, and O. I. Marichev, *Integrals and Series Vol.* 2*: Special Functions*, Gordon and Brech Science Publishers, London, 1986.
[29] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, Boston, 1996.
[30] M. A. Shubin, *Pseudodifferential Operators and Spectral Theory*, Springer Series in Soviet Mathematics, Springer-Verlag, Berlin, 1987.
[31] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Chapman & Hall, New York, 1989.
[32] V. H. Weston, *Factorization of the wave equation in higher dimensions*, J. Math. Phys., 28 (1987), pp. 1061–1068.
[33] V. H. Weston, *Time-domain wave splitting of Maxwell's equations*, J. Math. Phys., 34 (1993), pp. 1370–1392.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 1987–2017

A NOISE CONTROL PROBLEM ARISING IN A FLOW DUCT*

PHILIPPE DESTUYNDER† AND JOCELYNE VÉTILLARD†

Abstract. Let us consider a one dimensional flow in a cylindrical tube, a portion of which is flexible. It can be a shell structure or a membrane equipped with actuators in order to reduce and if possible to cancel the acoustic perturbations transferred by the flow. The goal of this paper is to discuss the efficiency of such a noise insulator in a mathematical framework.

Key words. aeroacoustics, fluid-structure, control of noise, flow duct

AMS subject classifications. 35L20, 49J20, 49N05

DOI. 10.1137/S0036139902418912

1. Introduction. The modelling of aeroacoustic waves in an air flow coupled with a flexible structure is an old but difficult problem, which still contains a lot of unrevealed secrets [18], [17]. A cornerstone of such systems is the reflection of different wave speeds on the boundaries or at a geometrical discontinuity. Such phenomena can occur in a much more complex way at the interface between an air flow and a flexible structure. For instance, local waves like Stoneley waves [9] can appear and be acoustically predominant, at least locally [7]. More generally the question of local waves is a difficult problem in fluid-structure interactions, and the mathematical tools available are not yet well adapted to such analysis. The reason is that the approaches which make use of Fourier series or eigenmodes expansions are based on energetic approximations which do not represent correctly the local behavior of wave models, especially when the media are inhomogeneous. Phenomena occurring at a low frequency but with a very short wave length can certainly be better analyzed using Fourier transform (with respect to space coordinates). In such a formulation local behavior is transformed into global behavior. However, special technical strategies are necessary because Fourier transforms cannot be applied straightforwardly to inhomegeneous media. Another alternative consists of using special eigenmodes like the so-called Steklov ones. This is the approach used in this paper, but on a very restricted model. The difficulties mentioned here are magnified when a control model is considered. A lot of new mechanical phenomena arise and significantly disturb the certainties of the engineers. Therefore we think that it could be better to restrict our ambitions to a simple one dimensional model coupling a flexible structure with a fluid. Our goal is to prove that very strange mechanical behaviours can occur in such a class of aeroacoustic problems and to discuss the efficiency of a control system in order to suppress exactly any acoustic perturbation in the flow or in the structure. The first section gives a presentation of the model used and few useful properties. The flutter analysis is carried out in the third section, and the control problem is defined in the fourth one. Then a mathematical study of the adjoint state functions, performed in section five, enables one to prove a few controllability results in the sixth section. The method is an adaptation of the H.U.M. method of Lions [21]. The characterization of spaces for the control functions is performed in section 7 and is based on an idea

*Received by the editors November 29, 2002; accepted for publication (in revised form) November 11, 2003; published electronically August 19, 2004.
http://www.siam.org/journals/siap/64-6/41891.html
†Chaire de Calcul Scientifique, Conservatoire National des Arts et Métiers, 292 rue Saint Martin, Paris 75003 (destuynd@cnam.fr, jocelyne.vetillard@free.fr).

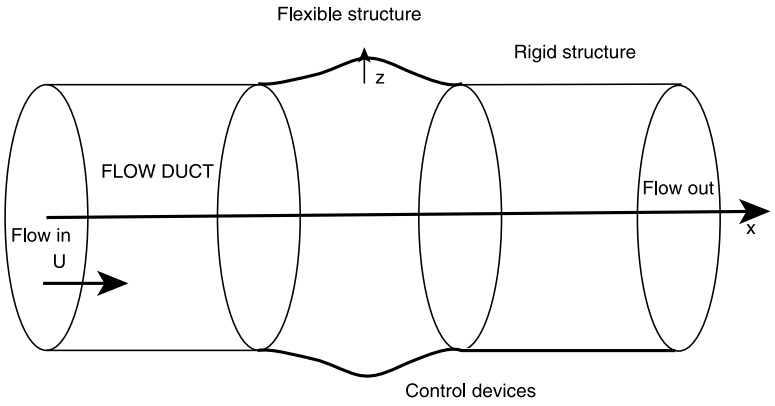


FIG. 1. *The flow duct and the flexible structure.*

introduced by Zuazua (quoted in [21]). The last part (section 8) suggests a discussion on the stability of the classical optimal control loop compared to the controllability results obtained previously.

2. The aeroacoustic model coupled with a flexible structure. We consider the steady flow of a fluid through a duct like that of Figure 1. A flow is travelling through it. It is assumed that this steady flow is uniform with the velocity U. The acoustic waves can be modelled by a potential function, say φ, which is a solution of the following partial differential equation, where c_f is the sound speed in the fluid:

$$\frac{\partial^2\varphi}{\partial t^2} + 2U\frac{\partial^2\varphi}{\partial x\partial t} + (U^2 - c_f^2)\frac{\partial^2\varphi}{\partial x^2} = ac_f^2\left(\frac{\partial z}{\partial t} + U\frac{\partial z}{\partial x}\right)\chi_{[L_1,L_2]}(x), \tag{1}$$

where z is the normal displacement of the flexible structure which is positioned between $x = L_1$ and $x = L_2$. (It is a wave equation but written in a moving frame at the velocity U ($x' = x - Ut$), and therefore we substitute the time derivative $\frac{\partial.}{\partial t}$ by $\frac{\partial.}{\partial t} + U\frac{\partial.}{\partial x}$.) The characteristic function of the segment $[L_1, L_2]$ is denoted by $\chi_{[L_1,L_2]}$, and a is a geometrical coefficient function of the cross section of the duct. The term $\frac{\partial z}{\partial t} + U\frac{\partial z}{\partial x}$ is due to the continuity of the normal velocity between the flexible structure and the wall of the duct (see Figure 2). More precisely, the term $U\frac{\partial z}{\partial x}$ comes from the rotation of the normal to the interface. Another kind of junction including a smart device will also be considered in this paper. It is obtained, for instance, by a direct control of the position of a part of the tube or by a prescribed acoustic pressure using skin loud speakers. This control is denoted by w. It is assumed for the sake of simplicity that its support is on a segment $]\alpha, \beta[$ inside the flexible structure, but its position could be anywhere along the flow duct. The full justification of such a model will be given in a forthcoming paper using the asymptotic method based on the small parameters representing the transverse dimensions of the duct and of the structure.

Then the new model is

$$\frac{\partial^2\varphi}{\partial t^2} + 2U\frac{\partial^2\varphi}{\partial x\partial t} + (U^2 - c_f^2)\frac{\partial^2\varphi}{\partial x^2} = ac_f^2\left(\frac{\partial z}{\partial t} + U\frac{\partial z}{\partial x}\right)\chi_{[L_1,L_2]}(x) + w\chi_{[\alpha,\beta]}(x). \tag{2}$$

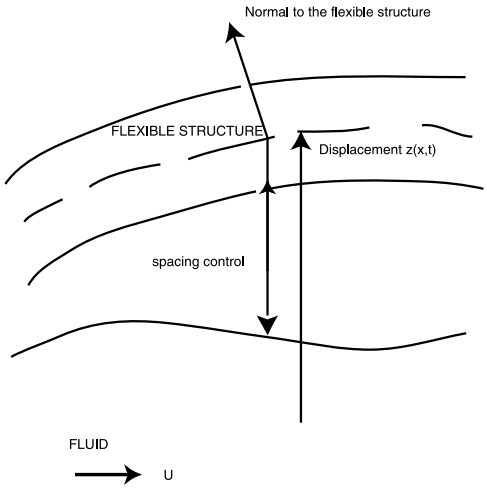


FIG. 2. *The smart system.*

The segment $[\alpha, \beta]$ can be quite small compared to the length of the structure. This is precisely one advantage of the system that we study: the mechanical effect of the control on the flow duct is extended to the whole length of the structure even if the control device is restricted to the subsegment $[\alpha, \beta]$. In fact, this control system is not necessary for proving the exact controllability of the coupled model, as we prove in the following. But it enables one to regularize the control functions used on the structure which is specified later on. Another point is that the boundary layer in the fluid is not taken into account in the model (1) or (2). It is true that the viscosity would contribute to changing the conclusions of the present study; however, it has been omitted in this paper for the sake of brevity and also because the parabolic behavior of the wall law is a real difficulty for the mathematical analysis [10]. At the extremities of the duct, the boundary conditions are not obvious at all from a mechanical point of view (in our opinion at least!). A first attempt in [5] was to prescribe homogeneous Dirichlet boundary conditions on the potential function φ. This facilitates the control analysis of the coupled system considerably, but it should be confessed that this is not a very realistic condition. A more physically founded one consists of prescribing the mass flow. Let us explain how it can be written mathematically. Let us first recall from classical fluid mechanics that the acoustic pressure p_a and the mass density variation ϱ_a are such that (ϱ_0 is the mass density of the steady flow)

$$p_a = c_f^2 \varrho_a = -\varrho_0 \left[\frac{\partial \varphi}{\partial t} + U \frac{\partial \varphi}{\partial x} \right]. \tag{3}$$

Let us denote by D the mass flow at the extremities of the duct. If S is the cross-section area, the variation of D at first order is

$$\delta D = S \left[\varrho_0 \frac{\partial \varphi}{\partial x} + U \varrho_a \right]. \tag{4}$$

Therefore, from (3) and (4), one deduces that (we set $M = U/c_f^2$, which is the Mach

number and is assumed to be smaller than 1 in all the text)

$$\delta D = -S\frac{U\varrho_0}{c_f^2}\left[\frac{\partial\varphi}{\partial t} + U\left(1-\frac{1}{M^2}\right)\frac{\partial\varphi}{\partial x}\right]. \tag{5}$$

Finally, the conservation of the mass flow at the extremities of the flow duct will be written as

$$\left[\frac{\partial\varphi}{\partial t} + U\left(1-\frac{1}{M^2}\right)\frac{\partial\varphi}{\partial x}\right](0,t) = \left[\frac{\partial\varphi}{\partial t} + U\left(1-\frac{1}{M^2}\right)\frac{\partial\varphi}{\partial x}\right](L,t) = 0 \qquad \forall t \in]0,T[. \tag{6}$$

From the mathematical point of view, the previous condition corresponds to the normal derivative with respect to the operator involved in (1) and (2). The function φ is defined up to an arbitrary constant (with respect to the coordinate x). This constant has no physical meaning, and we eliminate it by a convenient condition for the mathematical aspects of our analysis. Therefore we prescribe the additional condition that was not necessary for a Dirichlet condition:

$$\int_{L_1}^{L_2} \varphi = 0 \ \ \forall t \in]0,T[. \tag{7}$$

In the following it will be convenient to use the definition

$$H_m^1(]0,L[) = \left\{ \psi \in H^1(]0,L[),\ \int_{L_1}^{L_2} \psi = 0 \right\}. \tag{8}$$

Let us now introduce a variational formulation for (1) or (2), assuming that there is a smooth enough solution (it will be justified in the following). By multiplying (1) or (2) by an arbitrary element ψ in the space $H^1(]0,L[)$ and integrating by parts, we obtain, because of (6),

$$\begin{aligned} \forall \psi \in H^1(]0,L[), \qquad & \int_0^L \frac{\partial^2\varphi}{\partial t^2}\psi + U\int_0^L \left(\frac{\partial^2\varphi}{\partial x \partial t}\psi - \frac{\partial\varphi}{\partial t}\frac{\partial\psi}{\partial x}\right) + (c_f^2 - U^2)\int_0^L \frac{\partial\varphi}{\partial x}\frac{\partial\psi}{\partial x} \\ & = ac_f^2 \int_{L_1}^{L_2} \left(\frac{\partial z}{\partial t} + U\frac{\partial z}{\partial x}\right)\psi + \int_\alpha^\beta w\psi. \end{aligned} \tag{9}$$

Let us notice that if we choose $\psi = \frac{\partial\varphi}{\partial t}$ in (9), assuming again that there exists a solution which is smooth enough, one has the first energetic invariant relation

$$\begin{aligned} & \frac{\partial}{\partial t}\left[\frac{1}{2}\int_0^L \left(\frac{\partial\varphi}{\partial t}\right)^2 + \frac{c_f^2 - U^2}{2}\int_0^L \left(\frac{\partial\varphi}{\partial x}\right)^2\right] \\ & = ac_f^2 \int_{L_1}^{L_2} \frac{\partial z}{\partial t}\frac{\partial\varphi}{\partial t} + U\frac{\partial z}{\partial t}\frac{\partial\varphi}{\partial x} + aUc_f^2 \frac{\partial}{\partial t}\left[\int_{L_1}^{L_2} \frac{\partial z}{\partial x}\varphi\right] + \int_\alpha^\beta w\frac{\partial\varphi}{\partial t} \end{aligned} \tag{10}$$

or else

$$\begin{aligned} & \left[\frac{1}{2}\int_0^L \left(\frac{\partial\varphi}{\partial t}\right)^2 + \frac{c_f^2 - U^2}{2}\int_0^L \left(\frac{\partial\varphi}{\partial x}\right)^2 - aUc_f^2 \int_{L_1}^{L_2} \frac{\partial z}{\partial x}\varphi\right]_0^t \\ & = ac_f^2 \int_0^t \int_{L_1}^{L_2} \left(\frac{\partial\varphi}{\partial t} + U\frac{\partial\varphi}{\partial x}\right)\frac{\partial z}{\partial t} + \int_0^t \int_\alpha^\beta w\frac{\partial\varphi}{\partial t}. \end{aligned} \tag{11}$$

Concerning the flexible structure, one can consider that it is a membrane or a shell. In the first case, the normal displacement z is a solution of

$$(12)\qquad \begin{cases} \dfrac{\partial^2 z}{\partial t^2} - c_s^2 \dfrac{\partial^2 z}{\partial x^2} = -b\left(\dfrac{\partial \varphi}{\partial t} + U\dfrac{\partial \varphi}{\partial x}\right) + u\chi_{[\alpha,\beta]} \quad \forall (x,t) \in]L_1, L_2[\times]0,T[\,, \\ z(L_1,t) = z(L_2,t) = 0 \quad \forall t \in]0,T[\,, \end{cases}$$

where c_s is the wave speed in the structure, b is a geometrical coefficient which also involves the ratio between the mass density of the fluid and that of the structure, and, finally, u is the external control force applied on the segment $[\alpha,\beta] \subset [L_1, L_2]$ of the structure. This control is the most important in our study. In fact, it is sufficient for proving an exact controllability of the solution of the coupled system.

If the structure is a clamped shell, then z is solution of

$$(13)\qquad \begin{cases} \dfrac{\partial^2 z}{\partial t^2} + D\left[\dfrac{\partial^4 z}{\partial x^4} + \gamma^4 z\right] = -b\left(\dfrac{\partial \varphi}{\partial t} + U\dfrac{\partial \varphi}{x}\right) + u\chi_{[\alpha,\beta]} \quad \forall (x,t) \in]L_1, L_2[\times]0,T[\,, \\ z(L_1,t) = \dfrac{\partial z}{\partial x}(L_1,t) = z(L_2,t) = \dfrac{\partial z}{\partial x}(L_2,t) = 0 \quad \forall t \in]0,T[\,, \end{cases}$$

where D is the bending modulus of the shell in the direction x_1 and γ is the Batdorf coefficient of the shell. Here again if z is a smooth enough solution of (12) or (13), one has the following:

(a) for a membrane structure,

$$(14)\qquad \frac{\partial}{\partial t}\left[\frac{1}{2}\int_{L_1}^{L_2}\left(\frac{\partial z}{\partial t}\right)^2 + \frac{c_s^2}{2}\int_{L_1}^{L_2}\left(\frac{\partial z}{\partial x}\right)^2\right] = -b\int_{L_1}^{L_2}\left(\frac{\partial \varphi}{\partial t} + U\frac{\partial \varphi}{\partial x}\right)\frac{\partial z}{\partial t} + \int_\alpha^\beta u\frac{\partial z}{\partial t};$$

(b) for a shell structure,

$$(15)\qquad \frac{\partial}{\partial t}\left[\frac{1}{2}\int_{L_1}^{L_2}\left(\frac{\partial z}{\partial t}\right)^2 + \frac{D}{2}\int_{L_1}^{L_2}\left(\frac{\partial^2 z}{\partial x^2}\right)^2 + \gamma^4 z^2\right] = -b\int_{L_1}^{L_2}\left(\frac{\partial \varphi}{\partial t} + U\frac{\partial \varphi}{\partial x}\right)\frac{\partial z}{\partial t} + \int_\alpha^\beta u\frac{\partial z}{\partial t}.$$

It is possible to combine (11)–(14) or (11)–(15) in order to derive an energy invariance. Let us set

$$(16)\qquad X = (\varphi, z)(x,t)$$

and

$$(17)\qquad \begin{cases} a^m(X,X) = (c_f^2 - U^2)\displaystyle\int_0^L \left(\frac{\partial \varphi}{\partial x}\right)^2 + \frac{ac_f^2c_s^2}{b}\int_{L_1}^{L_2}\left(\frac{\partial z}{\partial x}\right)^2 - 2aUc_f^2\int_{L_1}^{L_2}\frac{\partial z}{\partial x}\varphi, \\ a^c(X,X) = (c_f^2 - U^2)\displaystyle\int_0^L \left(\frac{\partial \varphi}{\partial x}\right)^2 + \frac{ac_f^2c_s^2}{b}\int_{L_1}^{L_2}\left(\frac{\partial^2 z}{\partial x^2}\right)^2 - 2aUc_f^2\int_{L_1}^{L_2}\frac{\partial z}{\partial x}\varphi. \end{cases}$$

A solution of the coupled model (9)–(12) or (13) is such that (we set $a = a^m$ *or* a^c, depending on whether the structure is a membrane or a shell)

$$(18)\qquad \frac{\partial}{\partial t}\left[\frac{1}{2}\int_0^L \left(\frac{\partial \varphi}{\partial t}\right)^2 + \frac{ac_f^2}{2b}\int_{L_1}^{L_2}\left(\frac{\partial z}{\partial t}\right)^2 + \frac{1}{2}a(X,X)\right](t) = \frac{ac_f^2}{b}\int_\alpha^\beta u\frac{\partial z}{\partial t} + \int_\alpha^\beta w\frac{\partial \varphi}{\partial t}.$$

Let us now introduce a change of variables by setting

$$
\begin{aligned}
&X(x,t) = e^{\lambda t}\tilde{X}(x,t),\ \ X=(\varphi,z),\ \tilde{X}=(\tilde{\varphi},\tilde{z}), \qquad \text{where}\\
&\lambda\in R,\ (\tilde{\varphi},\tilde{z})\in H^1(]0,L[)\times H_0^1(]L_1,L_2[).
\end{aligned}
\tag{19}
$$

Then $\tilde{X}$ satisfies

$$
\begin{aligned}
&\frac{\partial}{\partial t}\left[e^{2\lambda t}\left(\int_0^L\left\{\left(\frac{\partial\tilde{\varphi}}{\partial t}\right)^2+2\lambda\tilde{\varphi}\frac{\partial\tilde{\varphi}}{\partial t}\right\}+\frac{ac_f^2}{b}\int_{L_1}^{L_2}\left\{\left(\frac{\partial\tilde{z}}{\partial t}\right)^2+2\lambda\tilde{z}\frac{\partial\tilde{z}}{\partial t}\right\}+\frac{ac_f^2}{b}+a_\lambda(\tilde{X},\tilde{X})\right)\right]\\
&=\frac{2ac_f^2}{b}\int_\alpha^\beta e^{\lambda t}u(x,t)\left(\frac{\partial\tilde{z}}{\partial t}+\lambda\tilde{z}\right)dx+2\int_\alpha^\beta e^{\lambda t}w(x,t)\left(\frac{\partial\tilde{\varphi}}{\partial t}+\lambda\tilde{\varphi}\right)dx,
\end{aligned}
\tag{20}
$$

where we have used the notation

$$
a_\lambda(\tilde{X},\tilde{X})=\lambda^2\left(\int_0^L\tilde{\varphi}^2+\frac{ac_f^2}{b}\int_{L_1}^{L_2}\tilde{z}^2\right)+a(\tilde{X},\tilde{X}).
\tag{21}
$$

It is useful to notice that, from physical arguments, the coefficients a and b are both strictly positive. The stability of a solution can therefore be discussed with $u=w=0\ \forall(x,t)\in]\alpha,\beta[\times]0,T[$. The basic point is to characterize the eigenvalues of the bilinear form $a(.,.)$. If all of them are positive, then the system is stable because the global energy remains bounded with respect to time. But if there exist negative eigenvalues, one can always choose λ large enough that the following quantity remains positive:

$$
\text{pseudoenergy}=\int_0^L\left\{\left(\frac{\partial\tilde{\varphi}}{\partial t}\right)^2+2\lambda\tilde{\varphi}\frac{\partial\tilde{\varphi}}{\partial t}\right\}+\frac{ac_f^2}{b}\int_{L_1}^{L_2}\left\{\left(\frac{\partial\tilde{z}}{\partial t}\right)^2+\lambda\tilde{z}\frac{\partial\tilde{z}}{\partial t}\right\}+a_\lambda(\tilde{X},\tilde{X}).
$$

The terms $\int_{L_1}^{L_2}\tilde{\varphi}\frac{\partial\tilde{\varphi}}{\partial t}$ and $\int_{L_1}^{L_2}\tilde{z}\frac{\partial\tilde{z}}{\partial t}$ can be bounded using the Cauchy–Schwarz inequality. For instance,

$$
\lambda\int_{L_1}^{L_2}\tilde{\varphi}\frac{\partial\tilde{\varphi}}{\partial t}\le\frac{\lambda^2}{2}\int_0^L(\tilde{\varphi})^2+\frac{1}{2}\int_0^L\left(\frac{\partial\tilde{\varphi}}{\partial t}\right)^2,
$$

and the same thing can be used for z. In fact, it will appear in the following that the generalized Steklov eigenvalue problem can be useful in characterizing the velocities U_c at which a flutter phenomenon (i.e., an unstability) can occur.

3. Existence, uniqueness, and stability of solutions.

3.1. Eigenmodes for the fluid. The first step is to define the eigenmodes of vibration for the fluid. Assuming that the structure is rigid, we set

$$
\begin{cases}
-(c_f^2-U^2)\dfrac{d^2\Phi^f}{dx^2}=\lambda^f\Phi^f,\ \ 0<x<L,\ \ \displaystyle\int_{L_1}^{L_2}\Phi^f=0,\\
\dfrac{d\Phi^f}{dx}(0)=\dfrac{d\Phi^f}{dx}(L)=0,\ \ \displaystyle\int_0^L|\Phi^f|^2=1.
\end{cases}
\tag{22}
$$

It is classical that in this particular case the solution (Φ_n^f,λ_n^f) can be computed analytically. Another useful set of functions for our analysis is the generalized Steklov

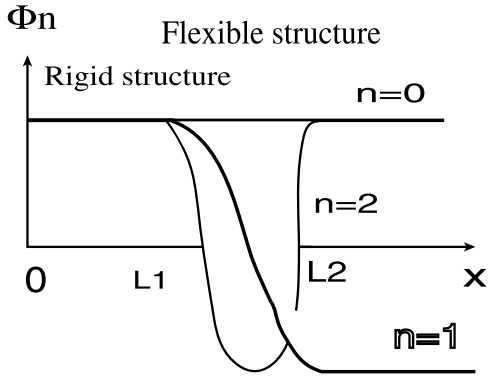


FIG. 3. *The generalized Steklov eigenmodes.*

one, which is defined as the solution of

$$\text{(23)}\qquad \begin{cases} -(c_f^2 - U^2)\dfrac{d^2\Phi^{sk}}{dx^2} = \lambda^{sk}\Phi^{sk}\chi_{[L_1,L_2]}, \;\; 0 < x < L, \\ \dfrac{d\Phi^{sk}}{dx}(0) = \dfrac{d\Phi^{sk}}{dx}(L) = 0, \;\; \displaystyle\int_{L_1}^{L_2} |\Phi^{sk}|^2 = 1, \; \Phi_n^{sk} \in H_m^1(]0, L[), \end{cases}$$

where $\chi_{[L_1,L_2]}$ is the characteristic function of the segment $[L_1, L_2]$. Here again the eigenmodes can be computed analytically (see Figure 3). It is interesting for our purpose to give their expressions:

$$\text{(24)}\qquad \begin{cases} \Phi_n^{sk}(x) = \pm A, \; 0 < x < L_1, \; L_2 < x < L, \\ \Phi_n^{sk}(x) = A\cos\left(\dfrac{n\pi(x - L_1)}{L_2 - L_1}\right), \;\; A = \sqrt{\dfrac{2}{L_2 - L_1}}, \\ \lambda_n^{sk} = \dfrac{n^2\pi^2(c_f^2 - U^2)}{(L_2 - L_1)^2}. \end{cases}$$

An important property for our purpose is the next one:

$$\forall n \geq 1, \qquad \int_{L_1}^{L_2} \Phi_n^{sk}(x)dx = 0.$$

Let us now consider a function φ in the space $H^1(]0, L[)$. We associate the new function $\delta\varphi$ defined by

$$\text{(25)}\qquad \delta\varphi = \varphi - \frac{1}{L_2 - L_1}\int_{L_1}^{L_2} \varphi(x)dx.$$

The second term on the right-hand side is a constant function on $]0, L[$. Furthermore, we have the property

$$\int_{L_1}^{L_2} \delta\varphi = 0, \quad \text{and hence} \quad \delta\varphi \in H_m^1(]0, L[),$$

which proves that $\delta\varphi$ is orthogonal to constant functions in the space $L^2(]L_1, L_2[)$. Therefore, from the min-max theorem applied to the generalized Steklov problem, one obtains (let us recall that we assumed that $U < c_f$!)

$$\forall \varphi \in H^1(]0, L[),\ (c_f^2 - U^2)\int_0^L \left|\frac{d\varphi}{dx}\right|^2 = (c_f^2 - U^2)\int_{L_1}^{L_2} \left|\frac{d\delta\varphi}{dx}\right|^2 \geq \lambda_1^{sk}\int_{L_1}^{L_2} |\delta\varphi|^2. \tag{26}$$

Furthermore,

$$\begin{cases} \forall z \in H_0^1(]L_1, L_2[),\ \dfrac{ac_s^2c_f^2}{b}\displaystyle\int_{L_1}^{L_2}\left|\frac{dz}{dx}\right|^2 - 2aUc_f^2\int_{L_1}^{L_2}\frac{dz}{dx}\varphi \\ \qquad\qquad = \dfrac{ac_f^2c_s^2}{b}\left[\displaystyle\int_{L_1}^{L_2}\left(\frac{dz}{dx} - \frac{Ub}{c_s^2}\delta\varphi\right)^2 - \frac{U^2b^2}{c_s^4}\int_{L_1}^{L_2}|\delta\varphi|^2\right]. \end{cases} \tag{27}$$

Therefore ($\delta\varphi$ is defined from φ by (25)),

$$\begin{cases} \forall X = (\varphi, z) \in H^1(]0, L[) \times H_0^1(]L_1, L_2[), \\ a^m(X, X) \geq \left[\lambda_1^{sk} - \dfrac{abc_f^2U^2}{c_s^2}\right]\displaystyle\int_{L_1}^{L_2}|\delta\varphi|^2 + \frac{ac_f^2c_s^2}{b}\int_{L_1}^{L_2}\left(\frac{\partial z}{\partial x} - \frac{Ub}{c_s^2}\delta\varphi\right)^2. \end{cases} \tag{28}$$

Finally, if the following condition is satisfied (we set $\eta_1^{sk}(c_f^2 - U^2) = \lambda_1^{sk}$),

$$(c_f^2 - U^2)\eta_1^{sk} > abU^2\left(\frac{c_f}{c_s}\right)^2,$$

or else (U_c is called a critical velocity for the steady flow)

$$U < U_c = \frac{c_f}{\sqrt{1 + \dfrac{ab}{\eta_1^{sk}}\left(\dfrac{c_f}{c_s}\right)^2}} < c_f, \tag{29}$$

then one obtains from (18)

$$a^m(X, X) \geq c_0[|\varphi|_{1,0L}^2 + ||z||_{1,L_1L_2}^2],$$

and therefore there exists a positive constant c_1 such that

$$\begin{aligned} &\left(\left|\left|\frac{\partial\varphi}{\partial t}\right|\right|_{0,0L}^2 + \left|\left|\frac{\partial z}{\partial t}\right|\right|_{0,0L}^2 + |\varphi|_{1,0L}^2 + ||z||_{1,L_1L_2}^2\right)(t) \\ &\leq c_1\left[\left(\left|\left|\frac{\partial\varphi}{\partial t}\right|\right|_{0,0L}^2 + \left|\left|\frac{\partial z}{\partial t}\right|\right|_{0,0L}^2 + |\varphi|_{1,0L}^2 + ||z||_{1,L_1L_2}^2\right)(0) + \int_0^t\int_\alpha^\beta u^2(x,\xi)dxd\xi\right]. \end{aligned} \tag{30}$$

Remark 1. From the expression of the Steklov eigenvalue λ_1^{sk}, we deduce that

$$\eta_1^{sk} = \frac{\pi^2}{(L_2 - L_1)^2},$$

that $\lambda_1^{sk} \to \infty$ when $|L_2 - L_1| \to 0$, and therefore, from (29), $U_c \to c_f$. Conversely, when $|L_2 - L_1|\sqrt{ab} \to \infty$, one has the asymptotic behavior for the critical value U_c:

$$U_c \simeq c_s\frac{\pi}{|L_2 - L_1|}\frac{1}{\sqrt{ab}}.$$

Remark 2. For a shell structure, one has

$$\forall z \in H_0^2(]L_1, L_2[), \qquad \frac{Dac_f^2}{b}\left[\int_{L_1}^{L_2}\left|\frac{d^2z}{dx^2}\right|^2 + \gamma^4 z^2\right] \geq \frac{Dac_f^2\pi^2}{b(L_2-L_1)^2}\int_{L_1}^{L_2}\left|\frac{dz}{dx}\right|^2,$$

and thus similar results can be obtained. For instance, one obtains the new critical value for the steady flow velocity:

$$U_c = \frac{c_f}{\sqrt{1 + \dfrac{abc_f^2(L_2-L_1)^2}{D\pi^2}}}. \tag{31}$$

3.2. The eigenmodes for the flexible structure. Let us introduce now the eigenvectors of the flexible structure (for a membrane, for instance), which can be explicited analytically:

$$\begin{cases} -c_s^2\dfrac{d^2Z^s}{dx^2} = \lambda^s Z^s, \quad L_1 < x < L_2, \\ Z^s(L_1) = Z^s(L_2) = 0, \quad \displaystyle\int_{L_1}^{L_2}|Z^s|^2 = 1. \end{cases} \tag{32}$$

The solutions are

$$Z_n^s(x) = \sqrt{\frac{2}{L_2-L_1}}\sin\left(\frac{n\pi(x-L_1)}{L_2-L_1}\right). \tag{33}$$

A similar system can be explicated for a shell structure.

3.3. Existence and uniqueness of solutions for $U < c_f$. In order to simplify the presentation, we restrict the analysis to the case of a membrane structure, but it could be extended to a shell. Let us introduce the following approximation spaces:

$$\begin{cases} V^N = \left\{\varphi = \displaystyle\sum_{1\leq n\leq N}\alpha_n\Phi_n^f\right\}, \\ Z^N = \left\{z = \displaystyle\sum_{1\leq n\leq N}\beta_n Z_n^s\right\}. \end{cases} \tag{34}$$

Then the approximate solution $(\varphi^N, z^N) \in C^1([0,T]; V^N \times Z^N)$ is defined as the unique solution of the finite dimensional differential equation

$$\begin{cases} \forall\psi \in V^N, & \displaystyle\int_0^L \frac{\partial^2\varphi^N}{\partial t^2}\psi + U\int_0^L\left(\frac{\partial^2\varphi^N}{\partial x\partial t}\psi - \frac{\partial\varphi^N}{\partial t}\frac{\partial\psi}{\partial x}\right) \\ & \qquad + (c_f^2 - U^2)\displaystyle\int_0^L\frac{\partial\varphi^N}{\partial x}\frac{\partial\psi}{\partial x} \\ & \qquad = ac_f^2\displaystyle\int_{L_1}^{L_2}\left(\frac{\partial z^N}{\partial t} + U\frac{\partial z^N}{\partial x}\right)\psi + \int_\alpha^\beta w\psi, \\ \forall v \in Z^N, & \displaystyle\int_{L_1}^{L_2}\frac{\partial^2 z^N}{\partial t^2}v + c_s^2\int_{L_1}^{L_2}\frac{\partial z^N}{\partial x}\frac{\partial v}{\partial x} \\ & \qquad = -b\displaystyle\int_{L_1}^{L_2}\left(\frac{\partial\varphi^N}{\partial t} + U\frac{\partial\varphi^N}{\partial x}\right)v + \int_\alpha^\beta uv. \end{cases} \tag{35}$$

The initial conditions satisfied by (φ^N, z^N) are defined from those of the continuous model by taking the L^2 projection of both (φ^N, z^N) and $(\frac{\partial \varphi^N}{\partial t}, \frac{\partial z^N}{\partial t})$. The existence and uniqueness theorem is then derived from a priori energy estimates by setting $\psi = \frac{\partial \varphi^N}{\partial t}$ and $v = \frac{\partial z^N}{\partial t}$ in (35) when $U < U_c$. Then the weak convergence of a subsequence to a weak solution of the variational coupled model can be proved from classical strategies. When $U_c < U < c_f$, a spectrum translation can be used as we mentioned in (19)–(20). But even if existence and uniqueness are still true, an unstability can appear. For details concerning the existence and uniqueness theorem, we refer to the paper [6]. In order to characterize more precisely the unstabilities which can occur for $U > U_c$ (U_c is only a lower bound), let us consider the particular case where $L_1 = 0$ and $L_2 = L$. In such a configuration the Steklov basis Φ_n^{sk} is identical to the one used for the fluid Φ_n^f. The solution (φ, z) of the coupled model can be expanded in the basis of the eigenmodes of the fluid, on the one hand, and of the structure, on the other hand. Therefore we set (because $\{\Phi_n^f\}$ and $\{Z_n^s\}$ are Hilbert bases of, respectively, $H^1(]0, L[)$ and $H_0^1(]0, L[)$)

$$\left\{ \begin{array}{l} \varphi(x,t) = \displaystyle\sum_{n\geq 1} \alpha_n(t)\Phi_n^f(x), \\ \\ z(x,t) = \displaystyle\sum_{n\geq 1} \beta_n(t) Z_n^s(x), \end{array} \right. \tag{36}$$

where we recall that

$$\Phi_n^f(x) = \sqrt{\frac{2}{L}} \cos\left(\frac{n\pi x}{L}\right), \qquad Z_n^s(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right).$$

Introducing these expressions of φ and z into the variational formulation, we obtain (the control functions u and w are supposed to be zero and one can observe that the second term in the fluid equation disappears)

$$\left\{ \begin{array}{l} \dfrac{\partial^2 \alpha_n}{\partial t^2} + \mu_n^2(c_f^2 - U^2)\alpha_n - a c_f^2 U \mu_n \beta_n = 0, \\ \\ \dfrac{\partial^2 \beta_n}{\partial t^2} + \mu_n^2 c_s^2 \beta_n - bU\mu_n \alpha_n = 0 \end{array} \right. \tag{37}$$

(we set $\mu_n = \frac{n\pi}{L}$ for the sake of brevity). The solutions $(\alpha_n, \beta_n)(t)$ can be analyzed from the eigenvalues of the *stiffness matrix* of (37). Let us denote them by λ. They are the solutions of the characteristic equation of (37):

$$\lambda^2 + \lambda\mu_n^2(c_s^2 + c_f^2 - U^2) + \mu_n^4(c_f^2 - U^2)c_s^2 - ab\mu_n^2 U^2 c_f^2 = 0. \tag{38}$$

If $\lambda \in R$, then one can check directly that $\lambda > 0$ (see the product and the sum of the roots). Then the solutions (α_n, β_n) are stable (i.e., sine and cosine functions of time with pulsation equal to $\sqrt{\lambda}$). But if λ is a complex number, the solutions of (37) can be exponentially increasing with respect to time. In fact, using an eigenvector basis (assuming that the two eigenvalues are distinct), (37) can also be written as

follows:

$$\frac{\partial^2 \xi_n}{\partial t^2} + (d_0 \pm i d_1)\xi_n = 0, \tag{39}$$

where $d_0 \pm i d_1$ are the roots of (38). Thus, whatever is the sign of d_1, one of the two roots of the characteristic equations of (37) leads to an exponentially increasing solution with respect to time. The point is to characterize the critical velocity U_f at which the discriminant of (38) is zero, because for $U = 0$ the two roots are real and positive, and their product remains also positive. One obtains

$$U^4\mu_n^4 - 2\mu_n^2 U^2(\mu_n^2(c_f^2 - c_s^2) - 2abc_f^2) + \mu_n^4(c_f^2 - c_s^2)^2 = 0. \tag{40}$$

There exist real and positive roots in U to this equation if and only if

$$\begin{cases} \text{(i) } \mu_n^2(c_f^2 - c_s^2) - abc_f^2 < 0 & \text{(real roots)}, \\ \text{(ii) } \mu_n^2(c_f^2 - c_s^2) - 2abc_f^2 < 0 & \text{(positivity of the real parts of the roots)}. \end{cases} \tag{41}$$

The first condition implies the second one. Let us recall that we assumed that $U < c_f$ and for physical reasons the numbers a and b are both positive. Therefore the flutter phenomenon occurs if and only if $c_s < c_f$ and for the values of n (if there are any!) such that

$$\mu_n < \frac{\sqrt{ab}}{\sqrt{1 - \left(\dfrac{c_s}{c_f}\right)^2}}. \tag{42}$$

Hence, only a finite number of values of n lead to a flutter mechanism. All of them correspond to homogeneous boundary conditions at the extremities of the tube ($x = 0$ and $x = L$). However, the acoustic pressure is not zero at these points. More realistic cases (i.e., $0 < L_1 < L_2 < L$) can be treated numerically using a finite element method, but unfortunately, not analytically. From a mechanical point of view, the flutter phenomenon is a coupling between two eigenmodes which have the same frequencies. One captures the energy from the steady flow, and the other stores it. In fact, the flutter mechanism is generally destructive and thus avoided. Nevertheless it can be interesting to use the phenomenon in order to increase the efficiency of the control system that we discuss in the following. The principle would be to transfer the acoustic energy from the flow into the structure and to kill it there. But clearly such a strategy can be catastrophic whenever the control system fails. It is worth noting that the condition $c_s < c_f$ is precisely the same as that which enables the existence of local waves in two dimensional or three dimensional fluid-structure interaction. This surprising aspect has been discussed in [11]. Concerning the Stoneley waves in inhomogeneous structures, one can refer to Stoneley [27], Cagniard [1], or Fung [9]. From a mechanical point of view the existence of local waves enables us to store energy in a close neighborhood of the structure. This is also what happens in the flutter mechanism.

In the next section we discuss the control law for a steady flow velocity which is smaller than the critical value U_c. Using the change of unknowns defined at (19), one can extend the result to the more general case: $U_c \leq U < c_f$. Unfortunately the analysis that has been developed in this paper cannot be extended to the supersonic case.

4. The control problem. Let us consider a system of initial conditions for the coupled model

$$\text{(43)}\qquad \begin{cases} \varphi(x,0)=\varphi_0(x), \quad \dfrac{\partial\varphi}{\partial t}(x,0)=\varphi_1(x) & \forall x\in\,]0,L[\,,\\ z(x,0)=z_0(x), \quad \dfrac{\partial z}{\partial t}(x,0)=z_1(x) & \forall x\in\,]L_1,L_2[\,, \end{cases}$$

Furthermore we assume that

$$\varphi_0\in H^1(]0,L[),\ \varphi_1\in L^2(]0,L[),\ z_0\in H_0^1(]L_1,L_2[),\ z_1\in L^2(]L_1,L_2[).$$

Then the solution of the coupled model is such that (it is assumed that $(u,w)\in [L^2(\,]\alpha,\beta[\times]0,T[\,)]^2$)

$$\text{(44)}\qquad \begin{cases} \varphi\in C^0([0,T];H^1(]0,L[))\cap C^1([0,T];L^2(]0,L[)),\\ z\in C^0([0,T];H_0^1(]L_1,L_2[))\cap C^1([0,T];L^2(]L_1,L_2[)). \end{cases}$$

Let us now define the following criterion as a function of the control variables $(u,w)\in[L^2(\,]\alpha,\beta[\times]0,T[\,)]^2$, where A,B,C,D and ε are positive constants:

$$\text{(45)}\qquad \begin{cases} J^\varepsilon(u,w)=\dfrac{A}{2}\displaystyle\int_0^L\left|\frac{\partial\varphi}{\partial t}(x,T)\right|^2+\frac{B}{2}\int_0^L\left|\frac{\partial\varphi}{\partial x}(x,T)\right|^2\\ \qquad\quad +\dfrac{C}{2}\displaystyle\int_{L_1}^{L_2}\left|\frac{\partial z}{\partial t}(x,T)\right|^2+\frac{D}{2}\int_{L_1}^{L_2}\left|\frac{\partial z}{\partial x}(x,T)\right|^2+\frac{\varepsilon}{2}\int_0^T\int_\alpha^\beta(u^2+w^2)(x,t). \end{cases}$$

Let us point out that the two control variables have been located at the same place. This is not necessary for our analysis, but it enables a few helpful simplifications. The existence and uniqueness of a solution to the next optimization problem is clear because of the strict convexity, the continuity, and the coerciveness of J^ε :

$$\text{(46)}\qquad \begin{cases} \min\ \ J^\varepsilon(u,w),\\ (u,w)\in[L^2(]\alpha,\beta[\times]0,T[)]^2. \end{cases}$$

In order to characterize the solution of (46), it is very convenient (in fact, it is necessary) to introduce the adjoint state function—say, ψ and d, which are the Lagrange multipliers of the fluid and of the structural equation, respectively. They are the solution of

$$\text{(47)}\qquad \begin{cases} \dfrac{\partial^2\psi}{\partial t^2}+2U\dfrac{\partial^2\psi}{\partial x\partial t}+(U^2-c_f^2)\dfrac{\partial^2\psi}{\partial x^2}\\ \qquad =b\left(\dfrac{\partial d}{\partial t}+U\dfrac{\partial d}{\partial x}\right)\chi_{[L_1,L_2]}(x)\quad \forall(x,t)\in\,]0,L[\times]0,T[\,,\\ \left[\dfrac{\partial\psi}{\partial t}+U\left(1-\dfrac{1}{M^2}\right)\dfrac{\partial\psi}{\partial x}\right](0,t)=\left[\dfrac{\partial\psi}{\partial t}+U\left(1-\dfrac{1}{M^2}\right)\dfrac{\partial\psi}{\partial x}\right](L,t)=0\quad \forall t\in\,]0,T[\,,\\ \dfrac{\partial^2 d}{\partial t^2}-c_s^2\dfrac{\partial^2 d}{\partial x^2}=-ac_f^2\left(\dfrac{\partial\psi}{\partial t}+U\dfrac{\partial\psi}{\partial x}\right)\quad \forall(x,t)\in\,]L_1,L_2[\times]0,T[\,,\\ d(L_1,t)=d(L_2,t)=0\ \forall t\in\,]0,T[\,, \end{cases}$$

and the final conditions are adjusted such that the gradient of J^ε could be represented

simply in the space $[L^2(]\alpha,\beta[\times]0,T[)]^2$. Thus we set

$$
\left\{\begin{array}{l}
d(x,T) = C\dfrac{\partial z}{\partial t}(x,T)\ \forall x \in]L_1,L_2[\,,\\
\dfrac{\partial d}{\partial t}(x,T) = -ac_f^2 A\dfrac{\partial \varphi}{\partial t}(x,T) - Dz(x,T)\ \forall x \in]L_1,L_2[\,,\\
\psi(x,T) = A\dfrac{\partial \varphi}{\partial t}(x,T)\ \forall x \in]0,L[\,,\\
\dfrac{\partial \psi}{\partial t}(x,T) = -2AU\dfrac{\partial^2 \varphi}{\partial x \partial t}(x,T) + bC\dfrac{\partial z}{\partial t}(x,T)\chi_{[L_1,L_2]} + B\dfrac{\partial^2 \varphi}{\partial x^2}(x,T)\\
\qquad\qquad + \left(AU + \dfrac{BM^2}{U(M^2-1)}\right)\left[\dfrac{\partial \varphi}{\partial t}(L,T)\delta_L(x) - \dfrac{\partial \varphi}{\partial t}(0,T)\delta_0(x)\right].
\end{array}\right. \tag{48}
$$

It is worth noting that if

$$
B = A(c_f^2 - U^2) \tag{49}
$$

(where $M = \frac{U}{c_f} < 1$), then the Dirac distributions which appear in the expression of $\frac{\partial \psi}{\partial t}(x,T)$ are canceled. Anyway, a simple computation leads to the following expressions for the derivatives of J^ε with respect to the control functions u and w:

$$
\left\{\begin{array}{l}
\dfrac{\partial J^\varepsilon}{\partial u}(u,w) = d + \varepsilon u \quad \text{on }]\alpha,\beta[\times]0,T[\,,\\[2ex]
\dfrac{\partial J^\varepsilon}{\partial w}(u,w) = \psi + \varepsilon w \quad \text{on }]\alpha,\beta[\times]0,T[\,.
\end{array}\right. \tag{50}
$$

The existence and uniqueness of a solution to the dual system (47)–(48) are not obvious, because the initial data for the primal model are not smooth enough. Usually one can use a series of eigenvectors in order to characterize the solution of a time dependent system (cf. Lions [20]). In our case this method cannot be applied because there is no spectral theorem for this coupled model. This is due to a loss of symmetry in the coupling terms and also to the first order derivatives with respect to the time variable. But another strategy can be used. We set

$$
\left\{\begin{array}{l}
\tilde{\psi}(x,t) = \tilde{\psi}_0(x) + \displaystyle\int_T^t \psi(x,s)ds,\\[2ex]
\tilde{d}(x,t) = \tilde{d}_0(x) + \displaystyle\int_T^t d(x,s)ds,
\end{array}\right. \tag{51}
$$

and from a quite classical computation we obtain that $(\tilde{\psi},\tilde{d})$ is a solution of

$$
\left\{\begin{array}{l}
\dfrac{\partial^2 \tilde{\psi}}{\partial t^2} + 2U\dfrac{\partial^2 \tilde{\psi}}{\partial x \partial t} + (U^2 - c_f^2)\dfrac{\partial^2 \tilde{\psi}}{\partial x^2}\\
\qquad = b\left(\dfrac{\partial \tilde{d}}{\partial t} + U\dfrac{\partial \tilde{d}}{\partial x}\right)\chi_{[L_1,L_2]}(x) \quad \forall (x,t) \in]0,L[\times]0,T[\,,\\
\left[\dfrac{\partial \tilde{\psi}}{\partial t} + U\left(1 - \dfrac{1}{M^2}\right)\dfrac{\partial \tilde{\psi}}{\partial x}\right](0,t) = \left[\dfrac{\partial \tilde{\psi}}{\partial t} + U\left(1 - \dfrac{1}{M^2}\right)\dfrac{\partial \tilde{\psi}}{\partial x}\right](L,t) = 0\ \forall t \in]0,T[\,,\\
\dfrac{\partial^2 \tilde{d}}{\partial t^2} - c_s^2\dfrac{\partial^2 \tilde{d}}{\partial x^2} = -ac_f^2\left(\dfrac{\partial \tilde{\psi}}{\partial t} + U\dfrac{\partial \tilde{\psi}}{\partial x}\right)\ \forall (x,t) \in]L_1,L_2[\times]0,T[\,,\\
\tilde{d}(L_1,t) = \tilde{d}(L_2,t) = 0\ \forall t \in]0,T[\,,
\end{array}\right. \tag{52}
$$

and the boundary conditions are

$$
(53)\quad \begin{cases} \tilde{d}(L_1,t)=\tilde{d}(L_2,t)=0 \quad \forall t\in\,]0,T[\,, \\ \left(\dfrac{\partial\tilde{\psi}}{\partial t}+U\left(1-\dfrac{1}{M^2}\right)\dfrac{\partial\tilde{\psi}}{\partial x}\right)(x,t) \\ \qquad\qquad =\psi(x,T)+U\left(1-\dfrac{1}{M^2}\right)\dfrac{\partial\tilde{\psi}_0}{\partial x}(x) \quad \forall (x,t)\in\{0,L\}\times\,]0,T[\,. \end{cases}
$$

The functions $(\tilde{\psi}_0,\tilde{d}_0)$ are chosen such that the right-hand sides of (52)–(53) are zero. Hence we set

$$
(54)\quad \begin{cases} -(c_f^2-U^2)\dfrac{\partial^2\tilde{\psi}_0}{\partial x^2}-Ub\dfrac{\partial\tilde{d}_0}{\partial x}=-\left(\dfrac{\partial\psi}{\partial t}(x,T)+2U\dfrac{\partial\psi}{\partial x}(x,T)-bd(x,T)\right), \\ -c_s^2\dfrac{\partial^2\tilde{d}_0}{\partial x^2}+aUc_f^2\dfrac{\partial\tilde{\psi}_0}{\partial x}=-\left(\dfrac{\partial d}{\partial t}(x,T)+ac_f^2\psi(x,T)\right), \\ \tilde{d}_0\in H_0^1(]L_1,L_2[), \qquad \tilde{\psi}_0\in H^1(]0,L[), \\ \dfrac{\partial\tilde{\psi}_0}{\partial x}(x)=\dfrac{M^2}{U(1-M^2)}\psi(x,T)=A\dfrac{\partial\varphi}{\partial x}(x,T)\ \forall x\in\{0,L\}. \end{cases}
$$

The final conditions that should be satisfied by $(\tilde{\psi},\tilde{d})$ are

$$
(55)\quad \begin{cases} \tilde{\psi}(x,T)=\tilde{\psi}_0(x)\ \forall x\in\,]0,L[\,, \\ \dfrac{\partial\tilde{\psi}}{\partial t}(x,T)=\psi(x,T)=A\dfrac{\partial\varphi}{\partial x}(x,T)\ \forall x\in\,]0,L[\,, \end{cases}
$$

$$
(56)\quad \begin{cases} \tilde{d}(x,T)=\tilde{d}_0(x)\ \forall x\in\,]L_1,L_2[\,, \\ \dfrac{\partial\tilde{d}}{\partial t}(x,T)=d(x,T)=C\dfrac{\partial z}{\partial t}(x,T)\ \forall x\in\,]L_1,L_2[\,. \end{cases}
$$

The existence and uniqueness of $(\tilde{\psi},\tilde{d})$ can then be obtained by the same result as that used for (φ,z) (i.e., the primal solution). Furthermore, one can check directly that (ψ,d), obtained by taking the time derivative of $(\tilde{\psi},\tilde{d})$, is the solution of the dual model (47)–(48). The only point to be checked is the existence and uniqueness of $(\tilde{\psi}_0,\tilde{d}_0)$. Let us therefore first introduce the notation

$$
X_0=(\tilde{\psi}_0,\tilde{d}_0)\in H^1(]0,L[)\times H_0^1(]L_1,L_2[),
$$

and then the bilinear form

$$
(57)\quad a_0(X_0,X_0)=(c_f^2-U^2)\int_0^L\left(\frac{\partial\tilde{\psi}_0}{\partial x}\right)^2+\frac{bc_s^2}{ac_f^2}\int_{L_1}^{L_2}\left(\frac{\partial\tilde{d}_0}{\partial x}\right)^2-2bU\int_{L_1}^{L_2}\left(\frac{\partial\tilde{d}_0}{\partial x}\right)\tilde{\psi}_0,
$$

and finally the linear form $l_0(.)$ defined by

$$
(58)\quad \begin{cases} l_0(X_0)=A(c_f^2-U^2)\displaystyle\int_0^L\frac{\partial\varphi}{\partial x}(x,T)\frac{\partial\tilde{\psi}_0}{\partial x}(x)-2AU\int_0^L\frac{\partial\varphi}{\partial t}(x,T)\frac{\tilde{\psi}_0}{\partial x}(x) \\ \qquad\qquad +bC\displaystyle\int_{L_1}^{L_2}\int_{L_1}^{L_2}\frac{\partial z}{\partial t}(x,T)\tilde{\psi}_0(x). \end{cases}
$$

First, $a_0(.,.)$ and $l_0(.)$ are respectively bilinear and linear. Furthermore, they are both continuous on the space $H^1(]0,L[)\times H_0^1(]L_1,L_2[)$ as soon as (φ,z), which appears in

the definition of $l_0(.)$, is a solution of the primal model with a finite energy initial condition. The last point to be checked in order to apply the Lax–Milgram theorem concerns the coerciveness of $a_0(.,.)$. In fact, it has already been proved in (28) when $U < U_c$. For $U > U_c$ one can use the Garding inequality based on a compactness argument [15]. Let us now discuss the regularity of the solution (ψ, d) and/or that of (φ, z). It is dependent on the initial conditions. If the energy of the initial condition is finite, then the solution (φ, z) is in the space

$$C^0([0,T]; H^1(]0,L[) \times H_0^1(]L_1,L_2[)) \cap C^1([0,T]; L^2(]0,L[) \times L^2(]L_1,L_2[)).$$

But taking the time or space derivatives of (φ, z) (and also of (ψ, d)), the same regularity can be obtained if the initial conditions and the control functions are smooth enough. Unfortunately a restriction appears again for the space derivatives because at $x = L_1$ and $x = L_2$ the functions z (and d) are continuous, but this is not true for the first order derivative. Hence one can only apply the previous method to first order derivatives with respect to the space coordinate. Nevertheless this is a sufficient regularity for our purpose in the multiplier method that we use in the next section.

5. Energy estimates on the adjoint state. Several a priori estimates can be obtained from the multiplier method of Lions [21]. Then from several additional tricks introduced by Lions [21] and Zuazua [31] we can derive local estimates which are useful for characterizing the initial data, and which can be exactly controlled with control functions in the space $L^2(]\alpha,\beta[\times]0,T[)$. All the following computations are performed assuming that the dual fields (ψ, d) are smooth enough. This is possible because of the regularity results mentioned previously. Then the inequalities obtained are extended by a density argument. The method being quite standard, we refer to the book of Lions [21] for the details.

5.1. Lagrangian invariants. Let us multiply (47) by $\frac{\partial \psi}{\partial t}$ and $\frac{\partial d}{\partial t}$. Then by integrating with respect to time from 0 to T and with respect to the space coordinate x, one obtains the following identities:

$$\epsilon(t) = \epsilon(0) \quad \forall t \in [0,T], \tag{59}$$

where ϵ is a pseudoenergy defined by

$$\left\{ \begin{array}{l} \epsilon(t) = \dfrac{1}{2}\displaystyle\int_0^L \left(\frac{\partial \psi}{\partial t}\right)^2 + \frac{b}{2ac_f^2}\int_{L_1}^{L_2} \left(\frac{\partial d}{\partial t}\right)^2 \\ \qquad + \dfrac{(c_f^2 - U^2)}{2}\displaystyle\int_0^L \left(\frac{\partial \psi}{\partial x}\right)^2 + \frac{bc_s^2}{2ac_f^2}\int_{L_1}^{L_2} \left(\frac{\partial d}{\partial x}\right)^2 - bU\int_{L_1}^{L_2} \frac{\partial d}{\partial x}\psi. \end{array} \right. \tag{60}$$

Then from (28), and if $U < U_c$, there exist two positive constants—say, c_0 and c_1—such that

$$c_0 E(t) \le \epsilon(t) \le c_1 E(t), \tag{61}$$

where $E(t)$ is the global energy of the fluid and the structure (but without the coupling term) and which is defined by

$$\left\{ \begin{array}{l} E(t) = \dfrac{1}{2}\displaystyle\int_0^L \left(\frac{\partial \psi}{\partial t}\right)^2 + \frac{b}{2ac_f^2}\int_{L_1}^{L_2} \left(\frac{\partial d}{\partial t}\right)^2 \\ \qquad + \dfrac{(c_f^2 - U^2)}{2}\displaystyle\int_0^L \left(\frac{\partial \psi}{\partial x}\right)^2 + \frac{bc_s^2}{2ac_f^2}\int_{L_1}^{L_2} \left(\frac{\partial d}{\partial x}\right)^2. \end{array} \right. \tag{62}$$

The difference between $\epsilon(t)$ and $E(t)$ is the coupling term, which is connected to the transfer of energy from the steady flow to the flexible structure through the compressible flow (transient waves), or conversely. The mechanical meaning of $E(t)$ is clear because it is just the sum of the energy of the fluid and that of the structure up to multiplicative constants.

Remark 3. The quantity $E(t)$ is the square of a norm as soon as $U < c_f$. However, from (59) and the inequalities (61), we proved that there exist four constants c_2, c_3, c_4, and c_5 such that

$$\text{(63)} \qquad \begin{cases} c_2 E(0) \le \epsilon(t) \le c_3 E(0), \\ c_4 E(0) \le E(t) \le c_5 E(t). \end{cases}$$

5.2. Eulerian invariants. Let us consider a point x_0 of the axis $\overrightarrow{ox}$; it will be specified later on. Let us now set $q = x - x_0$ ($\frac{\partial q}{\partial x} = 1$!). Then, by multiplying (47) by $\frac{\partial \psi}{\partial x} q$ for the fluid equation and by $\frac{\partial d}{\partial x} q$ for the structural equation, we obtain

$$\text{(64)} \quad \begin{cases} \text{(a)} \displaystyle\int_0^T \int_0^L \frac{\partial^2 \psi}{\partial t^2} \frac{\partial \psi}{\partial x} q + 2U \int_0^T \int_0^L \frac{\partial^2 \psi}{\partial x \partial t} \frac{\partial \psi}{\partial x} q + (U^2 - c_f^2) \int_0^T \int_0^L \frac{\partial^2 \psi}{\partial x^2} \frac{\partial \psi}{\partial x} q \\ \qquad\qquad = b \displaystyle\int_0^T \int_{L_1}^{L_2} \left(\frac{\partial d}{\partial t} + U \frac{\partial d}{\partial x} \right) \frac{\partial \psi}{\partial x} q, \\ \text{(b)} \displaystyle\int_0^T \int_{L_1}^{L_2} \frac{\partial^2 d}{\partial t^2} \frac{\partial d}{\partial x} q - c_s^2 \int_0^T \int_{L_1}^{L_2} \frac{\partial^2 d}{\partial x^2} \frac{\partial d}{\partial x} q \\ \qquad\qquad = -a c_f^2 \displaystyle\int_0^T \int_{L_1}^{L_2} \left(\frac{\partial \psi}{\partial t} + U \frac{\partial \psi}{\partial x} \right) \frac{\partial d}{\partial x} q. \end{cases}$$

Then from several integrations by parts one deduces that

$$\text{(65)} \quad \begin{cases} \text{(a)} \left[\displaystyle\int_0^L \frac{\partial \psi}{\partial t} \frac{\partial \psi}{\partial x} q \right]_0^T + U \left[\int_0^L \left(\frac{\partial \psi}{\partial x} \right)^2 q \right]_0^T - \frac{1}{2} \left[\left(\frac{\partial \psi}{\partial t} \right)^2 q \right]_0^L \\ + \dfrac{U^2 - c_f^2}{2} \left[\displaystyle\int_0^T \left(\frac{\partial \psi}{\partial x} \right)^2 q \right]_0^L + \frac{1}{2} \int_0^T \int_0^L \left(\frac{\partial \psi}{\partial t} \right)^2 \frac{\partial q}{\partial x} - \frac{U^2 - c_f^2}{2} \int_0^T \int_0^L \left(\frac{\partial \psi}{\partial x} \right)^2 \frac{\partial q}{\partial x} \\ \qquad\qquad = -b \left[\displaystyle\int_{L_1}^{L_2} \left(\frac{\partial d}{\partial x} \right) \psi q \right]_0^T - b \int_0^T \int_{L_1}^{L_2} \frac{\partial d}{\partial t} \psi \frac{\partial q}{\partial x} \frac{\partial q}{\partial x} \\ \qquad\qquad + b \displaystyle\int_0^T \int_{L_1}^{L_2} \frac{\partial d}{\partial x} \frac{\partial \psi}{\partial t} q + bU \int_0^T \int_{L_1}^{L_2} \frac{\partial \psi}{\partial x} \frac{\partial d}{\partial x} q, \\ \text{(b)} \left[\displaystyle\int_{L_1}^{L_2} \frac{\partial d}{\partial t} \frac{\partial d}{\partial x} q \right]_0^T - \frac{1}{2} \left[\left(\frac{\partial d}{\partial t} \right)^2 q \right]_{L_1}^{L_2} + \frac{1}{2} \int_0^T \int_{L_1}^{L_2} \left(\frac{\partial d}{\partial t} \right)^2 \frac{\partial q}{\partial x} \\ \qquad\qquad - \dfrac{c_s^2}{2} \left[\displaystyle\int_0^T \left(\frac{\partial d}{\partial x} \right)^2 q \right]_{L_1}^{L_2} + \frac{c_s^2}{2} \int_0^T \int_{L_1}^{L_2} \left(\frac{\partial d}{\partial x} \right)^2 \frac{\partial q}{\partial x} \\ \qquad\qquad = -a c_f^2 \displaystyle\int_0^T \int_{L_1}^{L_2} \left(\frac{\partial \psi}{\partial t} + U \frac{\partial \psi}{\partial x} \right) \frac{\partial d}{\partial t} q. \end{cases}$$

Multiplying the second relation by the coefficient $\frac{b}{ac_f^2}$ and adding this to (a), we obtain

$$
\begin{cases}
\dfrac{1}{2}\displaystyle\int_0^T\int_0^L\left(\frac{\partial\psi}{\partial t}\right)^2+\frac{c_f^2-U^2}{2}\int_0^T\int_0^L\left(\frac{\partial\psi}{\partial x}\right)^2+\frac{b}{2ac_f^2}\int_0^T\int_{L_1}^{L_2}\left(\frac{\partial d}{\partial t}\right)^2 \\
+\dfrac{bc_s^2}{2ac_f^2}\displaystyle\int_0^T\int_{L_1}^{L_2}\left(\frac{\partial d}{\partial x}\right)^2+b\int_0^T\int_{L_1}^{L_2}\left(\frac{\partial d}{\partial t}\psi\right) \\
+\left[\displaystyle\int_0^L\frac{\partial\psi}{\partial t}\frac{\partial\psi}{\partial x}q+U\int_0^L\left(\frac{\partial\psi}{\partial x}\right)^2q+b\int_{L_1}^{L_2}\left(\frac{\partial d}{\partial x}\psi\right)q+\frac{b}{ac_f^2}\int_{L_1}^{L_2}\frac{\partial d}{\partial t}\frac{\partial d}{\partial x}q\right]_0^T \\
=\dfrac{1}{2}\left[\displaystyle\int_0^T\left(\frac{\partial\psi}{\partial t}\right)^2q+(c_f^2-U^2)\left(\frac{\partial\psi}{\partial x}\right)^2q\right]_0^L+\frac{b}{2ac_f^2}\left[\int_0^T\left(\frac{\partial d}{\partial t}\right)^2q+c_s^2\left(\frac{\partial d}{\partial x}\right)^2q\right]_{L_1}^{L_2}.
\end{cases}
\tag{66}
$$

This equality enables one to prove a regularity result on the boundary term (as an element of the space $L^2(]0,T[)$), but the main point is the inverse inequality, which will give very interesting information on the control law. The method is similar to the one introduced by Lions [21] and his coworkers [14], [29]. The new difficulty in our study concerns the coupling term

$$b\int_0^T\int_{L_1}^{L_2}\frac{\partial d}{\partial t}\psi.$$

The first point consists in noting that for any $\alpha>0$

$$\left|b\int_0^T\int_{L_1}^{L_2}\frac{\partial d}{\partial t}\psi\right|\le\frac{b\alpha}{2}\int_0^T\int_{L_1}^{L_2}\left(\frac{\partial d}{\partial t}\right)^2+\frac{b}{2\alpha}\int_0^T\int_{L_1}^{L_2}\psi, \tag{67}$$

and if we introduce the smallest eigenvalue $\lambda_1^{sk}=\eta_1^{sk}(c_f^2-U^2)$ of the generalized Steklov problem (see (24)), we get

$$\left|b\int_0^T\int_{L_1}^{L_2}\frac{\partial d}{\partial t}\psi\right|\le\frac{b\alpha}{2}\int_0^T\int_{L_1}^{L_2}\left(\frac{\partial d}{\partial t}\right)^2+\frac{b}{2\alpha\eta_1^{sk}}\int_0^T\int_{L_1}^{L_2}\left(\frac{\partial\psi}{\partial x}\right)^2. \tag{68}$$

Let us assume that the following geometrical condition is satisfied for a given Mach number M (one has $\eta_1^{sk}=\frac{\pi^2}{(L_2-L_1)^2}$):

$$ab<\eta_1^{sk}(1-M^2)\frac{\pi^2}{(L_2-L_1)^2}(1-M^2).$$

Then if the steady flow velocity U satisfies

$$U<c_f\sqrt{1-\frac{ab}{\eta_0^{sk}}}=U_0, \tag{69}$$

then one can find a number $\alpha > 0$ such that

$$
\begin{cases}
\dfrac{1}{2}\displaystyle\int_0^T\int_0^L\left(\frac{\partial\psi}{\partial t}\right)^2+\left(\frac{c_f^2-U^2}{2}-\frac{b}{2\alpha\eta_1^{sk}}\right)\int_0^T\int_0^L\left(\frac{\partial\psi}{\partial x}\right)^2 \\
+\left(\dfrac{b}{2ac_f^2}-\dfrac{b\alpha}{2}\right)\displaystyle\int_0^T\int_{L_1}^{L_2}\left(\frac{\partial d}{\partial t}\right)^2+\frac{c_s^2 b}{2ac_f^2}\int_0^T\int_{L_1}^{L_2}\left(\frac{\partial d}{\partial x}\right)^2 \\
-2T_0\displaystyle\sup_{t\in[0,T]}\left[\frac{1}{2}\int_0^L\left(\frac{\partial\psi}{\partial t}\right)^2\right. \\
\qquad\left.+\left(\dfrac{c_f^2-U^2}{2}\right)\displaystyle\int_0^L\left(\frac{\partial\psi}{\partial x}\right)^2+\frac{b}{ac_f^2}\left(\frac{1}{2}\int_{L_1}^{L_2}\left(\frac{\partial d}{\partial t}\right)^2+\frac{c_s^2}{2}\int_{L_1}^{L_2}\left(\frac{\partial d}{\partial x}\right)^2\right)\right] \\
\le\dfrac{1}{2}\left[\displaystyle\int_0^T\left(\frac{\partial\psi}{\partial t}\right)^2 q+(c_f^2-U^2)\left(\frac{\partial\psi}{\partial x}\right)^2 q\right]_0^L+\frac{b}{2ac_f^2}\left[\int_0^L\left(\frac{\partial d}{\partial t}\right)^2 q+c_s^2\left(\frac{\partial d}{\partial x}\right)^2 q\right]_{L_1}^{L_2},
\end{cases}
\tag{70}
$$

where we have set

$$
T_0=\min_{\theta>0}\left(\max\left(\frac{L}{c_f-U}+\frac{b}{\theta\eta_1^{sk}},\left(\frac{L_2-L_1}{ac_s c_f^2}+\theta\right)b\right)\right). \tag{71}
$$

Finally, setting

$$
\kappa=\max_{\alpha>0}\left(\min\left(1-\frac{b}{\alpha\eta_1^{sk}(c_f^2-U^2)},1-a\alpha c_f^2\right)\right) \tag{72}
$$

and because of (59), one obtains (see (69))

$$
\begin{cases}
(\kappa c_5 T-2T_0)E(0) \\
\le\dfrac{1}{2}\left[\displaystyle\int_0^T\left(\frac{\partial\psi}{\partial t}\right)^2 q+(c_f^2-U^2)\int_0^T\left(\frac{\partial\psi}{\partial x}\right)^2 q\right]_0^L \\
+\dfrac{b}{2ac_f^2}\left[\displaystyle\int_0^T\left(\frac{\partial d}{\partial t}\right)^2 q+\ c_s^2\int_0^T\left(\frac{\partial d}{\partial x}\right)^2 q\right]_{L_1}^{L_2}.
\end{cases}
\tag{73}
$$

Remark 4. The estimate (73) makes sense only if $\kappa > 0$. But one can notice that this condition is obvious if a or b is zero. This points out that the restriction is fully connected to the coupling between the fluid and the structure.

Remark 5. The assumptions required for justifying (73) are

$$
\begin{cases}
U<c_f\sqrt{1-\dfrac{ab}{\eta_1^{sk}}}<c_f, \\
ab<\eta_1^{sk}(1-M^2)=\dfrac{\pi^2}{(L_2-L_1)^2}(1-M^2), \\
U<\dfrac{c_f}{\sqrt{1+\dfrac{abc_f^2}{\eta_1^{sk}c_s^2}}}=U_c\ \ (<c_f!).
\end{cases}
\tag{74}
$$

But from the boundary conditions (recalling that $M = \frac{U}{c_f} < 1$),

$$\left\{ \begin{array}{ll} \dfrac{\partial d}{\partial t} = 0 & \text{for } x = L_1 \text{ and } L_2, \\ \dfrac{\partial \psi}{\partial t} = U \left(\dfrac{1}{M^2} - 1 \right) \dfrac{\partial \psi}{\partial x} & \text{for } x = 0 \text{ and } L, \end{array} \right. \tag{75}$$

and then assuming that (74) is satisfied, one deduces, setting $q = x - \frac{L_1+L_2}{2}$, that

$$\left\{ \begin{array}{l} (\kappa c_5 T - 2T_0) E(0) \leq \dfrac{L}{2} c_f^2 \left(\dfrac{1-M^2}{M^2} \right) \left[\displaystyle\int_0^T \left(\frac{\partial \psi}{\partial x} \right)^2 (0) + \left(\frac{\partial \psi}{\partial x} \right)^2 (L) \right] \\ \qquad + \dfrac{b(L_2 - L_1)}{2ac_f^2} \left[\displaystyle\int_0^T c_s^2 \left(\frac{\partial d}{\partial x} \right)^2 (L_1) + c_s^2 \left(\frac{\partial d}{\partial x} \right)^2 (L_2) \right]. \end{array} \right. \tag{76}$$

Remark 6. The coefficients a and b can be expressed with respect to the mass density of the structure—say, ϱ_s—and of the fluid (at rest)—say, ϱ_f—but also using the thickness of the structure—say, 2ε—and the inner radius of the flow duct—say, R. One has $ab = \frac{1}{\varepsilon R} \frac{\varrho_f}{\varrho_s}$.

6. Asymptotic behavior of the control problem when $\varepsilon \to 0$. Let us first introduce a formal asymptotic expansion of $(\varphi^\varepsilon, z^\varepsilon, u^\varepsilon)$ which is solution of the optimal control problem. Thus we set

$$\left\{ \begin{array}{l} (\varphi^\varepsilon, z^\varepsilon, u^\varepsilon) = (\varphi^0, z^0, u^0) + \varepsilon(\varphi^1, z^1, u^1) + \cdots, \\ (\psi^\varepsilon, d^\varepsilon) = (\psi^0, d^0) + \varepsilon(\psi^1, d^1) + \cdots. \end{array} \right. \tag{77}$$

Nothing guarantees the validity of this expansion. However, a convergence result will be proved when $\varepsilon \to 0$ as soon as the exact controllability conditions are satisfied. The first step consists in identifying the terms of order zero and one. The terms of order one enable us to characterize the optimal control u^0. In fact, this limit control is exactly the one given by the so-called H.U.M. method of Lions [21]. A few additional difficulties arise because of the coupling and the instabilities which can appear when the velocity U is large enough.

6.1. Identification of terms of order zero. By introducing (77) into the equations satisfied by $(\varphi^\varepsilon, z^\varepsilon, u^\varepsilon)$ and by equating the terms of same order in ε, one obtains the following necessary conditions:

(a) For the primal model,

$$\left\{ \begin{array}{l} \dfrac{\partial^2 \varphi^0}{\partial t^2} + 2U \dfrac{\partial^2 \varphi^0}{\partial x \partial t} + (U^2 - c_f^2) \dfrac{\partial^2 \varphi^0}{\partial x^2} \\ \qquad = ac_f^2 \left(\dfrac{\partial z^0}{\partial t} + U \dfrac{\partial z^0}{\partial x} \right) \chi_{[L_1, L_2]}(x) \quad \forall (x,t) \in]0, L[\times]0, T[, \\ \left[\dfrac{\partial \varphi^0}{\partial t} + U \left(1 - \dfrac{1}{M^2} \right) \dfrac{\partial \varphi^0}{\partial x} \right] (0,t) \\ \qquad = \left[\dfrac{\partial \varphi^0}{\partial t} + U \left(1 - \dfrac{1}{M^2} \right) \dfrac{\partial \varphi^0}{\partial x} \right] (L,t) = 0 \quad \forall t \in]0, T[, \\ \dfrac{\partial^2 z^0}{\partial t^2} - c_s^2 \dfrac{\partial^2 z^0}{\partial x^2} = -b \left(\dfrac{\partial \varphi^0}{\partial t} + U \dfrac{\partial \varphi^0}{\partial x} \right) \quad \forall (x,t) \in]L_1, L_2[\times]0, T[, \\ z^0(L_1, t) = z^0(L_2, t) = 0 \quad \forall t \in]0, T[. \end{array} \right. \tag{78}$$

The initial conditions satisfied by (φ^0, z^0) are those defined in (43).

(b) For the adjoint model,

$$
\left\{
\begin{array}{l}
\dfrac{\partial^2\psi^0}{\partial t^2} + 2U\dfrac{\partial^2\psi^0}{\partial x\partial t} + (U^2 - c_f^2)\dfrac{\partial^2\psi^0}{\partial x^2} \\
\qquad = b\left(\dfrac{\partial d^0}{\partial t} + U\dfrac{\partial d^0}{\partial x}\right)\chi_{[L_1,L_2]}(x) \quad \forall (x,t) \in]0,L[\times]0,T[\,, \\
\left[\dfrac{\partial\psi^0}{\partial t} + U\left(1 - \dfrac{1}{M^2}\right)\dfrac{\partial\psi^0}{\partial x}\right](0,t) \\
\qquad = \left[\dfrac{\partial\psi^0}{\partial t} + U\left(1 - \dfrac{1}{M^2}\right)\dfrac{\partial\psi^0}{\partial x}\right](L,t) = 0 \quad \forall t \in]0,T[\,, \\
\dfrac{\partial^2 d^0}{\partial t^2} - c_s^2\dfrac{\partial^2 d^0}{\partial x^2} = -ac_f^2\left(\dfrac{\partial\psi^0}{\partial t} + U\dfrac{\partial\psi^0}{\partial x}\right) \quad \forall (x,t) \in]L_1,L_2[\times]0,T[\,, \\
d^0(L_1,t) = d^0(L_2,t) = 0 \quad \forall t \in]0,T[\,.
\end{array}
\right. \tag{79}
$$

The final conditions satisfied by (ψ^0, d^0) are those defined in (48).

(c) For the optimality equations when two controls functions are used,

$$
d^0(x,t) = \psi^0(x,t) = 0 \ \ \forall (x,t) \in]\alpha,\beta[\times]0,T[\,. \tag{80}
$$

Remark 7. The existence and uniqueness of a solution to systems (78) and (79) can be obtained by the same method used for $(\varphi^\varepsilon, z^\varepsilon)$ and $(\psi^\varepsilon, d^\varepsilon)$.

Remark 8. From (80) and using the inverse inequality, we prove in the following that:

$$
(\psi^0, d^0) = 0.
$$

However, this inverse inequality will give much more information on the control and also on the space of initial data which can be controlled. Many of technical tricks are necessary, which are developed in section 7.

6.2. Exact controllability of noise in the duct with only the structural control. In this section we consider that there is only one control, which is the one applied on the structure (i.e., $w = 0$ and only u is active). The optimality condition is now restricted to the following one:

$$
d^0(x,t) = 0 \ \ \forall (x,t) \in]\alpha,\beta[\times]0,T[\,. \tag{81}
$$

This corresponds to a very simple version of the Holmgrem theorem for a coupled fluid-structure model. From (80) and using the structural equation, one deduces that

$$
\frac{\partial\psi^0}{\partial t} + U\frac{\partial\psi^0}{\partial x}(x,t) = 0 \ \ \forall (x,t) \in]\alpha,\beta[\times]0,T[\,.
$$

Therefore there exists a function k such that on $\ ]\alpha,\beta[\times]0,T[$

$$
\psi^0(x,t) = k(x - Ut).
$$

But from the fluid equation one has

$$
\frac{\partial^2\psi^0}{\partial t^2} + 2U\frac{\partial^2\psi^0}{\partial x\partial t} + (U^2 - c_f^2)\frac{\partial^2\psi^0}{\partial x^2} = -c_f^2\frac{\partial^2 k}{\partial x^2}(x - Ut) = 0 \ \ \forall (x,t) \in]\alpha,\beta[\times]0,T[\,,
$$

and therefore there exist two constants F and G such that

$$\psi^0(x,t) = F(x-Ut) + G \ \ \forall (x,t) \in]\alpha,\beta[\times]0,T[\,. \tag{82}$$

Let us now multiply the equations of the adjoint state by $(\frac{\partial \psi^0}{\partial t}, \frac{\partial d^0}{\partial t})$, and let us integrate on $]0,\alpha[$ and $]\beta,L[$ separately. Because $d(\alpha,t) = d(\beta,t) = 0 \ \forall t \in]0,T[$ and $d^0(x,t) = 0 \ \forall (x,t) \in]\alpha,\beta[\times]0,T[$, from the expression of ψ^0 on $]\alpha,\beta[\times]0,T[$ given at (82), we obtain

$$\begin{cases} \dfrac{\partial .}{\partial t}[\epsilon^{0\alpha}] = -UF^2c_f^2, \\ \dfrac{\partial .}{\partial t}[\epsilon^{\beta L}] = UF^2c_f^2, \end{cases} \tag{83}$$

where $\epsilon^{0\alpha}$ (respectively, $\epsilon^{\beta L}$) is the quantity defined at (60), but where the integrals are restricted to $]0,\alpha[$ (respectively, $]\beta,L[$). By integrating (83) from 0 to t, we deduce that

$$\begin{cases} \epsilon^{0\alpha}(t) = \epsilon^{0\alpha}(0) - UFc_f^2t, \\ \epsilon^{\beta L}(t) = \epsilon^{\beta L}(0) + UFc_f^2t. \end{cases} \tag{84}$$

It is worth noting that for $U < U_c$ (see (29)) the quantity $\epsilon^{0\alpha}(t)$ is positive. Hence for t large enough, one has necessarily $F = 0$. But it is necessary to get rid of this time condition which depends on F. Therefore we use again the multiplier method with the equations satisfied by the adjoint state. The multipliers are $\frac{\partial \psi^0}{\partial x}(x-x_0)$ and $\frac{\partial d^0}{\partial x}(x-x_0)$ with, respectively, $x_0 = 0$ for the set $]0,\alpha[$ and $x_0 = L$ for $]\beta,L[$. Assuming again that $U < U_c$ (see (29)), one deduces from a computation similar to the one which led to (66) that the following inequalities hold:

$$\begin{cases} \displaystyle\int_0^T \epsilon^{0\alpha}(t) - T_0(\epsilon^{0\alpha}(0) + \epsilon^{0\alpha}(T)) \le \frac{\alpha}{2}F^2Tc_f^2, \\ \displaystyle\int_0^T \epsilon^{\beta L}(t) - T_0(\epsilon^{\beta L}(0) + \epsilon^{\beta L}(T)) \le \frac{L-\beta}{2}F^2Tc_f^2, \end{cases} \tag{85}$$

where T_0 is a constant homogeneous to time and which could be adjusted separately on $]0,\alpha[$ and $]\beta,L[$. From (85) and using (84), we deduce that, for instance, on $]\beta,L[$ one has

$$\int_0^T \epsilon^{\beta L}(t) - 2T_0\epsilon^{\beta L}(0) - UT_0TF^2c_f^2 \le \frac{L-\beta}{2}F^2Tc_f^2, \tag{86}$$

or else

$$(T-2T_0)\epsilon^{\beta L}(0) + \frac{UF^2c_f^2}{2}T^2 - T_0TUFc_f^2 \le \frac{L-\beta}{2}F^2Tc_f^2. \tag{87}$$

One can arrange the previous expression such that

$$(T-2T_0)\epsilon^{\beta L}(0) + \frac{F^2c_f^2}{2}T(UT - 2T_0U - (L-\beta)) \le 0. \tag{88}$$

Let us consider that $T > 2T_0 + \frac{L-\beta}{U}$. Such a condition is possible if and only if $U > 0$. Then we conclude that $F = 0$. Thus we can also conclude that $\epsilon^{\beta L}(0) = 0$. From the

final condition which should be satisfied by $\psi^0(x,T)$ (see (48)), and because $\varphi^0(x,t)$, we prescribe on ϕ^0 the following condition ((see (7)):

$$\int_{L_1}^{L_2} \varphi^0(x,t)dx = 0 \ \forall t \in [0,T]; \tag{89}$$

one can also ensure that $G = 0$. Finally, we proved that

$$(\psi^0, d^0) = 0 \quad \forall (x,t). \tag{90}$$

When the steady velocity U is zero, then the function ψ^0 is linear with respect to the coordinate x on $]\alpha,\beta[\times]0,T[$. This proves that steady flows cannot be controlled even if they satisfy all the equations of the model for ψ^0, d^0. Let us summarize the previous results in the following theorem.

THEOREM 1. *Let us consider a set of initial data for the coupled model* $(\varphi_0, \varphi_1, z_0, z_1)$*, which are assumed to be smooth enough and such that* φ^0 *satisfies* (7)*. Then if* d^0 *is zero on the set* $]\alpha,\beta[\times]0,T[$*, and if*

(i) $T > 2T_0 + \frac{L-\beta}{U}$*, where* T_0 *is defined at* (71)*,*

(ii) U *satisfies* (74)*,*

then

$$(\psi^0, d^0) \equiv 0.$$

If $U = 0$*, then one only has* $\psi^0(x,t) = Fx + G$*.*

Remark 9. The results obtained in (89) are still true even if the control w^0 is also applied. But the proof is much easier in this latter case because the optimality condition implies directly that $\psi^0 = 0 \ \forall (x,t) \in]\alpha,\beta[$. This can be checked directly on the proof given previously. If the controllability is a general result which doesn't require the introduction of w (the second control function), it will enable us to characterize an exact control (u,w) in the space $[L^2(]\alpha,\beta[\times]0,T[)]^2$ for finite energy initial data.

Remark 10. The condition on T is not classical in control theory (see Lions [21]), but it can also be physically interpreted. First, for $U = 0$ it cannot be satisfied, because the information cannot travel with the particles along the duct and because the boundary conditions that we chose for the steady flow (i.e., uniform velocity) cannot be controlled. For $U \neq 0$ the condition on T means that a perturbation introduced at one extremity of the flow duct can cross the full length of the duct and return at the entrance. But it is also necessary to add the time necessary for the steady flow to cross the uncontrolled portion of the duct $(L - \beta)$.

Remark 11. A simple consequence of (89) is that the control (u^0, w^0) is exact. However, it is not yet defined. This is the goal of the next section.

Remark 12. A similar result to the one given in this section can be obtained for a control which is only applied to the fluid (i.e., $u = 0$ and $w \neq 0$).

6.3. Identification of terms of order one. The model characterizing $(\varphi^1, z^1, \psi^1 d^1)$ is similar to the one characterizing the terms of order zero. But only few of the equations are necessary in order to define (u^0, w^0), which is the formal limit (at this step) of $(u^\varepsilon, w^\varepsilon)$ when ε tends to zero. First, let us note that the optimality conditions give

$$u^0 + d^1 = 0, \ w^0 + \psi^1 = 0 \qquad \forall (x,t) \in]\alpha,\beta[\times]0,T[\,. \tag{91}$$

Then, setting $(u^0, w^0) = -(d^1, \psi^1)$ in the equations satisfied by (φ^0, z^0), we obtain

$$
\left\{
\begin{array}{l}
\forall \delta\varphi \in H^1(]0,L[),\ \displaystyle\int_0^L \frac{\partial^2\varphi^0}{\partial t^2}\delta\varphi + U\int_0^L \left(\frac{\partial^2\varphi^0}{\partial x\partial t}\delta\varphi - \frac{\partial\varphi^0}{\partial t}\frac{\partial\delta\varphi}{\partial x}\right) \\
\qquad\qquad + (c_f^2 - U^2)\displaystyle\int_0^L \frac{\partial\varphi^0}{\partial x}\frac{\partial\delta\varphi}{\partial x} \\
\qquad = ac_f^2\displaystyle\int_{L_1}^{L_2}\left(\frac{\partial z^0}{\partial t} + U\frac{\partial z^0}{\partial x}\right)\delta\varphi - \int_\alpha^\beta \psi^1\delta\varphi, \\
\forall \delta z \in H_0^1(]L_1, L_2[),\ \displaystyle\int_{L_1}^{L_2}\frac{\partial^2 z^0}{\partial t^2}\delta z + c_s^2\int_{L_1}^{L_2}\frac{\partial z^0}{\partial x}\frac{\partial\delta z}{\partial x} \\
\qquad = -b\displaystyle\int_{L_1}^{L_2}\left(\frac{\partial\varphi^0}{\partial t} + U\frac{\partial\varphi^0}{\partial x}\right)\delta z - \int_\alpha^\beta d^1\delta z.
\end{array}
\right. \tag{92}
$$

Let us assume that $(\delta\varphi, \delta z)$ is a solution of the adjoint state (as for (φ^1, d^1)). Thus we obtain (because at $t = T$, φ^0, $\frac{\partial\varphi^0}{\partial t}$, z^0, $\frac{\partial z^0}{\partial t}$ are zero)

$$
\left\{
\begin{array}{l}
\displaystyle\int_0^T\int_\alpha^\beta d^1\delta z + \psi^1\delta\varphi = \int_0^L \frac{\partial\varphi^0}{\partial t}(x,0)\delta\varphi(x,0) - \int_0^L \varphi^0(x,0)\frac{\partial\delta\varphi}{\partial t}(x,0) \\
+ 2U\displaystyle\int_0^L \frac{\partial\varphi^0}{\partial x}(x,0)\delta\varphi(x,0) - ac_f^2\int_{L_1}^{L_2} d^0(x,0)\delta\varphi(x,0) \\
+\displaystyle\int_{L_1}^{L_2}\frac{\partial z^0}{\partial t}(x,0)\delta z(x,0) - \int_{L_1}^{L_2} z^0(x,0)\frac{\partial\delta z^0}{\partial t}(x,0) + b\int_{L_1}^{L_2}\varphi^0(x,0)\delta z(x,0).
\end{array}
\right. \tag{93}
$$

Setting

$$
\left\{
\begin{array}{ll}
\Phi_0 = (\psi^1, d^1)(x,0), & \Phi_1 = \left(\dfrac{\partial\psi^1}{\partial t}, \dfrac{d^1}{\partial t}\right)(x,0), \\
\delta\Phi_0 = (\delta\psi, \delta z)(x,0), & \delta\Phi_1 = \left(\dfrac{\partial\delta\psi^1}{\partial t}, \dfrac{\partial\delta d^1}{\partial t}\right)(x,0),
\end{array}
\right. \tag{94}
$$

we introduce the following bilinear and linear forms:

$$
\left\{
\begin{array}{l}
\Lambda(\Phi, \delta\Phi) = \displaystyle\int_0^T\int_\alpha^\beta d^1\delta z + \psi^1\delta\varphi, \\
\qquad \text{where } \Phi = (\Phi_0, \Phi^1) \text{ and } \delta\Phi = (\delta\Phi_0, \delta\Phi^1), \\
L(\delta\Phi) = \displaystyle\int_0^L \frac{\partial\varphi^0}{\partial t}(x,0)\delta\varphi(x,0) - \int_0^L \varphi^0(x,0)\frac{\partial\varphi}{\partial t}(x,0) \\
+ 2U\displaystyle\int_0^L \frac{\partial\varphi^0}{\partial x}(x,0)\delta\varphi(x,0) - ac_f^2\int_{L_1}^{L_2} z^0(x,0)\delta\varphi(x,0) \\
+\displaystyle\int_{L_1}^{L_2}\frac{\partial z^0}{\partial t}(x,0)\delta z(x,0) - \int_{L_1}^{L_2} z^0(x,0)\frac{\partial\delta z}{\partial t}(x,0) + b\int_{L_1}^{L_2}\varphi^0(x,0)\delta z(x,0).
\end{array}
\right. \tag{95}
$$

Hence (92) can be formulated as follows:

$$
\left\{
\begin{array}{l}
\text{find } \Phi \in V^* \text{ such that} \\
\forall \delta\Phi \in V^*,\ \Lambda(\Phi, \delta\Phi) = L(\delta\Phi).
\end{array}
\right. \tag{96}
$$

The space V^* is not yet defined; it is the completed space with respect to the norm

$$\Phi \in [H^1(]0,L[) \times L^2(]0,L[), H_0^1(]L_1,L_2[) \times L^2(]L_1,L_2[)] \ \rightarrow \ \sqrt{\Lambda(\Phi,\Phi)}.$$

Conversely, if Φ solution of (96) can be found, one has

$$\left\{\begin{array}{l} \forall \delta\Phi \in V^*, \ \displaystyle\int_0^L \frac{\partial \varphi^0}{\partial t}(x,T)\delta\varphi(x,T) - \int_0^L \varphi^0(x,T)\frac{\partial \varphi}{\partial t}(x,T) \\ \quad + 2U\displaystyle\int_0^L \frac{\partial \varphi^0}{\partial x}(x,T)\delta\varphi(x,T) - ac_f^2\int_{L_1}^{L_2} z^0(x,T)\delta\varphi(x,T) \\ \quad + \displaystyle\int_{L_1}^{L_2} \frac{\partial z^0}{\partial t}(x,T)\delta z(x,T) - \int_{L_1}^{L_2} z^0(x,T)\frac{\partial \delta z}{\partial t}(x,T) \\ \quad + b\displaystyle\int_{L_1}^{L_2} \varphi^0(x,T)\delta z(x,T) = 0. \end{array}\right. \tag{97}$$

Because, on the one hand, the space V^* contains $[H^1(]0,L[)\times L^2(]0,L[), H_0^1(]L_1,L_2[)\times L^2(]L_1,L_2[)]$, and on the other hand, the final value (at time $t=T$) of $(\delta\varphi,\delta z)$ can be chosen arbitrarily in this space (it is sufficient to reverse the time and to choose the initial value obtained for $(\delta\Phi)$!), one deduces from (97) that (as soon as the initial data are such that the expression (97) make sense, i.e., the initial data should be smooth enough)

$$\varphi^0(x,T) = \frac{\partial \varphi^0}{\partial t}(x,T) = 0 \quad \forall x \in]0,L[,$$

and

$$z^0(x,T) = \frac{\partial z^0}{\partial t}(x,T) = 0 \ \ \forall x \in]L_1,L_2[.$$

There are still two important points to justify: the characterization of the space V^* and the construction of the exact control, and then the convergence of the sequences $(u^\varepsilon, w^\varepsilon)$ to (u^0,w^0) when $\varepsilon \to 0$ (and even the one of u^ε to u^0 if only the structural control is used).

7. Construction of an exact control. In this section we make use of a technical method given in the book of Lions [21] and due to Zuazua. There are four steps which basically make use of the inequalities obtained by the multiplier method as in section 4. Let us first state the results that we are going to prove.

THEOREM 2. *Let $\Lambda(.,.)$ be the symmetrical and bilinear form defined at* (95). *For T large enough, there exists a strictly positive constant—say, c_0—such that*

$$\forall X \in V^\#, \ \ \Lambda(\Phi,\Phi) \geq c_0\left[||\Phi_0||_{0,0L}^2 + ||\Phi_1||_{[H^1(]O,L[)]'}^2 + ||D_0||_{0,0L}^2 + ||D_1||_{-1,0L}^2\right], \tag{98}$$

where $X = (\Phi_0,\Phi_1,D_0,D_1)$ is the initial condition for the adjoint state for (ϕ^1,d^1).

THEOREM 3. *Let*

$$\varphi^0(x,0) \in H^1(]0,L[), \ \ \frac{\partial \varphi}{\partial t}(x,0) \in L^2(]0,L[),$$

and

$$z^0(x,0) \in H_0^1(]L_1,L_2[), \ \ \frac{\partial z^0}{\partial t}(x,0) \in L^2(]L_1,L_2[).$$

Then $L(.)$ defined at (96) *is a linear and continuous form on the space*

$$V^{\#} = L^2(]0,L[) \times [H^1(]0,L[)]' \times L^2(]L_1,L_2[) \times H^{-1}(]L_1,L_2[).$$

Remark 13. The dual space $(H^1(]0,L[))'$ is isomorphic to $H^{-1}(]0,L[) \times R^2$. The two scalar components are the coefficients of the Dirac distributions at both end of the segment $]0,L[$. Therefore the term $\int_0^L \varphi(x,0)\frac{\partial\delta\varphi}{\partial t}(x,0)$ should be written more precisely using a duality product between $H^1(]0,L[)$ and its dual space. But this would not lead to new phenomena or new difficulties. The discussion would be quite different for the two dimensional cases that are considered in [7].

From Theorems 2 and 3, we can deduce the following controllability result.

THEOREM 4. *Let $(\varphi_0,\ \varphi_1\ z_0,\ z_1)$ be a set of initial conditions for the coupled system lying in the space*

$$H_m^1(]0,L[) \times L^2(]0,L[) \times H_0^1(]L_1,L_2[) \times L^2(]L_1,L_2[).$$

Then there exists an exact control (u^0,w^0) in the space $[L^2(]\alpha,\beta[\times]0,T[)]^2$.

Proof of the Theorem 4. Because of Theorems 2 and 3, one can claim that the variational equation (96) has a unique solution, say (φ^1,d^1). The exact control $(u^0,w^0)(x,t)$ is then given by $-(\varphi^1,d^1)(x,t)\chi(x)$ because of (97). □

Proof of the Theorem 2. There are several steps in our proof. They are rather technical and mainly rest upon the multiplier method.

Step 1. Let us set $q=(x-x_0)t(t-T)$, where $x_0 \in]\alpha,\beta[$. Then we consider the equations of the coupled dual model, which should be satisfied by (ψ^1,d^1). They can be written as follows:

$$(99)\quad \begin{cases} \dfrac{\partial^2\psi^1}{\partial t^2} + 2U\dfrac{\partial^2\psi^1}{\partial x\partial t} + (U^2-c_f^2)\dfrac{\partial^2\psi^1}{\partial x^2} \\ \qquad = b\left(\dfrac{\partial d^1}{\partial t} + U\dfrac{\partial d^1}{\partial x}\right)\chi_{[L_1,L_2]}(x) \quad \forall (x,t)\in]0,L[\times]0,T[, \\ \left[\dfrac{\partial\psi^1}{\partial t} + U\left(1-\dfrac{1}{M^2}\right)\dfrac{\partial\psi^1}{\partial x}\right](0,t) \\ \qquad = \left[\dfrac{\partial\psi^1}{\partial t} + U\left(1-\dfrac{1}{M^2}\right)\dfrac{\partial\psi^1}{\partial x}\right](L,t) = 0 \quad \forall t\in]0,T[, \\ \dfrac{\partial^2 d^1}{\partial t^2} - c_s^2\dfrac{\partial^2 d^1}{\partial x^2} = -ac_f^2\left(\dfrac{\partial\psi^1}{\partial t} + U\dfrac{\partial\psi^1}{\partial x}\right) \quad \forall (x,t)\in]L_1,L_2[\times]0,T[, \\ d^1(L_1,t) = d^1(L_2,t) = 0 \quad \forall t\in]0,T[. \end{cases}$$

Let us multiply each of these equations by $\frac{\partial\psi^1}{\partial x}q$ and $\frac{\partial d^1}{\partial x}q$, and for $\eta>0$ let us integrate on $]\alpha,\beta[\times]\eta,T-\eta[$, where $]\alpha,\beta[$ is the segment on which the controls are applied. Thus from a computation similar to the one we did previously, we obtain, where c_1 is a constant which tends to the infinity when $\eta\to 0$,

$$(100)\quad \begin{cases} \displaystyle\int_\eta^{T-\eta}\left[\left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2-U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2\right]_\alpha^\beta + \frac{ac_f^2}{b}\left[\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right]_\alpha^\beta \\ \displaystyle\le c_1\left[\int_0^T\int_\alpha^\beta \left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2-U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\}\right]. \end{cases}$$

But one can also apply the multiplier method on the set $]0,\eta[\times]\alpha,\beta[$ and then, by upper bounding the energy on $]0,\eta[\times]0,L[$, we obtain $\left(\text{with } q = x - \frac{\alpha+\beta}{2}\right)$

$$\left\{\begin{array}{l} \displaystyle\int_0^\eta \left[\left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2\right]_\alpha^\beta + \frac{ac_f^2}{b}\left[\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right]_\alpha^\beta \\ \leq 2\eta E(0), \end{array}\right. \tag{101}$$

where $E(0)$ is the energy of the global system defined at (62) and for $t = 0$. Thus by adding (100) and (101), one can conclude that

$$\left\{\begin{array}{l} \displaystyle\int_0^T \left[\left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2\right]_\alpha^\beta + \frac{ac_f^2}{b}\left[\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right]_\alpha^\beta \\ \displaystyle\leq c_1\left[\int_0^T\int_\alpha^\beta \left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\}\right] \\ + c_2\eta E(0). \end{array}\right. \tag{102}$$

Step 2. Let us now consider the segment $]0,\alpha[$, and we define two functions, say, $q_f = x$ and $q_s = x - L_1$. Then by multiplying the two equations (99) by $\frac{\partial\psi^1}{\partial x}q_f$ and $\frac{\partial d^1}{\partial x}q_s$, respectively, one obtains that (for $U < U_c$) there exists a positive constant c and a time delay T_0 such that

$$\left\{\begin{array}{l} \displaystyle\int_0^T\int_0^\alpha \left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\} \\ \displaystyle - T_0\left\{\left[\int_0^\alpha \left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\}\right](0)\right. \\ \displaystyle\left. + \left[\int_0^\alpha \left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\}\right](T)\right\} \\ \displaystyle\leq \left[\int_0^T \left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\}\right](\alpha). \end{array}\right. \tag{103}$$

The time T_0 (or at least an upper bound) has been defined similarly in (71). The same inequality as (103) can also be written on the segment $[\beta, L]$. By adding these two and choosing for T_0 the largest one (one could compute the best constant depending on the geometry of the system), we obtain the inverse inequality for the coupled system:

$$\left\{\begin{array}{l} \displaystyle\left[\int_0^T \left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\}\right](\alpha) \\ \displaystyle + \left[\int_0^T \left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\}\right](\beta) \\ \displaystyle + \left[\int_0^T\int_\alpha^\beta \left(\frac{\partial\psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial\psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\}\right] \\ \displaystyle\geq \tilde{c}(T - 2T_0 - 4\eta)E(0), \end{array}\right. \tag{104}$$

where $\tilde{c}$ is a new positive constant. Finally, we proved that for a constant $\tilde{\tilde{c}}$, one has

$$\left\{ \begin{array}{l} \left[\displaystyle\int_0^T \int_\alpha^\beta \left(\frac{\partial \psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial \psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\}\right] \\ \qquad\qquad \geq \tilde{\tilde{c}}(T - 2T_0 - 4\eta)E(0). \end{array} \right. \tag{105}$$

Step 3. Following again a strategy introduced by Lions [21], we apply the multiplier method to (99) with the multipliers $\psi^1 t(t-T)(x-x_0)$ and $d^1 t(t-T)(x-x_0)$ and by integrating over $]\alpha, \beta[\times]0, T[$. Thus we obtain from a standard computation (see [21])

$$\left\{ \begin{array}{l} \left[\displaystyle\int_0^T \int_\alpha^\beta \left(\frac{\partial \psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)\left(\frac{\partial \psi^1}{\partial x}\right)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2\left(\frac{\partial d^1}{\partial x}\right)^2\right\}\right] \\ \leq \left[\displaystyle\int_0^T \int_\alpha^\beta \left(\frac{\partial \psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)(\psi^1)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2 (d^1)^2\right\}\right] \\ \qquad\qquad + 2\eta E(0). \end{array} \right. \tag{106}$$

Step 4. From (104) and (105) we can conclude that there exists a new constant—say, c—such that

$$\left\{ \begin{array}{l} (T - 2T_0 - 6\eta)E(0) \\ \leq \left[\displaystyle\int_0^T \int_\alpha^\beta \left(\frac{\partial \psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)(\psi^1)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2 (d^1)^2\right\}\right]. \end{array} \right. \tag{107}$$

Hence the right-hand side of (106) defines the square of a norm on the initial data $(\varphi_0, \varphi_1, z_0, z_1)$ in the space $(E(0) < \infty)$:

$$V = [H^1(]0, L[) \times L^2(]0, L[), H_0^1(]L_1, L_2[) \times L^2(]L_1, L_2[)].$$

Step 5. The last step of the proof of Theorem 4 rests on a compactness argument. Thus we are going to prove that there exists a positive constant—say again, c—such that

$$\left\{ \begin{array}{l} \left[\displaystyle\int_0^T \int_\alpha^\beta \left(\frac{\partial \psi^1}{\partial t}\right)^2 + (c_f^2 - U^2)(\psi^1)^2 + \frac{ac_f^2}{b}\left\{\left(\frac{\partial d^1}{\partial t}\right)^2 + c_s^2 (d^1)^2\right\}\right] \\ \leq c\left[\displaystyle\int_0^T \int_\alpha^\beta \left(\frac{\partial \psi^1}{\partial t}\right)^2 + \left(\frac{\partial d^1}{\partial t}\right)^2\right]. \end{array} \right. \tag{108}$$

The method is classical in numerical analysis; therefore, we only sketch it. Let us assume that (108) is false. Then for any integer number n there exists an element $X^n \in V$ such that (where $E(0)(X^n)$ defined at (62) is the square of a norm on V)

$$\forall n > 0,\ E(0)(X^n) \simeq ||X^n||_V^2 = 1,\ \int_0^T \int_\alpha^\beta \left(\frac{\partial d^n}{\partial t}\right)^2 + \left(\frac{\partial \psi^n}{\partial t}\right)^2 \leq \frac{1}{n}, \tag{109}$$

where (Ψ^n, d^n) is the solution of (99) with the initial conditions X^n. From the weak compactness of unit balls in Hilbert spaces, one can extract from X_n a subsequence denoted $X^{n'}$ and such that

$$X^n \to X^* \text{ in } V \text{ weakly.}$$

But one has also

$$(\psi^{n'}, d^{n'}) \to (\psi^*, d^*) \text{ weakly in the space of solutions (see section 3),}$$

where (ψ^*, d^*) is the solution of (99) with the initial condition $X^* \in V$. However, from (109), one deduces that

$$\int_0^T \int_\alpha^\beta \left(\frac{\partial d^*}{\partial t}\right)^2 + \left(\frac{\partial \psi^*}{\partial t}\right)^2 \leq \varliminf_{n' \to \infty} \int_0^T \int_\alpha^\beta (d^{n'})^2 + (\psi^{n'})^2 = 0,$$

and therefore

$$\frac{\partial \psi^*}{\partial t} = \frac{\partial d^*}{\partial t} = 0 \qquad \forall (x,t) \in]\alpha, \beta[\times]0, T[\,.$$

Setting

$$\bar{\psi}^* = \frac{\partial \psi^*}{\partial t} \text{ and } \bar{d}^* = \frac{\partial d^*}{\partial t},$$

one deduces that, on the one hand,

$$\bar{\psi}^* = \bar{d}^* = 0 \ \ \forall (x,t) \in]\alpha, \beta[\times]0, T[\,;$$

following the proof given in section 3, one deduces that

$$\bar{\psi}^* = 0 \ \ \forall (x,t) \in]0, L[\times]0, T[\text{ and } \bar{d}^* = 0 \ \ \forall (x,t) \in]L_1, L_2[\times]0, T[\,.$$

Thus, ψ^* and d^* are both time independent. From (99) we obtain that

$$\text{(110)} \qquad \begin{cases} -(c_f^2 - U^2)\dfrac{\partial^2 \psi^*}{\partial x^2} = bU \dfrac{\partial d^*}{\partial x} \chi_{[L_1, L_2]} \ \ \forall (x,t) \in]0, L[\times]0, T[\,, \\ -c_s^2 \dfrac{\partial^2 d^*}{\partial x^2} = -a c_f^2 \dfrac{\partial \psi^*}{\partial x} \ \ \forall (x,t) \in]L_1, L_2[\times]0, T[\,, \\ \dfrac{\partial \psi^*}{\partial x}(0) = \dfrac{\partial \psi^*}{\partial x}(L) = d^*(L_1) = d(L_2) = 0. \end{cases}$$

Because $U < U_c$, we can deduce that

$$d^* = 0, \ \ \psi^* = \text{constant},$$

and because $\psi^* \in H_m^1(]0, L[)$, one has (see (7))

$$\int_{L_1}^{L_2} \psi^* = 0$$

and finally

$$\psi^* = 0.$$

This is in contradiction to (109), and the inequality (108) is true.

Step 6. Let us consider for a given set of initial conditions—say, $(\Phi_0, \Phi_1, D_0, D_1)$—the solution (φ, z) of the coupled system. Then we set

$$\text{(111)} \qquad \begin{cases} \tilde{\psi}(x,t) = \tilde{\psi}_0(x) + \displaystyle\int_T^t \psi(x,s) ds, \\ \tilde{d}(x,t) = \tilde{d}_0(x) + \displaystyle\int_T^t d(x,s) ds. \end{cases}$$

The initial terms $(\tilde{\psi}_0, \tilde{d}_0)$ are chosen such that $(\tilde{\psi}, \tilde{d})$ is also a solution of the adjoint system (52). The definitions of these terms are specified in (54). By applying the inequalities (76) and (108), we deduce that there exists a constant c such that

$$\left\{ \begin{array}{l} c\big[||\Phi_0||^2_{0,OL} + ||\Phi_1||^2_{[H^1(]O,L[)]'} + ||D_0||^2_{0,L_1L_2} + ||D_1||^2_{1,L_1L_2}\big] \\ \qquad \leq \displaystyle\int_0^T \int_\alpha^\beta \left(\frac{\partial \tilde{\psi}}{\partial t}\right)^2 + \left(\frac{\partial \tilde{d}}{\partial t}\right)^2, \end{array} \right. \tag{112}$$

and finally from (111), we deduce that Theorem 2 is proved. □

8. Convergence of the least square control to the exact control when $\varepsilon \to 0$. In practical application it is more convenient to use the optimal control $(u^\varepsilon, w^\varepsilon)$. However, the stability of the method is strongly dependent on the convergence when ε tends to zero to the exact control, when it exists. Let first state the result that we are proving in this section.

THEOREM 5. *Let us assume that the assumptions of Theorem* 4 *are satisfied. Then one has the following convergence result:*

$$\lim_{\varepsilon \to 0} ||u^\varepsilon - u^0||^2_{L^2(]\alpha,\beta[\times]0,T[)} + ||w^\varepsilon - w^0||^2_{L^2(]\alpha,\beta[\times]0,T[)} = 0.$$

Remark 14. The order of convergence with respect to ε can be obtained by the computation of the term of order one in the asymptotic expansion with respect to ε. But it depends on additional regularity of the initial data in order to make sure that u^1 and w^1 are effectively in the space $L^2(]\alpha,\beta[\times]0,T[)$. Nevertheless, from interpolation techniques, one can adjust the order of convergence with respect to ε when this regularity is not satisfied. We refer to the book by Lions [22].

Proof of Theorem 5. Let us set $(u, w) = (u^0, w^0)$ in the criterion J^ε. Because it is an exact control, the corresponding final state is zero. Thus we obtain the following upper bound:

$$J^\varepsilon(u^\varepsilon, w^\varepsilon) \leq J^\varepsilon(u^0, w^0) = \frac{\varepsilon}{2} \int_0^T \int_\alpha^\beta (u^0)^2 + (w^0)^2.$$

Therefore the sequence $(u^\varepsilon, w^\varepsilon)$ is bounded with respect to ε in the space $\big[L^2(]\alpha,\beta[\times]0,T[)\big]^2$. Furthermore, one has

$$\begin{aligned} &\frac{A}{2}\int_0^L \left|\frac{\partial \varphi}{\partial t}(x,T)\right|^2 + \frac{B}{2}\int_0^L \left|\frac{\partial \varphi}{\partial x}(x,T)\right|^2 \\ &+ \frac{C}{2}\int_{L_1}^{L_2} \left|\frac{\partial z}{\partial t}(x,T)\right|^2 + \frac{D}{2}\int_{L_1}^{L_2} \left|\frac{\partial z}{\partial x}(x,T)\right|^2 \leq c\varepsilon. \end{aligned}$$

Thus, from classical analysis, we deduce the "strong" continuity of the solution of the coupled model with respect to the right-hand side $(u^\varepsilon, w^\varepsilon)$ in $\big[L^2(]\alpha,\beta[\times]0,T[)\big]^2$, and therefore the strong convergence of the solution $(\varphi^\varepsilon, z^\varepsilon)(x,t)$ tends also to $(\varphi^0, z^0)(x,t)$ with ε in the space mentioned in Theorem 5. The last point concerns the strong convergence of the control $(u^\varepsilon, w^\varepsilon)$ to the H.U.M. control (u^0, w^0) in the space $\big[L^2(]\alpha,\beta[\times]0,T[)\big]^2$. From the upper bound mentioned at the beginning of the proof, one has the following:

$$\int_0^T \int_\alpha^\beta (u^\varepsilon - u^0)^2 + (w^\varepsilon - w^0)^2$$
$$= \int_0^T \int_\alpha^\beta (u^\varepsilon)^2 + (w^\varepsilon)^2 - 2\int_0^T \int_\alpha^\beta u^\varepsilon u^0 + w^\varepsilon w^0 + \int_0^T \int_\alpha^\beta (u^0)^2 + (w^0)^2$$
$$\leq 2\left[\int_0^T \int_\alpha^\beta (u^0)^2 + (w^0)^2 - (u^\varepsilon u^0 + w^\varepsilon w^0)\right] \to 0 \quad \text{when } \varepsilon \to 0.$$

This completes the proof of Theorem 5. ∎

Remark 15. In practical applications it is usually more convenient to use the least square control because it can be more easily computed. But this is only true if the system can be exactly controlled. In other words, when ε tends to zero, there could be a singular perturbation known as a stiff problem if there were no exact control in the space $[L^2(\alpha, \beta[\times]0, T[)]^2$. Nevertheless, if only one control is used (the structural one), we proved that the system could be exactly controlled, but nothing has been proved concerning the space of initial data which could be exactly controlled with a control u^0 in $L^2(\alpha, \beta[\times]0, T[)$. There exist a few tricks in order to obtain an exact control for any initial data with finite energy (for both the fluid and the structure). But unfortunately the control are no longer in L^2. They contain Dirac distributions at both time extremities. The details of the method mentioned also in [7] for fluid-structure problems can be found again in [21].

9. Conclusion. A simple one dimensional model in aeroacoustics, coupled with a structure, has been discussed in this paper. The goal was to analyze the solutions of the coupled model and to point out the possibility of a flutter phenomenon. Then we discussed the exact controllability of any perturbations in the fluid or/and in the structure. For sake of simplicity in the explanation we have restricted the discussion to a membrane model for the structure. It has been proved that one control applied to a small part of the structure is sufficient for obtaining an exact control. However, a stability (inverse inequality) result has only been proved for a couple of controls: one applied to the flexible structure and the other on the walls of the flow duct. This restriction is due to mathematical difficulties but also to a subsonic shock wave which can appear at the extremity of the structure where the rotations can be discontinuous. (In fact, the terms involved are $U\frac{\partial z}{\partial x}(L_1, t)$ and $U\frac{\partial z}{\partial x}(L_2, t)$.) When $U = 0$ (i.e., no steady flow), the difficulty disappears. One basic point is that the exact control still works if the critical velocity for the flutter apparition is overtaken. But obviously the cost of the control can also be exponentially increasing with respect to time. A lot of improvements can be suggested. First one could introduce a wall law in order to take into account the viscosity of the fluid. But the mathematical structure of the operator is changed and nonlocal energy bounds from the control area are much more complicated to derive (see, for instance, the difficulties encountered for Stokes equations [8], [2], [3] [10]). Another point is that the one dimensional model is certainly insufficient in order to reproduce reality. Some improvements could be forecast by a better modelling of the control devices themselves. Two dimensional modelling is clearly more realistic, especially for reproducing local waves which can appear at the interface between the fluid and the structure. One possibility seems to use an asymptotic method, where the small parameters are the transverse dimensions of the structure compared to the length of the flow duct and the ratio between the smallest period of the coupled system and the time control delay T.

Acknowledgment. The authors are very grateful to Caroline Fabre for her valuable advice and suggestions when we were writing this paper.

REFERENCES

[1] L. Cagniard, *Reflection and Refraction of Progressive Seismic Wave* (translated and revisited by A. Flinn and C. H. Dix), McGraw–Hill, New York, 1962.

[2] J. M. Coron, *Sur la contrôlabilité approchée des équations de Naviers–Stokes* 2-*D avec des conditions de glissement de Navier*, ESAIM Control Optim. Calc. Var., 1 (1996), pp. 35–75.

[3] J. M. Coron, *Return method and flow control*, Communication au colloque à la mémoire de J. L. Lions, Collège de France, Paris, 2002.

[4] Ph. Destuynder, *A mathematical analysis of a smart-beam which is equipped with piezoelectric actuators*, Control Cybernet., 28 (1999), pp. 503–530.

[5] Ph. Destuynder, *Strange aeroacoustic waves in a flow duct with flexible walls*, in Proceedings of the XVII Congreso de Ecuaciones Diferenciales y Aplicationes, L. Ferragut and A. Santos, eds., Universidad de Salamanca, Spain, 2001, pp. 67–90.

[6] Ph. Destuynder and E. Gout-d'Hénin, *Existence and uniqueness of a solution to an aeroacoustic model*, Chinese Ann. Math. Ser. B, 23 (2002), pp. 11–24.

[7] Ph. Destuynder, *Few remarks on noise control in fluid-structure modelling*, Revue Européenne des Eléments Finis, 11 (2002), pp. 149–171.

[8] C. Fabre, *Quelques problèmes de controllabilité approché pour des problèmes paraboliques linéaires et non linéaires*, Habilitation à diriger des recherches en Mathématiques, Université Pierre et Curie, Paris VI, Paris, 1996.

[9] Y. C. Fung, *Foundations of Solid Mechanics*, Prentice–Hall, London, 1965.

[10] A. V. Fursikov and O. Y. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Series 34, Research Institute of Mathematics, Global Analysis Research Center, Seoul National University, Seoul, Korea, 1996.

[11] E. Gout d'Henin, *Caractérisation, Analyse et Contrôle des Ondes de Stoneley en Interaction Fluide-Structure*, Thèse de l'université de Poitiers, Poitiers, France, 2002.

[12] M. A. Galland, O. Sellen, and O. Hillbrunner, *Experimental and numerical investigation of noise reduction in a lined duct by hybrid/passive control*, in Proceedings of the 5th CEA-ASC Workshop on Turbomachinery Noise and Duct Acoustics, Eindhoven, The Netherlands, 2001.

[13] D. S. Jones and J. D. Morgan, *A linear model of a finite Helmholtz instability*, Proc. Roy. Soc. London, A338 (1974), pp. 17–41.

[14] V. Komornik, *La méthode des Multiplicateurs*, Rech. Math. Appl. 27, Masson, Paris, 1994.

[15] J. Necas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson, Paris, 1967.

[16] I. Lasiecka and C. Lebiedzik, *Uniform stability in structural acoustic systems with thermal effects and nonlinear boundary damping*, Control Cybernet., 28 (1999), pp. 557–581.

[17] S. Lévy, *Acoustique Industrielle et Aéroacoustique*, Hermès, Paris, 2001.

[18] J. L. Lighthill, *Waves in Fluids*, Cambridge University Press, Cambridge, UK, 1978.

[19] J. L. Lions and E. Magenes, *Problèmes aux Limites non Homogènes et Applications*, Dunod, Paris, 1968.

[20] J. L. Lions, *Contrôle Optimal de Systèmes Gouvernés par des Équations aux Dérivées Partielles*, Dunod, Paris, 1969.

[21] J. L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués*, Rech. Math. Appl. 8, Masson, Paris, 1988.

[22] J. L. Lions, *Perturbations singulières dans les problèmes aux limites et en contrôle optimal*, Lecture Notes in Math. 323, Springer-Verlag, Berlin, 1973.

[23] A. E. H. Love, Proceedings of the London Mathematical Society, Series 1, 1903, pp. 37–62.

[24] A. E. H. Love, *Some Problems of Geodynamics*, Cambridge University Press, Cambridge, UK, 1911, pp. 165–178.

[25] S. Madanshetty and B. T. Chu, *Active sound extraction for noise control in the presence of duct mean flow*, in Proceedings of the 130th ASA Meeting, St. Louis, MO, 1995.

[26] J. Miklowitz, *The Theory of Elastic Waves and Waveguides*, North–Holland, Amsterdam, 1984.

[27] R. Stoneley, *Elastic waves at the surface of separation of two solids*, Proc. Roy. Soc. London Ser. A, 106 (1924), pp. 416–428.

[28] S. C. Snyder, *Active Noise Control Primer*, AIP Press, Springer-Verlag, Berlin, 2002.

[29] E. Zuazua, *Stability and decay for a class of nonlinear hyperbolic problems*, Asymptot. Anal., 1 (1988), pp. 1–28.

[30] E. Zuazua, *Exact controllability for semilinear wave equations in one space dimension*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 109–129.

[31] E. Zuazua, *Exact boundary controllability of a semilinear wave equation*, in Nonlinear Partial Differential Equations and Their Applications, Pitman Res. Notes Math. Ser. 220, H. Brezis and J. L. Lions, eds., Longman, Harlow, UK, 1991, pp. 357–391.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 2018–2033

GUIDED WAVES IN A PHOTONIC BANDGAP STRUCTURE WITH A LINE DEFECT*

HABIB AMMARI† AND FADIL SANTOSA‡

Abstract. Numerical simulations have shown that when a line of defects is introduced into a photonic bandgap structure, waves can be guided along the line. It has been conjectured that the mechanism responsible for the guidance phenomenon is the introduction of a spectrum in the bandgap by the defect. The purpose of this work is to give a mathematical framework for understanding this phenomenon. We show that there exist solutions of the scalar wave equation that is localized near the line defects that behave like guided modes. Moreover, these solutions can be parameterized by a frequency spectrum that is continuous and can cover parts of the original bandgap. The frequency of the guided modes depends on a wave number parameter and can be interpreted as a dispersion relation. We illustrate the main findings of the investigation in a numerical example.

Key words. photonic bandgap structures, line defect, guided waves, spectral analysis

AMS subject classifications. 35P99, 78A10, 78A48

DOI. 10.1137/S0036139902404025

1. Introduction.

The purpose of this work is the study of wave propagation in an infinite periodic structure with a line defect. The medium under consideration is periodic with the exception of a row of identical defects. Without the defect, the periodic medium is assumed to have a bandgap, which is an interval of frequencies at which waves cannot propagate.

It has been observed in numerical simulations that line defects can support guided modes which propagate along the row of defect. Moreover, these modes are highly confined near the row, with frequencies lying in the bandgap of the infinite periodic structure.

When a defect is introduced into the perfect array, i.e., a perturbation with compact support, it is possible to create a midgap defect mode, which is a highly localized standing wave whose frequency ω lies in the bandgap [6, 7]. What is less clear is how a medium with an infinite line of defects behaves. It has been observed through numerical experiments that propagating modes which are localized near the line of defects can be produced [4, 14, 13]. The goal of this work is to understand how these propagating modes are created and what properties they possess. For this purpose, we use the theory developed by Figotin and Klein [6, 7]. This work is a first step in rigorously explaining the guidance phenomenon in photonic bandgap structures.

In this work, we model wave propagation using the scalar Helmholtz equation in two dimensions, corresponding to transverse electric (TE) mode electromagnetic waves. The spectrum, which is the frequency parameter of Helmholtz's equation, is analyzed. What this work does not address is the fundamental question of whether the spectrum in question is absolutely continuous. Instead, we prove that guided

*Received by the editors March 13, 2002; accepted for publication (in revised form) January 12, 2004; published electronically August 19, 2004.

http://www.siam.org/journals/siap/64-6/40402.html

†Centre de Mathématiques Appliquées, CNRS UMR 7641 & Ecole Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr). The work of this author was partially supported by ACI Jeunes Chercheurs (0693) from the Ministry of Education and Scientific Research, France.

‡School of Mathematics, University of Minnesota, 127 Vincent Hall, 206 Church Street SE, Minneapolis, MN 55455 (santosa@math.umn.edu). The work of this author was partially supported by the National Science Foundation.

modes can be created by introducing line defects. We also provide a description for these guided modes. Our work does not answer the question of whether the bandgap could be filled in by the new spectrum.

The paper is organized as follows. The problem statement is given in section 2. This is followed by a brief review of known results about waves in an infinite periodic medium. Section 4 considers the spectral problem in an infinite strip with Bloch conditions on the sides. The strip problem is the basis from which we build a framework to analyze the spectral problem in a medium with a line defect. The strip problem with a defect is analyzed in section 5. We remark on the construction of guided waves in section 6. Section 7 contains numerical calculations that illustrate the main findings of this work. The paper ends with a discussion.

2. Problem statement. Consider a two-dimensional periodic medium characterized by the dielectric constant $\epsilon_{\rm p}(x_1, x_2)$. We assume that it is an L^∞ function satisfying

$$0 < \epsilon_- \le \epsilon_{\rm p}(x) \le \epsilon_+ < \infty$$

and is unit periodic, i.e.,

$$\epsilon_{\rm p}(x_1 + 1, x_2) = \epsilon_{\rm p}(x_1, x_2), \quad \epsilon_{\rm p}(x_1, x_2 + 1) = \epsilon_{\rm p}(x_1, x_2).$$

To this perfect array, we introduce a line defect which is represented by a perturbation to the dielectric property $\delta\epsilon(x_1, x_2)$. The perturbation is confined to the cells over the x_1-axis and is periodic in x_1,

$$\begin{aligned} &\delta\epsilon(x_1, x_2) = 0, && |x_2| > 1/2, \\ &\delta\epsilon(x_1 + 1, x_2) = \delta\epsilon(x_1, x_2). \end{aligned}$$

The medium with defect then has dielectric constant

$$\epsilon(x_1, x_2) = \epsilon_{\rm p}(x_1, x_2) + \delta\epsilon(x_1, x_2).$$

It is assumed that ϵ is still a strictly positive bounded measurable function. The object of this work is the study of Helmholtz equation

$$\Delta u + \omega^2 \epsilon u = 0, \quad (x_1, x_2) \in \mathbb{R}^2. \tag{1}$$

The Helmholtz equation is a model for TE-mode electromagnetic wave propagation in two dimensions. We view the frequency squared, ω^2, as the spectral parameter and investigate the spectrum of the operator

$$-\frac{1}{\epsilon}\Delta.$$

The approach is to compare the spectrum with that when the medium is periodic, i.e., one with dielectric constant $\epsilon_{\rm p}(x)$.

It is well known that the medium $\epsilon_{\rm p}(x)$ can have bandgaps, i.e., intervals of values ω for which propagating waves cannot exist [5, 10]. While necessary conditions under which bandgaps exist in general are not known, Figotin and Kuchment have produced an example of a high-contrast periodic medium where bandgaps exist and can be characterized [8, 9].

Moreover, it is also known that when a defect is introduced into the perfect array, i.e., a perturbation to $\epsilon_{\rm p}$ with compact support, it is possible to create a defect

mode, which is a solution to Helmholtz's equation with exponential decay, and with frequency ω which lies in the bandgap [6, 7]. The work cited also provided estimates of the decay rates, which we do not use in the present work. A lot less is known about the case in which there is an infinite line of defects such as $\epsilon(x)$. Although numerical experiments have provided evidence that waves that are localized near the line of defects can be produced, little analytical results are known for this case. The goal of this work is to investigate the spectral problem of (1) to reveal the guidance phenomena.

3. Periodic problem. We first consider the properties of the periodic structure without defect. The known results, particularly those that will be useful for this work, are presented. We consider the Bloch waves $w(x, \alpha)$ satisfying

$$\Delta w + \omega^2 \epsilon_{\mathrm{p}} w = 0, \quad x \in \mathbb{R}^2, \tag{2a}$$

$$w(x_1 + 1, x_2, \alpha_1, \alpha_2) = w(x_1, x_2, \alpha_1, \alpha_2) e^{i\alpha_1}, \tag{2b}$$

$$w(x_1, x_2 + 1, \alpha_1, \alpha_2) = w(x_1, x_2, \alpha_1, \alpha_2) e^{i\alpha_2}. \tag{2c}$$

When a solution to the above exists for a given vector α, the function $w(x, \alpha)$ corresponds to plane wave-like solutions with the vector $\alpha = (\alpha_1, \alpha_2)$ playing the role of wave number.

Another way to characterize this eigenvalue problem is by introducing

$$w(x, \alpha) = \psi(x, \alpha) e^{i\alpha \cdot x}.$$

It can be shown that ψ satisfies

$$(\nabla + i\alpha) \cdot (\nabla + i\alpha)\psi + \omega^2 \epsilon_{\mathrm{p}} \psi = 0 \tag{3a}$$

with periodic boundary conditions

$$\psi(x_1 + 1, x_2, \alpha_1, \alpha_2) = \psi(x_1, x_2, \alpha_1, \alpha_2), \tag{3b}$$

$$\psi(x_1, x_2 + 1, \alpha_1, \alpha_2) = \psi(x_1, x_2, \alpha_1, \alpha_2). \tag{3c}$$

We can view the above as an eigenvalue problem to find ω given the vector α. It can be shown that this eigenvalue problem admits an infinity of solutions, with frequencies $\omega_n(\alpha)$, where the index $n = 1, 2, \ldots$ provides an ordering for the eigenvalues. The function $\omega_n(\alpha)$ is referred to as the dispersion relation for the nth modes. The corresponding eigenfunctions, the Bloch waves, are denoted by $w_n(x, \alpha)$, and it will be useful to consider the pairs $\{\omega_n(\alpha), w_n(x, \alpha)\}_{n=1}^{\infty}$.

The periodicity of the eigenvalue problem above induces periodicity in the dispersion relation $\omega_n(\alpha)$ as a function of α. That is, $\omega_n(\alpha)$ is periodic with period $P := [0, 2\pi]^2$. If $\epsilon_{\mathrm{p}} \in C^{\infty}([0, 1]^2)$, ω_n are analytic functions of α everywhere on P except on subsets of measure zero where their multiplicity changes. In particular, they are continuous for all α [15, 3]. The set of Bloch waves, as α varies in P, and $n = 1, 2, \ldots$, is complete in $L^2(\mathbb{R}^2)$ [11]. Therefore, any solution of the wave equation

$$\Delta u + \omega^2 \epsilon_{\mathrm{p}} u = 0$$

can be expressed as a linear combination of Bloch waves.

Let I_n denote the interval of values of $\omega_n^2(\alpha)$ for $\alpha \in P$. Then the spectrum of the periodic operator $-\Delta/\epsilon_{\mathrm{p}}$ is

$$\Sigma(\epsilon_{\mathrm{p}}) = \bigcup_{n=1}^{\infty} I_n.$$

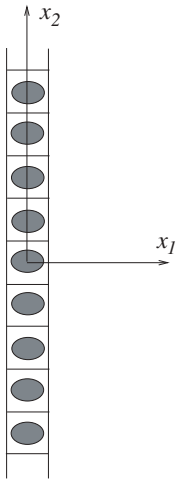


FIG. 1. *A vertical strip from the periodic medium.*

It is further known that for some $\epsilon_{\mathrm{p}}(x)$, there are bandgaps, i.e., $\Sigma(\epsilon_{\mathrm{p}})$ does not cover $[0,\infty)$. Let us suppose this is the case and let this gap be the interval $\Gamma(\epsilon_{\mathrm{p}}) =]a,b[$. Physically, it means that a wave of frequency $\lambda^2 \in \Gamma(\epsilon_{\mathrm{p}})$ cannot propagate in the medium. The existence of the gap is a problem studied in [8, 9], where it is shown that for a high-contrast medium of a particular geometry, the bandgap can be characterized. More general questions about gaps are still open [10].

3.1. Green's function. The Green's function for Helmholtz's equation with a periodic medium satisfies

$$\Delta G + \lambda^2 \epsilon_{\mathrm{p}} G = \delta(x-y).$$

The following is described in [6, 7, 2]:

- For $\lambda^2 \in \Gamma(\epsilon_{\mathrm{p}})$, the Green's function is exponentially decaying away from y:

$$|G(x,y;\lambda)| \leq C_1 e^{-C_2|x-y|}. \tag{4a}$$

- The function

$$G(x,y;\lambda) - \frac{1}{2\pi}\log|x-y| \tag{4b}$$

  is continuous for $|x-y| \to 0$.

4. The strip. With the introduction of the line defect as described in section 2, the medium loses periodicity in the x_2-direction. It is, however, still periodic in the x_1-direction. We will exploit this fact in our analysis. For now, we investigate the periodic problem on the strip as shown in Figure 1, and we view the x_1-direction quasi-momentum α_1 as a parameter on the interval $[0, 2\pi]$.

4.1. Characterization of the periodic problem in the strip. Let $O = (-\frac{1}{2}, \frac{1}{2}) \times \mathbb{R}$ denote the strip in Figure 1. Consider the Bloch wave $w(x,\alpha)$ as a function of x and α_2, parameterized by α_1. For each $\alpha_1 \in [0, 2\pi]$, we solve

$$\Delta u_{\alpha_1} + \nu_{\alpha_1}^2(\alpha_2)\epsilon_{\mathrm{p}}(x)u_{\alpha_1} = 0 \quad \text{in } O \tag{5a}$$

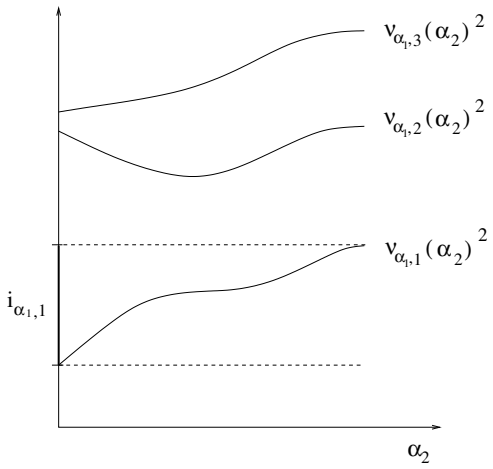


FIG. 2. *Dispersion relation $\nu_{\alpha_1,n}(\alpha_2)$ for a fixed α_1.*

with boundary conditions

$$u_{\alpha_1}(x_1+1,x_2) = u_{\alpha_1}(x_1,x_2)e^{i\alpha_1}, \tag{5b}$$

$$u_{\alpha_1}(x_1,x_2+1) = u_{\alpha_1}(x_1,x_2)e^{i\alpha_2}. \tag{5c}$$

Consider the problem for a fixed α_1. Now choose a value for α_2. The eigenvalue problem (5) admits an infinity of solutions $\{\nu_{\alpha_1,n}(\alpha_2), u_{\alpha_1,n}(x,\alpha_2)\}_{n=1}^{\infty}$. The dispersion relation and the Bloch waves in (2) are recovered as

$$\omega_n(\alpha_1,\alpha_2) = \nu_{\alpha_1,n}(\alpha_2), \quad w_n(x,\alpha_1,\alpha_2) = u_{\alpha_1,n}(x,\alpha_2).$$

For each fixed α_1, we view $\nu_{\alpha_1,n}(\alpha_2)$ as a dispersion relation. A sketch of what such a dispersion relation might look like is provided in Figure 2.

For a fixed α_1, the spectrum of the operator on the strip is

$$\sigma_{\alpha_1}(\epsilon_{\mathrm{p}}) = \bigcup_{n=1}^{\infty} i_{\alpha_1,n},$$

where $i_{\alpha_1,n}$ is the interval for the values of $\nu^2_{\alpha_1,n}(\alpha_2)$ for $\alpha_2 \in [0,2\pi]$. We recover

$$I_n = \bigcup_{\alpha_1} i_{\alpha_1,n} \quad \text{and} \quad \Sigma(\epsilon_{\mathrm{p}}) = \bigcup_{\alpha_1} \sigma_{\alpha_1}(\epsilon_{\mathrm{p}}).$$

The spectrum $\sigma_{\alpha_1}(\epsilon_{\mathrm{p}})$ may have a bandgap for a given α_1. Let us denote it by $\gamma_{\alpha_1}(\epsilon_{\mathrm{p}})$. The bandgap of the periodic medium $\Gamma(\epsilon_{\mathrm{p}})$ is contained in the intersection of γ_{α_1}

$$\Gamma(\epsilon_{\mathrm{p}}) \subset \bigcap_{\alpha_1} \gamma_{\alpha_1}(\epsilon_{\mathrm{p}}).$$

4.2. Green's function in the strip. We will next study the Green's function for the strip for a fixed α_1. For frequency λ such that λ^2 is in the gap $\Gamma(\epsilon_{\mathrm{p}})$, the Green's function can be constructed using the partial Floquet transform. The strip

Green's function satisfies

$$\Delta g_{\alpha_1} + \lambda^2 \epsilon_{\mathrm{p}} g_{\alpha_1} = \delta_{\alpha_1}(x-y) = \sum_{j\in\mathbb{Z}} \delta(x_1 + j - y_1, x_2 - y_2) e^{ij\alpha_1}, \tag{6a}$$

$$g_{\alpha_1}(x_1+1, x_2, y_1, y_2; \lambda) = g_{\alpha_1}(x_1, x_2, y_1, y_2; \lambda) e^{i\alpha_1}. \tag{6b}$$

LEMMA 1. *Suppose that $\lambda^2 \in \Gamma(\epsilon_{\mathrm{p}})$. Then the Green's function is related to the whole-space Green's function through*

$$g_{\alpha_1}(x_1, x_2, y_1, y_2; \lambda) = \sum_{j\in\mathbb{Z}} G(x_1 + j, x_2, y_1, y_2; \lambda) e^{ij\alpha_1}. \tag{7}$$

Moreover, $g_{\alpha_1}(\cdot, y; \lambda)$ is in $L^2_{\mathrm{loc}}(\mathbb{R}^2)$.

Proof. Suppose that $\lambda^2 \in \Gamma(\epsilon_{\mathrm{p}})$. Then, clearly, for any fixed $y = (y_1, y_2)$ $x = (x_1, x_2) \to \sum_{j\in\mathbb{Z}} G(x_1 + j, x_2, y_1, y_2; \lambda) e^{ij\alpha_1}$ defines a function in $L^2_{\mathrm{loc}}(\mathbb{R}^2)$ which satisfies (6b) almost everywhere on $\mathbb{R}^2$. Indeed, we have in a distribution sense

$$\begin{aligned}
&(\Delta + \lambda^2 \epsilon_{\mathrm{p}}) \sum_{j\in\mathbb{Z}} G(x_1 + j, x_2, y_1, y_2; \lambda) e^{ij\alpha_1} \\
&= \sum_{j\in\mathbb{Z}} (\Delta + \lambda^2 \epsilon_{\mathrm{p}}) G(x_1 + j, x_2, y_1, y_2; \lambda) e^{ij\alpha_1} \\
&= \sum_{j\in\mathbb{Z}} \delta(x_1 + j - y_1, x_2 - y_2) e^{ij\alpha_1}.
\end{aligned}$$

Since $\lambda^2 \in \Gamma(\epsilon_{\mathrm{p}})$, the strip Green's function g_{α_1} defined by (7) is the unique solution to (6). ∎

Let us define a regular part of g_{α_1} as the series

$$r_{\alpha_1}(x_1, x_2, y_1, y_2; \lambda) = \sum_{j\in\mathbb{Z}, j\neq 0} G(x_1 + j, x_2, y_1, y_2; \lambda) e^{ij\alpha_1}, \tag{8}$$

whose terms are well defined for $(x_1, y_1) \in (-\frac{1}{2}, \frac{1}{2})^2$. By (4), the series in (8) converges uniformly for (x, y) in compact subsets of $O \times O$. Therefore, as for the local behavior of g_{α_1}, i.e., when $|x - y| \to 0$, the following result of logarithmic singularity holds from (4).

LEMMA 2. *Suppose that $\lambda^2 \in \Gamma(\epsilon_{\mathrm{p}})$. Then the function*

$$g_{\alpha_1}(x, y; \lambda) - \frac{1}{2\pi} \log|x - y|$$

is continuous as $|x - y| \to 0$.

Proof. In any open set $\mathcal{K}$ that does not contain a point $(y_1 - j, y_2)$, $j \in \mathbb{Z}$, the regular part r_{α_1} satisfies

$$(\Delta + \lambda^2 \epsilon_{\mathrm{p}}) r_{\alpha_1} = 0 \text{ in } \mathcal{D}'(\mathcal{K}).$$

But r_{α_1} is in $L^2_{\mathrm{loc}}(\mathbb{R}^2)$ so that, applying classical results on elliptic regularity together with the Sobolev embedding theorem, we obtain that r_{α_1} is a continuous function in $\mathcal{K}$. Lemma 2 then follows immediately from the logarithmic behavior (4b) of $G(x, y; \lambda)$. ∎

For $(x_1, y_1) \in (-\frac{1}{2}, \frac{1}{2})^2$ it follows from the exponential decay property (4) that the regular part r_{α_1} defined by (8) is a continuous function of α_1 and therefore the following continuity result holds.

LEMMA 3. *If $\lambda^2 \in \Gamma(\epsilon_{\rm p})$, then the map $\alpha_1 \in [0, 2\pi] \mapsto$ the operator with kernel $g_{\alpha_1}(.,.;\lambda) \in \mathcal{D}'(\mathbb{R}^2) \times \mathcal{D}'(\mathbb{R}^2)$, mapping $L^2(O)$ to $L^2(O)$ is continuous.*

If $\lambda^2 \in \Gamma(\epsilon_{\rm p})$, then, by (4), all the terms in (7) are exponentially decaying as $|x_2 - y_2| \to +\infty$, and so is their sum.

LEMMA 4. *Suppose that $\lambda^2 \in \Gamma(\epsilon_{\rm p})$. Then there exist positive constants C_2 and C_3 such that*

$$|g_{\alpha_1}(x, y; \lambda)| \leq C_3 e^{-C_2|x_2 - y_2|} \quad \text{as } |x_2 - y_2| \to \infty.$$

Proof. Combining

$$\sqrt{(x_1 + j - y_1)^2 + (x_2 - y_2)^2} \leq |x_1 + j - y_1| + |x_2 - y_2|$$

with the fact that the series $\sum_{j\in\mathbb{Z}} e^{-C_2|x_1+j-y_1|}$ is uniformly bounded for $(x_1, y_1) \in (-\frac{1}{2}, \frac{1}{2})^2$, it is readily seen by using the exponential decay property (4a) of G that for $(x_1, y_1) \in (-\frac{1}{2}, \frac{1}{2})^2$ the estimates

$$\begin{aligned} |g_{\alpha_1}(x, y; \lambda)| &\leq C_1 \sum_{j\in\mathbb{Z}} e^{-C_2\sqrt{(x_1+j-y_1)^2+(x_2-y_2)^2}} \\ &\leq C_1 e^{-C_2|x_2-y_2|} \sum_{j\in\mathbb{Z}} e^{-C_2|x_1+j-y_1|} \leq C_3 e^{-C_2|x_2-y_2|} \end{aligned}$$

hold for some positive constant C_3. ∎

4.3. Preliminary results. The following preliminary results will be useful in the subsequent sections for understanding the behavior of the strip when a defect is introduced.

We shall use quite standard quasi-periodic L^2-based Sobolev spaces to measure regularity of functions which satisfy the Bloch condition in the x_1-direction. The space $L^2_{\alpha_1}(O)$ is the set of restrictions on O of functions u such that $e^{-i\alpha_1 x_1}u(x_1, x_2) \in L^2(\mathbb{R}^2/2\pi\mathbb{Z})$. The space $H^1_{\alpha_1}(O)$ is the set of restrictions on O of functions u which satisfy $e^{-i\alpha_1 x_1}u(x_1, x_2) \in H^1(\mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R})$. These functions are, along with all their first derivatives, in $L^2(O)$ and satisfy the Bloch condition

$$u(1, x_2) = u(0, x_2)e^{i\alpha_1}.$$

We also introduce for $\gamma > 0$ the weighted Sobolev space:

$$\mathcal{H}_{\alpha_1,\gamma}(O) = \left\{ f \in H^1_{\alpha_1}(O) : \int_O e^{\gamma|x_2|}(|f|^2 + |\nabla f|^2)(x_1, x_2)\, dx_1\, dx_2 < +\infty \right\}.$$

Let Ω be a region centered at the origin

$$\Omega \subset\subset \left\{ |x_1| < \frac{1}{2}, |x_2| < h, h > 0 \right\}.$$

The defect we will later introduce on the strip is supported in Ω. Define the operator

$$\begin{aligned} &A_{\alpha_1}(\lambda) : L^2_{\alpha_1}(O) \to L^2_{\alpha_1}(O), \\ &u \mapsto A_{\alpha_1}(u) = \int_\Omega g_{\alpha_1}(x, y; \lambda)q(y)u(y)dy, \end{aligned}$$

where $q \in L^\infty(\overline{\Omega})$ is either $q \geq 0$ or $q \leq 0$ almost everywhere in Ω.

Since the Green's function g_{α_1} decays exponentially in x_2 and has logarithmic singularity, we are guaranteed that there exists a constant $\gamma > 0$ such that $A_{\alpha_1}(\lambda)(L^2_{\alpha_1}(O)) \subset \mathcal{H}_{\alpha_1,\gamma}(O)$. Using the fact that the embedding $\mathcal{H}_{\alpha_1,\gamma}(O) \hookrightarrow L^2_{\alpha_1}(O)$ is compact, for any $\gamma > 0$ [1], we conclude that the operator $A_{\alpha_1}(\lambda)$ is a compact operator.

LEMMA 5. *For any fixed $\lambda^2 \in \Gamma(\epsilon_{\rm p})$ and $\alpha_1 \in [0, 2\pi]$, the operator $A_{\alpha_1}(\lambda)$ is a compact operator.*

The eigenvalue problem

$$A_{\alpha_1}(\lambda)(v) = \mu(\lambda, \alpha_1)v \tag{9}$$

can be rewritten as

$$\mu(\lambda, \alpha_1)v\chi_\Omega = \chi_\Omega\Big(\Delta + \lambda^2\epsilon_{\rm p}\Big)^{-1}(q\chi_\Omega v),$$

where χ_Ω is the characteristic function of the region Ω. If we set

$$\psi = \begin{cases} \sqrt{q}v\chi_\Omega & \text{if } q \geq 0, \\ \sqrt{-q}v\chi_\Omega & \text{if } q \leq 0, \end{cases}$$

we obtain that the eigenvalue problem (9) is equivalent to

$$\operatorname{sgn}(q)\mu(\lambda, \alpha)\psi = \mathcal{A}_{\alpha_1}(\lambda)\psi,$$

where

$$\mathcal{A}_{\alpha_1}(\lambda) = \begin{cases} \sqrt{q}\chi_\Omega\Big(\Delta + \lambda^2\epsilon_{\rm p}\Big)^{-1}\sqrt{q}\chi_\Omega & \text{if } q \geq 0, \\ \sqrt{-q}\chi_\Omega\Big(\Delta + \lambda^2\epsilon_{\rm p}\Big)^{-1}\sqrt{-q}\chi_\Omega & \text{if } q \leq 0. \end{cases}$$

Here $\big(\Delta + \lambda^2\epsilon_{\rm p}\big)^{-1} : L^2_{\alpha_1}(O) \to L^2_{\alpha_1}(O)$ is defined by $\big(\Delta + \lambda^2\epsilon_{\rm p}\big)^{-1}(f) = v$, where v is the unique solution in $H^1_{\alpha_1}(O)$ of $(\Delta + \lambda^2\epsilon_{\rm p})v = f$ in O.

LEMMA 6. *The operator $\mathcal{A}_{\alpha_1}(\lambda)$ is a compact self-adjoint operator.*

Since $\big(\Delta + \lambda^2\epsilon_{\rm p}\big)^{-1}$ is a monotonically decreasing, norm-continuous operator of $\lambda^2 \in \Gamma(\epsilon_{\rm p})$, the following holds.

LEMMA 7. *Let $\alpha_1 \in [0, 2\pi]$. The map $\lambda^2 \in \Gamma(\epsilon_{\rm p}) \mapsto \mathcal{A}_{\alpha_1}(\lambda)$ is norm-continuous and operator monotone decreasing for both positive and negative q.*

Now, given $\lambda^2 \in \Gamma(\epsilon_{\rm p})$ and $\alpha_1 \in [0, 2\pi]$, the spectrum of the self-adjoint compact operator $\mathcal{A}_{\alpha_1}(\lambda)$ consists of eigenvalues of finite multiplicity with 0 being the only possible point of accumulation. Let $\mu_1^+(\lambda, \alpha_1) \geq \mu_2^+(\lambda, \alpha_1) \geq \cdots \geq 0$ and $\mu_1^-(\lambda, \alpha_1) \leq \mu_2^-(\lambda, \alpha_1) \leq \ldots 0$ be infinite sequences of, respectively, the positive and negative eigenvalues of $\mathcal{A}_{\alpha_1}(\lambda)$, repeated according to their multiplicity. The sequence $\mu_i^-(\lambda, \alpha_1), i = 1, 2, \ldots$, can be obtained by applying the min-max principle to the operator $\chi_{(-\infty,0]}(\mathcal{A}_{\alpha_1}(\lambda))\mathcal{A}_{\alpha_1}(\lambda)$; similarly, we obtain $-\mu_i^+(\lambda, \alpha_1), i = 1, 2, \ldots$, by applying the min-max principle to the operator $-\chi_{[0,+\infty)}(\mathcal{A}_{\alpha_1}(\lambda))\mathcal{A}_{\alpha_1}(\lambda)$.

As a consequence of Lemma 7, the following λ-dependence of the eigenvalues $\mu_i^\pm(\lambda, \alpha_1)$ of $\mathcal{A}_{\alpha_1}(\lambda)$ holds.

LEMMA 8. *The functions $\mu_i^\pm(\lambda, \alpha_1)$ are monotonically decreasing continuous functions of $\lambda^2 \in \Gamma(\epsilon_{\rm p})$ for each $i = 1, 2, \ldots$.*

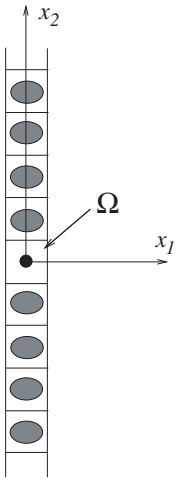


FIG. 3. *A vertical strip with a defect at the origin.*

On the other hand, the following α_1-dependence of the eigenvalues $\mu_i^{\pm}(\lambda, \alpha_1)$ on α_1 is an immediate consequence of Lemma 3.

LEMMA 9. *Let $\lambda^2 \in \Gamma(\epsilon_{\rm p})$. The functions $\mu_i^{\pm}(\lambda, \alpha_1)$ are continuous functions of $\alpha_1 \in [0, 2\pi]$ for each $i = 1, 2, \ldots$.*

5. Defect in the strip. To the strip in Figure 1 we introduce a defect centered at the origin. Let ϵ and $\epsilon_{\rm p}$ differ only in the region Ω which is contained in $(-\frac{1}{2}, \frac{1}{2}) \times (-h, h)$, as indicated in Figure 3. Again, we view α_1 as a parameter and consider the eigenvalue problem of finding $u_{\alpha_1} \in H^1_{\alpha_1}(O)$ satisfying

$$\Delta u_{\alpha_1} + \omega_{\alpha_1}^2 \epsilon(x) u_{\alpha_1} = 0 \quad \text{in } O. \tag{10}$$

Note that we have implicitly required u_{α_1} to satisfy the Bloch condition in the x_1-direction by putting it in $H^1_{\alpha_1}(O)$. We cannot apply the Bloch condition in the x_2-direction since the problem is no longer periodic in that direction. We will consider the spectrum $\sigma_{\alpha_1}(\epsilon)$ for a fixed α_1. In particular, we will look at defect modes introduced by the defect.

5.1. Stability of the essential spectrum. We define unbounded mappings

$$B^p_{\alpha_1} : L^2_{\alpha_1}(O, \epsilon_{\rm p}(x)dx) \to L^2_{\alpha_1}(O, \epsilon_{\rm p}(x)dx), \quad B^p_{\alpha_1} = -\Delta,$$
$$B_{\alpha_1} : L^2_{\alpha_1}(O, \epsilon(x)dx) \to L^2_{\alpha_1}(O, \epsilon(x)dx), \quad B_{\alpha_1} = -\Delta.$$

Denote by $\sigma_{\alpha_1}(\epsilon_{\rm p})$ and $\sigma_{\alpha_1}(\epsilon)$ the essential spectra of the operators $B^p_{\alpha_1}$ and B_{α_1}, respectively.

Instead of working directly with $B^p_{\alpha_1}$ and B_{α_1}, we will work with

$$A^p_{\alpha_1} = (-\Delta)^{-\frac{1}{2}} \epsilon_{\rm p} (-\Delta)^{-\frac{1}{2}} \text{ and } A_{\alpha_1} = (-\Delta)^{-\frac{1}{2}} \epsilon (-\Delta)^{-\frac{1}{2}}.$$

It is easily seen that $\omega^2_{\alpha_1} \in \sigma_{\alpha_1}(\epsilon_{\rm p})$ if and only if $\frac{1}{\omega^2_{\alpha_1}} \in \sigma_{\rm ess}(A^p_{\alpha_1})$. Similarly, $\omega^2_{\alpha_1} \in \sigma_{\alpha_1}(\epsilon)$ if and only if $\frac{1}{\omega^2_{\alpha_1}} \in \sigma_{\rm ess}(A_{\alpha_1})$. Next, let

$$C = A^p_{\alpha_1} - A_{\alpha_1} = (-\Delta)^{-\frac{1}{2}} (\epsilon_{\rm p} - \epsilon)(-\Delta)^{-\frac{1}{2}}.$$

Note that by assumption $(\epsilon_{\rm p} - \epsilon)$ has compact support. Therefore, the operator C maps functions in $L^2_{\alpha_1}(O)$ into $\mathcal{H}_{\alpha_1,\gamma}(O)$. By the compactness of the embedding $\mathcal{H}_{\alpha_1,\gamma}(O) \hookrightarrow L^2_{\alpha_1}(O)$, the perturbation C is a relatively compact perturbation of $A^p_{\alpha_1}$. Hence, as a consequence of Weyl's theorem [12], it follows that $\sigma_{\rm ess}(A^p_{\alpha_1}) = \sigma_{\rm ess}(A_{\alpha_1})$, and so the following lemma holds.

LEMMA 10. *For a fixed α_1, and perturbations*

$$q = \epsilon_{\rm p} - \epsilon$$

supported in $\Omega \subset\subset (-\frac{1}{2}, \frac{1}{2}) \times (-h, h)$, the essential spectrum $\sigma_{\alpha_1}(\epsilon_{\rm p})$ is stable; i.e.,

$$\sigma_{\alpha_1}(\epsilon_{\rm p}) = \sigma_{\alpha_1}(\epsilon).$$

5.2. Spectrum of the strip. From Lemma 10 it is readily seen that the spectrum of the strip (for a fixed $\alpha_1 \in [0, 2\pi]$) in the gap $\Gamma(\epsilon_{\rm p})$ consists only of isolated eigenvalues of finite multiplicity, which can accumulate only at the edges of the gap $\Gamma(\epsilon_{\rm p})$. Let us consider the eigenvalue problem (10), where the eigenvalue $\omega^2_{\alpha_1} \in \Gamma(\epsilon_{\rm p})$.

The eigenvector u_{α_1} satisfies

$$\Delta u_{\alpha_1} + \omega^2_{\alpha_1} \epsilon_{\rm p}(x) u_{\alpha_1} = \omega^2_{\alpha_1} (\epsilon_{\rm p}(x) - \epsilon(x)) \chi_\Omega u_{\alpha_1} \quad \text{in } O.$$

Since $\omega^2_{\alpha_1} \in \Gamma(\epsilon_{\rm p})$, u_{α_1} solves the Lippman–Schwinger integral equation

$$u_{\alpha_1}(x) = \omega^2_{\alpha_1} \int_\Omega g_{\alpha_1}(x, y; \omega_{\alpha_1}) q(y) u_{\alpha_1}(y)\, dy,$$

where $q(y) = \epsilon_{\rm p}(y) - \epsilon(y)$ is supported in Ω. This implies that

$$\operatorname{sgn}(q) \frac{1}{\omega^2_{\alpha_1}} \sqrt{|q|} \chi_\Omega u_{\alpha_1} = \mathcal{A}_{\alpha_1}(\omega_{\alpha_1})(\sqrt{|q|} \chi_\Omega u_{\alpha_1}).$$

Conversely, if $\varphi(\omega_{\alpha_1}, \alpha_1)$ is an eigenvector of $\mathcal{A}_{\alpha_1}(\omega_{\alpha_1})$ with eigenvalue $\operatorname{sgn}(q)\frac{1}{\omega^2_{\alpha_1}}$, then $\int_\Omega g_{\alpha_1}(x, y; \omega_{\alpha_1}) \varphi(\omega_{\alpha_1}, \alpha_1)(y)\, dy$ is an eigenvector of (10) with eigenvalue $\omega^2_{\alpha_1} \in \Gamma(\epsilon_{\rm p})$. As a consequence of the results stated in the above section for $\mathcal{A}_{\alpha_1}(\lambda)$, where $\lambda^2 \in \Gamma(\epsilon_{\rm p})$, the following results on the spectrum of the strip hold.

THEOREM 1. *For a fixed $\alpha_1 \in [0, 2\pi]$, and perturbations $q = \epsilon_{\rm p} - \epsilon$ supported in $\Omega \subset\subset (-\frac{1}{2}, \frac{1}{2}) \times (-h, h)$ we have*

- *if $q \geq 0$, the eigenvalues ω_{α_1} of (10) in the gap $\Gamma(\epsilon_{\rm p})$ coincide with the set of the solutions of the equations*

$$\mu^+_i(\lambda, \alpha_1) = \frac{1}{\lambda^2}, \qquad i = 1, 2, \dots,$$

  where $\mu^+_i(\lambda, \alpha_1)$ are the positive eigenvalues of the operator $\mathcal{A}_{\alpha_1}(\lambda)$; moreover, if $\varphi(\lambda, \alpha_1)$ is an eigenvector of the operator $\mathcal{A}_{\alpha_1}(\lambda)$ with eigenvalue $\mu^+_i(\lambda, \alpha_1) = \frac{1}{\lambda^2}$, then $\int_\Omega g_{\alpha_1}(x, y; \lambda) \varphi(\lambda, \alpha_1)(y)\, dy$ is an exponentially localized in x_2-direction eigenvalue of (10).
- *if $q \leq 0$, the eigenvalues ω_{α_1} of (10) in the gap $\Gamma(\epsilon_{\rm p})$ coincide with the set of the solutions of the equations*

$$\mu^-_i(\lambda, \alpha_1) = -\frac{1}{\lambda^2}, \qquad i = 1, 2, \dots,$$

  where $\mu^-_i(\lambda, \alpha_1)$ are the negative eigenvalues of the operator $\mathcal{A}_{\alpha_1}(\lambda)$. Moreover, if $\varphi(\lambda, \alpha_1)$ is an eigenvector of the operator $\mathcal{A}_{\alpha_1}(\lambda)$ with eigenvalue $\mu^-_i(\lambda, \alpha_1) = -\frac{1}{\lambda^2}$, then $\int_\Omega g_{\alpha_1}(x, y; \lambda) \varphi(\lambda, \alpha_1)(y)\, dy$ is an exponentially localized in x_2-direction eigenvalue of (10).

The following result is an immediate consequence of Theorem 1. It gives a criterion for the absence of eigenvalues of (10).

COROLLARY 1. *Problem* (10) *has no eigenvalues in the gap* $\Gamma(\epsilon_{\mathrm{p}})$ *for small* $||\epsilon_{\mathrm{p}} - \epsilon||_{L^\infty(\Omega)}$.

5.3. Filling in the gap. The result of Theorem 1 provides a way by which we can interpret the guided mode spectrum. However, it does not tell us about the general properties, such as absolute continuity, of the spectrum of the wave operator in a medium with a line defect. What we can say is that it is possible for the bandgap of the periodic medium, $\Gamma(\epsilon_{\mathrm{p}})$, to be filled in (partially or totally) by the introduction of the line defect. To see this, assume for now that $q > 0$, and let $\alpha_1^0 \in [0, 2\pi]$ be such that there exists a solution to

$$\mu_i^+(\lambda, \alpha_1^0) = \frac{1}{\lambda^2}$$

for some i, which we denote by $\omega^2_{\alpha_1^0,i}$. We argue by Lemma 9 that for α_1 in an open neighborhood of α_1^0, there exists a solution λ to

$$\mu_i^+(\lambda, \alpha_1) = \frac{1}{\lambda^2},$$

and this solution is a continuous function of α_1. Therefore, as α_1 is varied over the neighborhood of α_0, we trace out the solution set $\omega_{\alpha_1,i}$. The values $\omega^2_{\alpha_1,i}$ fill out part (or all) of the gap $\Gamma(\epsilon_{\mathrm{p}})$.

Let us draw a conceptual figure of the fill-in process. In Figure 4, the horizontal axis is α_1 and the vertical axis is frequency. For the periodic medium ϵ_{p}, we calculate the passband for each value of α_1. The passbands are intervals in frequency, which is

$$\bigcup_n i_{\alpha_1,n},$$

where $i_{\alpha_1,n}$ is as in section 4.1. To display the spectrum of the wave operator in the periodic medium, we draw these intervals vertically on the ω-α_1 plane at horizontal position α_1. When we are done, we are left with two (or more) disjoint regions, such as those shown in Figure 4. The dark regions correspond to values of α_1 and frequencies at which waves can propagate.

We have indicated the point α_0 on the figure and suppose there is a corresponding eigenvalue $\omega_{\alpha_1^0}$ of (10). As we vary α_1 in the neighborhood of α_1^0, we obtain solutions ω_{α_1}, which when displayed in the figure trace out a curve that fills (or partially fills) the gap. This is indicated in the figure. We denote the interval of α_1 over which such eigenvalues exist by A_1. The results we obtained so far conclude only that A_1 is nonempty. Detailed information, such as connectivity of A_1, is unknown. The picture gets a little more complicated if there are multiple solutions at a given α_1^0, producing multiple curves. It is hoped that the numerical examples, to be presented in the last section, will give some insight into this and other questions.

6. Guided propagation. Our medium, with a line defect lying along the x_1-axis, can guide waves so that they propagate along the x_1-direction. Such a wave will be a linear combination of the guided modes. Each guided mode is an eigenfunction of (10) and decays exponentially away from the origin in x_2. Let us suppose that we have found guided modes for α_1 in A_1 and the corresponding frequencies for these

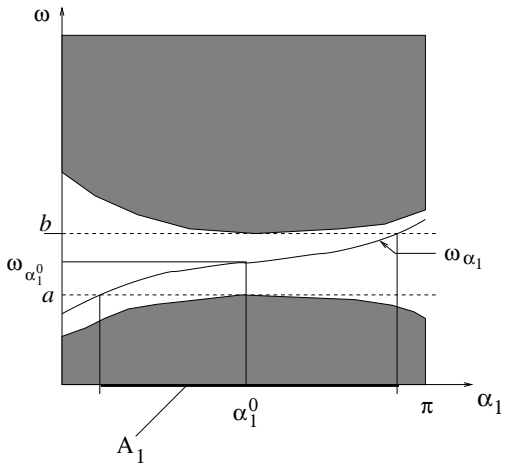


FIG. 4. *A conceptual figure depicting the fill-in process of the bandgap. The vertical axis is frequency. The dark regions correspond to values of* α_1 *and frequencies at which waves can propagate in the periodic medium. The gap above corresponds to the bandgap of the periodic medium. When a defect is introduced, a point spectrum is created for some values of* α_1*. These points trace out a curve in the gap. In the example, a complete fill-in is indicated.*

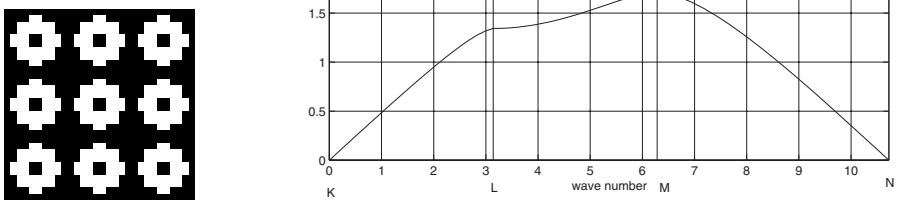


FIG. 5. *The periodic medium and its spectrum. Light region corresponds to* $\epsilon = 9$ *and dark region to* $\epsilon = 1$*. Note the presence of the gaps around* 1.7 *and* 2.7*.*

modes are ω_{α_1}. We denote the associated eigenfunctions by $u_{\alpha_1}(x)$. Then a guided wave in this medium has the representation

$$U(x,t) = \int_{\alpha_1 \in A_1} C_{\alpha_1} u_{\alpha_1}(x) e^{i\omega_{\alpha_1} t} d\alpha_1,$$

where C_{α_1} is the mode amplitude function. Such a wave would propagate like a pulse along the x_1-axis. Its dispersion relation along the x_1-direction is given by ω_{α_1}, with group velocity given by $d\omega_{\alpha_1}/d\alpha_1$.

7. Numerical experiment. We consider a periodic structure with permittivity ϵ_{p} taking a value of 1 or 9. The medium is shown in Figure 5 (left), where the dark region corresponds to low index. We use finite difference approximations to solve (3) for values of α on the boundary of the Brioullin zone:

$$KL = \{0 \le \alpha_1 \le \pi, \alpha_2 = 0\}, \quad LM = \{0 \le \alpha_2 \le \pi, \alpha_1 = \pi\},$$
$$MN = \{0 \le \alpha_1 \le \pi, \alpha_2 = \alpha_1\}.$$

The resulting values of $\omega_n(\alpha)$ are displayed in Figure 5 (right). Note the presence of a small gap near 1.7 and a larger one near 2.7.

The problem involving a line defect is considered next. We start by solving the strip eigenvalue problem (5) for the periodic medium ϵ_{p}. We do this by creating a strip of 19 cells and applying the Bloch condition with parameter α_2 at the top and at the bottom of the strip. The Bloch condition has to be modified to reflect the strip geometry:

$$u_{\alpha_1}(x_1, x_2 + 19) = u_{\alpha_1}(x_1, x_2)e^{i19\alpha_2}.$$

We choose α_1 and solve for eigenvalues $\nu_{\alpha_1,n}(\alpha_2)$ as α_2 is sampled over the interval $[0, \pi/19]$. The first 25 eigenvalues are found for each case and are displayed as points with horizontal coordinate α_1. The resulting picture is shown in Figure 6 (top). The points lie in the dark regions indicated in the conceptual version shown in Figure 4. Note that the gap in question is the one around 2.7 from Figure 5.

Next, a defect is introduced. The strip with the defect can be seen in Figure 6 (bottom). The calculation described above is repeated with the defect. With the Bloch condition at the top and bottom, we are essentially performing a supercell computation. The defect produces the fill-in described in section 5, which we display in Figure 6 (bottom). For each α_1 the eigenvalues calculated for different samples of α_2 are slightly different. The dense placement of points vertically has now become more broken. What can be clearly seen is that the boundary between the empty region and the region populated with dots remains in place. Moreover, we now see the presence of an additional set of points running from the lower left to the upper right. Of these points, those that lie within the gap correspond to the guided modes and trace out the dispersion curve of the modes. To verify that this is the case, we selected the points that lie in the bandgap of the periodic medium. Note that because we are using the supercell method, each defect spectrum is calculated as many times as we sample α_2. For each of these points, we calculate the associated eigenfunction $u_{\alpha_1}(x, \alpha_2)$ and display its absolute value. In Figure 7, we show the absolute value of the guided modes for different α_1. Note that the modes become more concentrated near the defect for higher values of α_1. In this example, the set of α_1 for which we have guided modes is smaller than the interval $[-\pi, \pi]$.

We ran two more examples with different defects. In the first case, shown in Figure 8 (top), the defect produces two defect modes in the gap at most values of α_1. The fill-in is total. In the second case, Figure 8, bottom, the fill-in is only partial. A bandgap, slightly smaller, still exists.

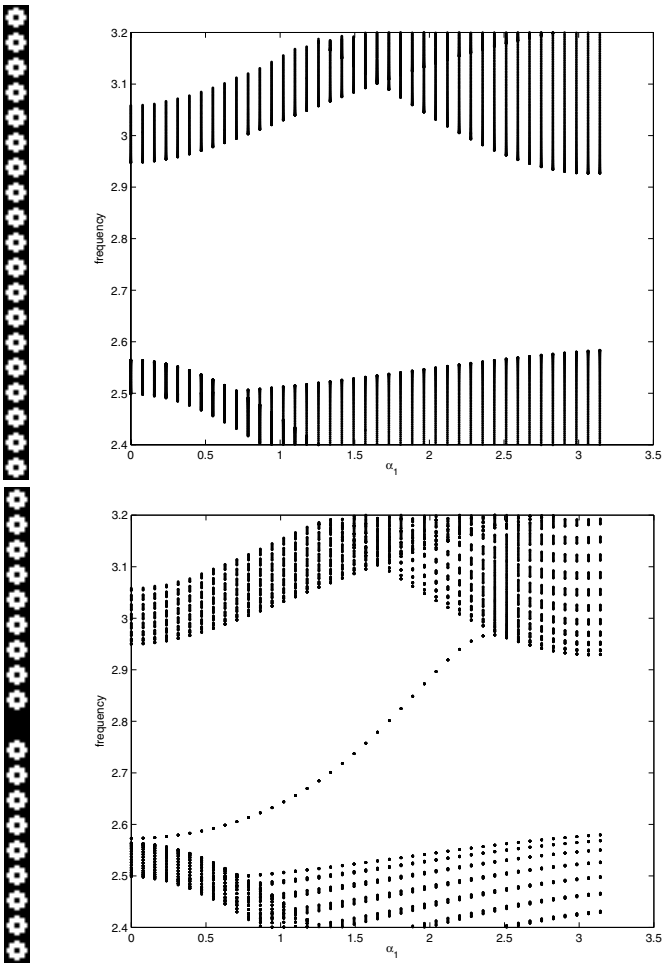


FIG. 6. *The spectrum of the periodic strip and the strip with a defect. The images of the strips are given on the left. The graph for the periodic medium is calculated for comparison with that of the medium with defect. On the top, note the white region corresponding to the gap. On the bottom, note that the boundary of the white region is not changed for the perturbed medium. The only difference is the creation of a point spectrum (a curve) which fills the gap.*

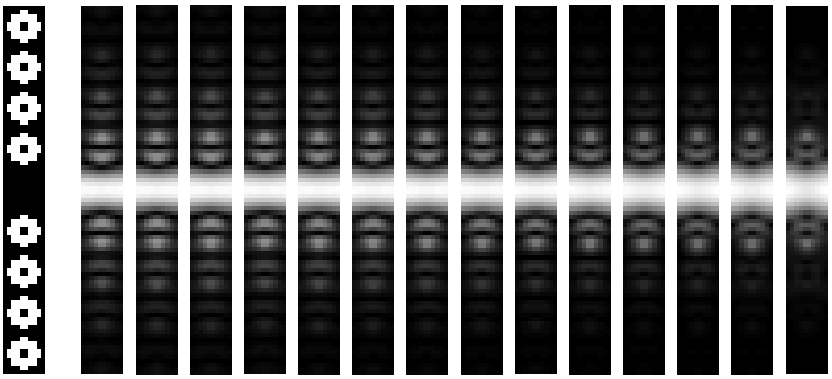


FIG. 7. *The medium and the guided modes intensity sampled at various values of* α_1.

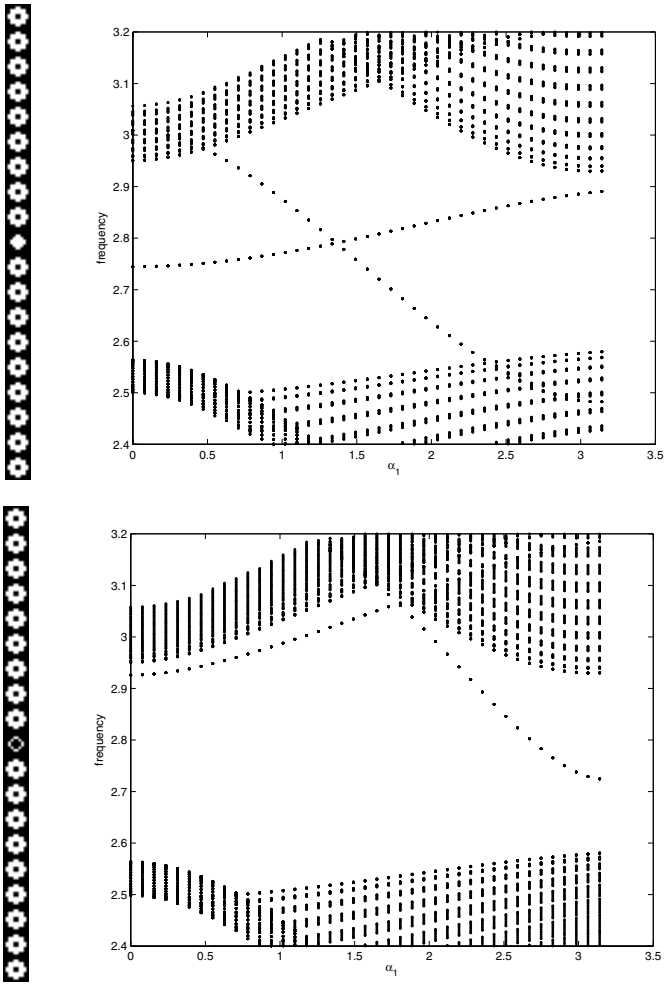


FIG. 8. *Spectra of the strip with two different defects. Images of the strip are given on the left. Note that a complete fill-in has occurred in the first case. Moreover, there are multiple guided modes at each* α_1. *In the second case, the fill-in is partial.*

8. Discussion. We have provided a mathematical framework by which wave guidance phenomena in a photonic bandgap structure with line defect can be understood. The key idea is to analyze the spectral properties of Helmholtz's equation in a strip. Using Weyl's theorem, we have shown that the essential spectrum of the periodic medium is stable to the introduction of the defect. The new spectrum created by the defect corresponds to guided modes and has a dispersion relation that can fill (or partially fill) the bandgap of the periodic medium. The main findings of this work are further illustrated in numerical examples.

The question of absolute continuity of the spectrum, while of great importance, is beyond the scope and techniques of the present work. Our theoretical result shows that if a line defect creates a guided mode, there should be a continuum of these modes parameterized by wave number α_1 over an open interval, covering an open interval of the bandgap in frequency. The numerical results show that for some perturbations, the coverage is complete, thus filling in the bandgap, while for other perturbations,

the fill-in is only partial. Our method also does not reveal what happens near the band edges. Numerical calculations presented here indicate that guided modes with frequencies up to the band edges are possible. It is not known how these modes decay in the x_2-direction.

Finally, we remark that it would be useful to take this work and produce a kind of approximate theory that is as simple to use as that employed in modeling guided waves in optical fibers. This and other issues mentioned indicate possibilities for further investigation.

Acknowledgments. This collaboration was initiated during the first author's visit to the Institute for Mathematics and Its Applications in July 2001. The authors express their gratitude to the referees who carefully read an earlier version of this paper. Their valuable suggestions have been incorporated into the present version.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] H. Ammari, N. Béreux, and E. Bonnetier, *Analysis of the radiation properties of a planar antenna on a photonic crystal substrate*, Math. Methods Appl. Sci., 24 (2001), pp. 1021–1042.
[3] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North–Holland, New York, 1978.
[4] I. El-Kady, M. Sigalas, R. Biswas, and K.-M. Ho, *Waveguides in two-dimensional photonic bandgap materials*, J. Lightwave Techn., 17 (1999), p. 2042.
[5] A. Figotin and A. Klein, *Localization of light in lossless inhomogeneous dielectrics*, J. Opt. Soc. Amer. A, 15 (1998), pp. 1423–1435.
[6] A. Figotin and A. Klein, *Localized classical waves created by defects*, J. Statist. Phys., 86 (1997), pp. 165–177.
[7] A. Figotin and A. Klein, *Midgap defect modes in dielectric and acoustic media*, SIAM J. Appl. Math., 58 (1998), pp. 1748–1773.
[8] A. Figotin and P. Kuchment, *Band-gap structure of spectra of periodic dielectric and acoustic media.* I: *Scalar model*, SIAM J. Appl. Math., 56 (1996), pp. 68–88.
[9] A. Figotin and P. Kuchment, *Band-gap structure of spectra of periodic dielectric and acoustic media.* II: *Two-dimensional photonic crystals*, SIAM J. Appl. Math., 56 (1996), pp. 1561–1620.
[10] P. Kuchment, *The mathematics of photonic crystals*, in Mathematical Modelling in Optical Science, G. Bao, L. Cowsar, and W. Masters, eds., Frontiers Appl. Math. 22, SIAM, Philadelphia, 2001, pp. 207–272.
[11] F. Odeh and J. B. Keller, *Partial differential equations with periodic coefficients and Bloch waves in crystals*, J. Math. Phys., 5 (1964), pp. 1499–1504.
[12] M. Reed and B. Simon, *Methods of Modern Physics, Vol.* IV*: Analysis of Operators*, Academic Press, New York, 1978.
[13] K. Sakoda, *Optical Properties of Photonic Crystals*, Springer-Verlag, Berlin, 2001.
[14] M. Sigalas, R. Biswas, K.-M. Ho, C. Soukoulis, and D. Crouch, *Waveguides in* 3*-D metallic photonic band gap materials*, Phys. Rev. B, 60 (1999), p. 4426–4429.
[15] C. Wilcox, *Theory of Bloch waves*, J. Anal. Math., 33 (1978), pp. 146–167.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 2034–2049

CLOSED-FORM SOLUTIONS FOR PERPETUAL AMERICAN PUT OPTIONS WITH REGIME SWITCHING*

X. GUO† AND Q. ZHANG‡

Abstract. This paper studies an optimal stopping time problem for pricing perpetual American put options in a regime switching model. An explicit optimal stopping rule and the corresponding value function in a closed form are obtained using the "modified smooth fit" technique. The solution is then compared with the numerical results obtained via a dynamic programming approach and also with a two-point boundary-value differential equation (TPBVDE) method.

Key words. Markov chain, optimal stopping time, American options, regime switching, modified smooth fit principle

AMS subject classifications. 90A09, 60J27

DOI. 10.1137/S0036139903426083

1. Introduction.

Given a probability space $(\Omega, \mathcal{F}, P)$, consider a process $X(t)$ which satisfies (in a strong sense) a stochastic differential equation of the following form:

$$dX(t) = X(t)\mu_{\epsilon(t)}dt + X(t)\sigma_{\epsilon(t)}dW(t), \qquad X(0) = x, \tag{1}$$

where $\epsilon(t) \in \{1, \ldots, S\}$ is a finite-state continuous-time Markov chain and $W(t)$ is a standard Wiener process. Here $\epsilon(t)$ and $W(t)$ are defined on $(\Omega, \mathcal{F}, P)$ and are independent. Moreover, for a given $\epsilon(t) = i$, μ_i and σ_i $(i = 1, \ldots, S)$ are constants and known.

The $X(t)$ governed by (1) is generally referred to as a process with "regime switching (or shifts)" or "a Markov modulated (geometric) Brownian motion." There is a substantial body of literature on this type of model studied from different perspectives. See, for instance, Di Masi, Kabanov, and Runggaldier [3] for mean variance hedging issues; Guo [5, 7] for closed-form solutions for pricing European and perpetual lookback options; Yao, Zhang, and Zhou [23] for numerical algorithms for computing European stock options; Zhang [24] for suboptimal selling rules for investors; and Zhang and Yin [25] for portfolio optimization problems.

In light of the celebrated Black–Scholes geometric Brownian motion model (see Black and Scholes [1] and Samuelson [20]), which corresponds to a special case of (1) with $\mu_1 = \cdots = \mu_S$ and $\sigma_1 = \cdots = \sigma_S$, the primary motivation for the incorporation of the Markov chain $\epsilon(t)$ is the conviction that various economic factors (e.g., interest rates, quarterly GDP) and general information (e.g., corporate news releases, quarterly earnings reports) could be major catalysts for stock fluctuations. In addition, a finite-state Markov chain has been proved to be simple yet rich enough to characterize the uncertainty in many discrete events. These convictions have been further substantiated by numerical studies: Yao, Zhang, and Zhou [23] showed that the infamous "volatility smile" can be created with a Markov chain of a single jump, instead of the more complicated stochastic volatility model by Renault and Touzi [17].

*Received by the editors April 13, 2003; accepted for publication (in revised form) February 10, 2004; published electronically August 19, 2004.
http://www.siam.org/journals/siap/64-6/42608.html
†School of ORIE, Cornell University, Ithaca, NY 14853 (xinguo@orie.cornell.edu).
‡Department of Mathematics, University of Georgia, Athens, GA 30602 (qingz@math.uga.edu).

Our results. In this paper we consider an optimal stopping problem that arises in pricing American put options, in the framework of this regime switching model. An American option is a derivative that gives its holder the option but not the obligation of exercising a share of stock at his/her choice of time τ ($T \geq \tau \geq 0$), with a payoff of $(K - X_\tau)^+ = \max(0, K - X_\tau)$. Here, T is the *expiration date* and K is the *strike price*. It is well known that under a risk-neutral measure, the value (or the price) of this option is the expected discounted value of its future cash flow. (For more details, readers are referred to Duffie [4] and the references therein for risk-neutral option pricing for general models, to Guo [6] for the regime switching models, and to Karatzas [10] for the mathematical formulation of the American option pricing problem in the context of optimal stopping problems.) In particular, when $T = \infty$, the option becomes perpetual, and our optimal stopping problem becomes the evaluation of

$$V^*(x,i) = \sup_{0\leq\tau\leq\infty} E[e^{-r\tau}(K - X(\tau))^+ \mid X(0) = x, \epsilon(0) = i]. \tag{2}$$

Here, $r > 0$ is the discounted factor, and τ is an $\mathcal{F}_t = \sigma\{(W(s), \epsilon(s)) \mid s \leq t\}$-stopping time.

We derive an optimal stopping rule for (2) and its corresponding value functions for $S = 2$ (see Remark 3.5). We show that the optimal stopping times are of threshold type, with the technique of modified smooth fit. The main ingredient of the optimality proof is Dynkin's formula.

It is worth mentioning that a special case of this problem with no switching (i.e., $\mu_1 = \mu_2, \sigma_1 = \sigma_2$) was solved by McKean [14], and it is referred to in what follows as "the McKean problem." His result is the earliest instance in which optimal stopping problems were related to option pricings. See also Jacka [9] and Robbins, Sigmund, and Chow [19] for related literature on optimal stopping.

Organization. In section 2, we provide a detailed derivation of the closed-form solution to (2). The optimality proof is given in section 3. In section 4, we numerically compare the closed-form solution with numerical results derived from other previous approaches, namely the dynamic programming approach (see Guo [7]) and the TPB-VDE (two-point boundary-value differential equation) method (see Zhang [24]). The paper concludes with additional discussion and open problems in section 5.

2. The derivation of solutions. Given (1), we will study problem (2) with a two-state Markov chain (see Remark 3.5) for the general case K. Without loss of generality, we assume that $\sigma_1 \neq \sigma_2$ (see Remark 3.1) and that the Markov chain has a generator of the form

$$\begin{pmatrix} -\lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2 \end{pmatrix}, \tag{3}$$

with $\lambda_1, \lambda_2 > 0$.

Recall that when there is no regime switching, this problem corresponds to a McKean problem [14] for which there exists a threshold x^* such that the optimal stopping rule is $\tau^* = \inf\{t > 0 : X(t) \notin (x^*, \infty)\}$, and the corresponding value function

$$\begin{aligned} V^*(x) &= \sup_{0\leq\tau\leq\infty} E[e^{-r\tau}(K - X(\tau))^+ | X(0) = x] \\ &= E[e^{-r\tau^*}(K - X(\tau^*))^+ | X(0) = x] \end{aligned}$$

is given by

$$V^*(x) = \begin{cases} (K - x^*)(x/x^*)^\gamma & \text{if } x > x^*, \\ K - x & \text{if } x \le x^*. \end{cases}$$

Now, with a two-state Markov chain and with $\sigma_1 \neq \sigma_2$, it is easy to see that $(X(t), \epsilon(t))$ is a joint Markov process (see Guo [7]). Therefore, it is natural to conjecture that the optimal stopping rule is also of threshold type, except that the threshold should vary depending on the state $\epsilon(t)$. In other words, we expect the existence of two thresholds $x_1, x_2 \le K$, so that the optimal stopping rule is given as

$$\tau^* = \inf\{t \ge 0 \ \mid \ (X(t), \epsilon(t)) \notin D\},$$

where

$$D = \{(x, i) \ \mid \ V^*(x, i) > (K - x)^+\}.$$

The set D is referred to as the *continuation region*. Using τ^*, the corresponding value functions are

$$V^*(x, i) = E[e^{-r\tau^*}(K - X(\tau^*))^+ \ \mid \ X(0) = x, \epsilon(0) = i]. \tag{4}$$

We consider the case when D can be represented by two threshold levels x_1 and x_2, i.e.,

$$D = \{(x, 1) \ \mid \ x \in (x_1, \infty)\} \cup \{(x, 2) \ \mid \ x \in (x_2, \infty)\}.$$

Notice that x_1 and x_2 should depend on r, K, $\mu_i, \sigma_i, \lambda_i$. For any x_1 and x_2, there are only three possibilities, $x_1 < x_2$, $x_1 > x_2$, and $x_1 = x_2$. In the next sections we discuss each of these cases and derive the values of these thresholds x_i as well as the corresponding value functions (denoted as $V_i(x)$) obtained from exercising this type of stopping rule. We will then prove the optimality of these value functions, i.e., $V^*(x, i) = V_i(x)$, in Theorem 3.1.

2.1. Case 1: $x_1 < x_2 \le K$. At any given time t, if $\epsilon(t) = 1$ and $X(t) \le x_1$, then one should stop immediately and obtain a payoff of $(K - X(t))^+$; this follows from the definition of x_1 and x_2. However, if $X(t) \le x_1$ with $\epsilon(t) = 2$, it is not optimal to stop until $X(t) \le x_2$. In view of Ito's differential rule, this is translated into a set of differential equations. For $x \in [x_1, x_2]$, we have

$$\begin{cases} (r + \lambda_1)V_1(x) &= x\mu_1 V_1'(x) + \dfrac{1}{2}x^2\sigma_1^2 V_1''(x) + \lambda_1(K - x), \\ V_2(x) &= K - x; \end{cases} \tag{5}$$

for $x \in [x_2, \infty)$,

$$\begin{cases} (r + \lambda_1)V_1(x) &= x\mu_1 V_1'(x) + \dfrac{1}{2}x^2\sigma_1^2 V_1''(x) + \lambda_1 V_2(x), \\ (r + \lambda_2)V_2(x) &= x\mu_2 V_2'(x) + \dfrac{1}{2}x^2\sigma_2^2 V_2''(x) + \lambda_2 V_1(x); \end{cases} \tag{6}$$

and for $x \in [0, x_1]$,

$$V_1(x) = V_2(x) = K - x. \tag{7}$$

Now, (6) has an associated characteristic function

$$g_1(\beta)g_2(\beta) = \lambda_1\lambda_2, \tag{8}$$

where

$$g_1(\beta) = \lambda_1 + r - \left(\mu_1 - \frac{1}{2}\sigma_1^2\right)\beta - \frac{1}{2}\sigma_1^2\beta^2,$$
$$g_2(\beta) = \lambda_2 + r - \left(\mu_2 - \frac{1}{2}\sigma_2^2\right)\beta - \frac{1}{2}\sigma_2^2\beta^2.$$

Moreover, this characteristic function has four distinct roots $\beta_1 < \beta_2 < 0 < \beta_3 < \beta_4$ (see Guo [7]), such that the general form of the solution to (6) is given by

$$\begin{aligned} V_1(x) &= \sum_{i=1}^{4} A_i x^{\beta_i}, \\ V_2(x) &= \sum_{i=1}^{4} B_i x^{\beta_i}, \end{aligned}$$

with $B_i = l_i A_i$ and $l_i = l(\beta_i) = \frac{g_1(\beta_i)}{\lambda_1} = \frac{\lambda_2}{g_2(\beta_i)}$.

Note that when $x \to \infty$, $V_1(x)$ and $V_2(x)$ are bounded. Thus, the positive powers of x should be eliminated so that

$$\begin{aligned} V_1(x) &= A_1 x^{\beta_1} + A_2 x^{\beta_2}, \\ V_2(x) &= B_1 x^{\beta_1} + B_2 x^{\beta_2}. \end{aligned} \tag{9}$$

Next, we turn our attention to (5). The first equation is an inhomogeneous equation whose solution can be written as

$$V_1(x) = C_1 x^{\gamma_1} + C_2 x^{\gamma_2} + \phi(x), \tag{10}$$

where $\phi(x)$ is a special solution and γ_1, γ_2 are the two real roots of

$$\mu_1\gamma + \frac{1}{2}\sigma_1^2\gamma(\gamma - 1) = r + \lambda_1.$$

In particular, when $r + \lambda_1 - \mu_1 \neq 0$, one can choose

$$\phi(x) = \frac{\lambda_1 K}{r + \lambda_1} - \frac{\lambda_1 x}{r + \lambda_1 - \mu_1}. \tag{11}$$

Now, we want to solve for A_1, A_2, C_1, C_2, x_1, and x_2. To this end, appropriate boundary conditions are needed. Applying the smooth fit at x_2, conditions $V_2(x+) = V_2(x-)$ and $V_2'(x+) = V_2'(x-)$ suggest

$$\begin{cases} l_1 A_1 x_2^{\beta_1} + l_2 A_2 x_2^{\beta_2} &= K - x_2, \\ \beta_1 l_1 A_1 x_2^{\beta_1} + \beta_2 l_2 A_2 x_2^{\beta_2} &= -x_2. \end{cases} \tag{12}$$

Similarly, the smoothness of $V_1(x)$ at x_1 and x_2 yields

$$\begin{cases} A_1 x_2^{\beta_1} + A_2 x_2^{\beta_2} &= C_1 x_2^{\gamma_1} + C_2 x_2^{\gamma_2} + \phi(x_2), \\ \beta_1 A_1 x_2^{\beta_1} + \beta_2 A_2 x_2^{\beta_2} &= \gamma_1 C_1 x_2^{\gamma_1} + \gamma_2 C_2 x_2^{\gamma_2} + x_2 \phi'(x_2), \end{cases} \tag{13}$$

and

$$\begin{cases} C_1 x_1^{\gamma_1} + C_2 x_1^{\gamma_2} + \phi(x_1) &= K - x_1, \\ \gamma_1 C_1 x_1^{\gamma_1} + \gamma_2 C_2 x_1^{\gamma_2} + x_1 \phi'(x_1) &= -x_1. \end{cases} \tag{14}$$

Combining the above three equations and following some algebraic manipulation, we obtain an algebraic equation for x_1 and x_2:

$$\begin{pmatrix} x_1^{-\gamma_1} & 0 \\ 0 & x_1^{-\gamma_2} \end{pmatrix} F_1(x_1) = \begin{pmatrix} x_2^{-\gamma_1} & 0 \\ 0 & x_2^{-\gamma_2} \end{pmatrix} F_2(x_2), \tag{15}$$

where

$$F_1(x_1) = \begin{pmatrix} 1 & 1 \\ \gamma_1 & \gamma_2 \end{pmatrix}^{-1} \begin{pmatrix} K - x_1 - \phi(x_1) \\ -x_1 - x_1 \phi'(x_1) \end{pmatrix}$$

and

$$F_2(x_2) = \begin{pmatrix} 1 & 1 \\ \gamma_1 & \gamma_2 \end{pmatrix}^{-1} \left[\begin{pmatrix} 1 & 1 \\ \beta_1 & \beta_2 \end{pmatrix} \begin{pmatrix} l_1 & l_2 \\ \beta_1 l_1 & \beta_2 l_2 \end{pmatrix}^{-1} \begin{pmatrix} K - x_2 \\ -x_2 \end{pmatrix} - \begin{pmatrix} \phi(x_2) \\ x_2 \phi'(x_2) \end{pmatrix} \right].$$

In particular, if $r + \lambda_1 - \mu_1 \neq 0$, where $\phi(x_1)$ is in the form of (11), then

$$F_1(x_1) = a_1 + a_2 x_1$$

and

$$F_2(x_2) = b_1 + b_2 x_2.$$

Here

$$a_1 = \begin{pmatrix} 1 & 1 \\ \gamma_1 & \gamma_2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{rK}{r+\lambda_1} \\ 0 \end{pmatrix}, \qquad a_2 = \begin{pmatrix} 1 & 1 \\ \gamma_1 & \gamma_2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{\mu_1 - r}{r+\lambda_1-\mu_1} \\ \frac{\mu_1 - r}{r+\lambda_1-\mu_1} \end{pmatrix},$$

$$b_1 = \begin{pmatrix} 1 & 1 \\ \gamma_1 & \gamma_2 \end{pmatrix}^{-1} \left[\begin{pmatrix} 1 & 1 \\ \beta_1 & \beta_2 \end{pmatrix} \begin{pmatrix} l_1 & l_2 \\ \beta_1 l_1 & \beta_2 l_2 \end{pmatrix}^{-1} \begin{pmatrix} K \\ 0 \end{pmatrix} + \begin{pmatrix} -\frac{\lambda_1 K}{r+\lambda_1} \\ 0 \end{pmatrix} \right],$$

$$b_2 = \begin{pmatrix} 1 & 1 \\ \gamma_1 & \gamma_2 \end{pmatrix}^{-1} \left[- \begin{pmatrix} 1 & 1 \\ \beta_1 & \beta_2 \end{pmatrix} \begin{pmatrix} l_1 & l_2 \\ \beta_1 l_1 & \beta_2 l_2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \frac{\lambda_1}{r+\lambda_1-\mu_1} \\ \frac{\lambda_1}{r+\lambda_1-\mu_1} \end{pmatrix} \right].$$

The coefficients are given by

$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} l_1 x_2^{\beta_1} & l_2 x_2^{\beta_2} \\ \beta_1 l_1 x_2^{\beta_1} & \beta_2 l_2 x_2^{\beta_2} \end{pmatrix}^{-1} \begin{pmatrix} K - x_2 \\ -x_2 \end{pmatrix}, \qquad \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} l_1 A_1 \\ l_2 A_2 \end{pmatrix},$$

and

$$\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} x_1^{\gamma_1} & x_1^{\gamma_2} \\ \gamma_1 x_1^{\gamma_1} & \gamma_2 x_1^{\gamma_2} \end{pmatrix}^{-1} \begin{pmatrix} K - x_1 - \phi(x_1) \\ -x_1 - x_1 \phi'(x_1) \end{pmatrix}.$$

With these coefficients, the value functions become

$$V_1(x) = \begin{cases} A_1 x^{\beta_1} + A_2 x^{\beta_2} & \text{if } x > x_2, \\ C_1 x^{\gamma_1} + C_2 x^{\gamma_2} + \phi(x) & \text{if } x_1 < x \le x_2, \\ K - x & \text{if } x \le x_1, \end{cases}$$

$$V_2(x) = \begin{cases} B_1 x^{\beta_1} + B_2 x^{\beta_2} & \text{if } x > x_2, \\ K - x & \text{if } x \le x_2. \end{cases} \tag{16}$$

2.2. Case 2: $x_2 < x_1 \le K$. The derivation of this case is analogous to that of $x_1 < x_2$, and we only summarize the results below.

Let $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ be the roots of

$$\mu_2 \gamma + \frac{1}{2}\sigma_2^2 \gamma(\gamma - 1) = r + \lambda_2,$$

and $\widetilde{\phi}(x)$ be a solution to

$$(r + \lambda_2) V_2(x) = x\mu_2 V_2'(x) + \frac{1}{2} x^2 \sigma_2^2 V_2''(x) + \lambda_2 (K - x).$$

Then, x_1, x_2 satisfy

$$\begin{pmatrix} x_1^{-\widetilde{\gamma}_1} & 0 \\ 0 & x_1^{-\widetilde{\gamma}_2} \end{pmatrix} \widetilde{F}_1(x_1) = \begin{pmatrix} x_2^{-\widetilde{\gamma}_1} & 0 \\ 0 & x_2^{-\widetilde{\gamma}_2} \end{pmatrix} \widetilde{F}_2(x_2), \tag{17}$$

with

$$\widetilde{F}_1(x_1) = \begin{pmatrix} 1 & 1 \\ \widetilde{\gamma}_1 & \widetilde{\gamma}_2 \end{pmatrix}^{-1} \left[\begin{pmatrix} 1 & 1 \\ \beta_1 & \beta_2 \end{pmatrix} \begin{pmatrix} \widetilde{l}_1 & \widetilde{l}_2 \\ \beta_1 \widetilde{l}_1 & \beta_2 \widetilde{l}_2 \end{pmatrix}^{-1} \begin{pmatrix} K - x_1 \\ -x_1 \end{pmatrix} - \begin{pmatrix} \phi(x_1) \\ x_1 \phi'(x_1) \end{pmatrix} \right]$$

and

$$\widetilde{F}_2(x_2) = \begin{pmatrix} 1 & 1 \\ \widetilde{\gamma}_1 & \widetilde{\gamma}_2 \end{pmatrix}^{-1} \begin{pmatrix} K - x_2 - \phi(x_2) \\ -x_2 - x_2\phi'(x_2) \end{pmatrix},$$

where $\widetilde{l}_i = 1/l_i$.

In particular, if $r + \lambda_2 - \mu_2 \neq 0$, then $\widetilde{\phi}(x)$ is given by

$$\widetilde{\phi}(x) = \frac{\lambda_2 K}{r + \lambda_2} - \frac{\lambda_2 x}{r + \lambda_2 - \mu_2},$$

and

$$\widetilde{F}_1(x_1) = \widetilde{a}_1 + \widetilde{a}_2 x_1,$$

$$\widetilde{F}_2(x_2) = \widetilde{b}_1 + \widetilde{b}_2 x_2,$$

where

$$\widetilde{a}_1 = \begin{pmatrix} 1 & 1 \\ \widetilde{\gamma}_1 & \widetilde{\gamma}_2 \end{pmatrix}^{-1} \left[\begin{pmatrix} 1 & 1 \\ \beta_1 & \beta_2 \end{pmatrix} \begin{pmatrix} \widetilde{l}_1 & \widetilde{l}_2 \\ \beta_1\widetilde{l}_1 & \beta_2\widetilde{l}_2 \end{pmatrix}^{-1} \begin{pmatrix} K \\ 0 \end{pmatrix} + \begin{pmatrix} -\frac{\lambda_2 K}{r+\lambda_2} \\ 0 \end{pmatrix} \right],$$

$$\widetilde{a}_2 = \begin{pmatrix} 1 & 1 \\ \widetilde{\gamma}_1 & \widetilde{\gamma}_2 \end{pmatrix}^{-1} \left[- \begin{pmatrix} 1 & 1 \\ \beta_1 & \beta_2 \end{pmatrix} \begin{pmatrix} \widetilde{l}_1 & \widetilde{l}_2 \\ \beta_1\widetilde{l}_1 & \beta_2\widetilde{l}_2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \frac{\lambda_2}{r+\lambda_2-\mu_2} \\ \frac{\lambda_2}{r+\lambda_2-\mu_2} \end{pmatrix} \right],$$

$$\widetilde{b}_1 = \begin{pmatrix} 1 & 1 \\ \widetilde{\gamma}_1 & \widetilde{\gamma}_2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{rK}{r+\lambda_2} \\ 0 \end{pmatrix}, \qquad \widetilde{b}_2 = \begin{pmatrix} 1 & 1 \\ \widetilde{\gamma}_1 & \widetilde{\gamma}_2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{\mu_2 - r}{r+\lambda_2-\mu_2} \\ \frac{\mu_1 - r}{r+\lambda_2-\mu_2} \end{pmatrix}.$$

In short, if $x_1 > x_2$, then the corresponding value functions are

$$V_1(x) = \begin{cases} \widetilde{A}_1 x^{\beta_1} + \widetilde{A}_2 x^{\beta_2} & \text{if } x > x_1, \\ K - x & \text{if } x \leq x_1, \end{cases}$$

$$V_2(x) = \begin{cases} \widetilde{B}_1 x^{\beta_1} + \widetilde{B}_2 x^{\beta_2} & \text{if } x > x_1, \\ \widetilde{C}_1 x^{\widetilde{\gamma}_1} + \widetilde{C}_2 x^{\widetilde{\gamma}_2} + \widetilde{\phi}(x) & \text{if } x_2 < x \leq x_1, \\ K - x & \text{if } x \leq x_2, \end{cases} \tag{18}$$

with

$$\begin{pmatrix} \widetilde{A}_1 \\ \widetilde{A}_2 \end{pmatrix} = \begin{pmatrix} x_1^{\beta_1} & x_1^{\beta_2} \\ \beta_1 x_1^{\beta_1} & \beta_2 x_1^{\beta_2} \end{pmatrix}^{-1} \begin{pmatrix} K - x_1 \\ -x_1 \end{pmatrix}, \qquad \begin{pmatrix} \widetilde{B}_1 \\ \widetilde{B}_2 \end{pmatrix} = \begin{pmatrix} l_1\widetilde{A}_1 \\ l_2\widetilde{A}_2 \end{pmatrix},$$

and

$$\begin{pmatrix} \widetilde{C}_1 \\ \widetilde{C}_2 \end{pmatrix} = \begin{pmatrix} x_2^{\widetilde{\gamma}_1} & x_2^{\widetilde{\gamma}_2} \\ \widetilde{\gamma}_1 x_2^{\widetilde{\gamma}_1} & \widetilde{\gamma}_2 x_2^{\widetilde{\gamma}_2} \end{pmatrix}^{-1} \begin{pmatrix} K - x_2 - \widetilde{\phi}(x_2) \\ -x_2 - x_2\widetilde{\phi}'(x_2) \end{pmatrix}.$$

2.3. Case 3: $x_1 = x_2 = x^* \le K$. In this case, we have, for $x \ge x^*$,

$$V_1(x) = A_1 x^{\beta_1} + A_2 x^{\beta_2},$$
$$V_2(x) = B_1 x^{\beta_1} + B_2 x^{\beta_2},$$

and $V_1(x) = V_2(x) = K - x$ for $x \in [0, x^*]$. The smooth fit scheme leads to

$$\begin{cases} A_1(x^*)^{\beta_1} + A_2(x^*)^{\beta_2} &= K - x^*, \\ \beta_1 A_1(x^*)^{\beta_1} + \beta_2 A_2(x^*)^{\beta_2} &= -x^*, \end{cases} \tag{19}$$

and

$$\begin{cases} B_1(x^*)^{\beta_1} + B_2(x^*)^{\beta_2} &= K - x^*, \\ \beta_1 B_1(x^*)^{\beta_1} + \beta_2 B_2(x^*)^{\beta_2} &= -x^*. \end{cases} \tag{20}$$

Necessarily, we have $A_1 = B_1$ and $A_2 = B_2$, and therefore, $V_1 = V_2$.

Defining $V(x) = V_1(x) = V_2(x)$, then for $x > x^*$, the first equation in (6) reduces to

$$rV(x) = x\mu_i V'(x) + \frac{1}{2}x^2\sigma_i^2 V''(x),$$

for both $i = 1, 2$. This implies

$$V_1(x) = V_2(x) = \begin{cases} \dfrac{(K - x^*)x^\beta}{(x^*)^\beta} & \text{if } x > x^*, \\ K - x & \text{if } x \le x^*, \end{cases}$$

where $x^* = K\beta/(\beta - 1)$ and β is the negative solution of

$$r - \left(\mu_i - \frac{1}{2}\sigma_i^2\right)\beta - \frac{1}{2}\sigma_i^2\beta^2 = 0$$

for $i = 1$ or $i = 2$.

3. Optimality of the solution. Now, we prove the optimality of $V_i(x)$ and x_i for $i = 1, 2$ derived in the previous section. For general results on stochastic calculus, we refer to the books by Karatzas and Shreve [11], McKean [15], and Revuz and Yor [18].

Recall

$$V^*(x, i) = \sup_\tau E[e^{-r\tau}(K - X(\tau))^+ \mid X(0) = x, \epsilon(0) = i].$$

Then we must prove the following claim.

THEOREM 3.1. *Suppose that* (15) (resp., (17)) *has a solution* (x_1^*, x_2^*) *such that* $0 < x_1^* \le K$ *and* $0 < x_2^* \le K$. *Assume* $V_i(x) > (K-x)^+$ *on* (x_i^*, ∞) *and* $\mu_i \ge 0$ *for* $i = 1, 2$. *Define*

$$D = \{(x, i) \mid V_i(x) > (K-x)^+\},$$

and let

$$\tau^* = \inf\{t \ge 0 \mid (X(t), \epsilon(t)) \notin D\}.$$

Then τ^* *is an optimal stopping time, and* $V_i(x)$ *are value functions (i.e.,* $V_i(x) = V^*(x, i)$*) and are given by* (16) (resp., (18)).

Proof. It is easy to see that $V_i(\infty) = 0$, $i = 1, 2$, and

$$D = \{(x, 1) \mid x \in (x_1^*, \infty)\} \cup \{(x, 2) \mid x \in (x_2^*, \infty)\}.$$

For any $v(x, i) \in C^2$, define

$$\mathcal{L}v(x,i) = x\mu_i \frac{\partial v(x,i)}{\partial x} + \frac{1}{2}x^2\sigma_i^2 \frac{\partial^2 v(x,i)}{\partial x^2} + \lambda_i(v(x, 3-i) - v(x,i)) - rv(x,i).$$

Let $v(x, i) = V_i(x)$. Then $\mathcal{L}v \le 0$ on $(x, i) \in D$. Using Dynkin's formula, we have

$$d(e^{-rt}v(X(t), \epsilon(t))) = e^{-rt}\mathcal{L}v(X(t), \epsilon(t))dt + d(\text{martingale}).$$

For any stopping time τ, it follows, from a smooth approximation approach for variational inequalities in Øksendal [16, p. 204], that

$$v(x,i) \ge E[e^{-r\tau}v(X(\tau), \epsilon(\tau))] \ge E[e^{-r\tau}(K - X(\tau))^+]. \tag{21}$$

To show the optimality of τ^*, note that if $\tau^* < \infty$, then $v(X(\tau^*), \epsilon(\tau^*)) = (K - X(\tau^*))^+$. In this case, Dynkin's formula yields $v(x, i) = E[e^{-r\tau^*}(K - X(\tau^*))^+]$. Otherwise, let

$$D_k = D \cap \{x < k\}, \quad \text{for } k = 1, 2, \ldots.$$

Let $\tau_k = \inf\{t \ge 0 \mid (X(t), \epsilon(t)) \notin D_k\}$. Then we can show that $\tau_k \to \tau^*$ a.s. Moreover, as in Zhang [24, Theorems 4.5 and 4.6], we can show that, for each k, $\tau_k < \infty$ a.s. Using the definition of τ_k, we have, for $k > K$,

$$v(X(\tau_k), \epsilon(\tau_k)) = v(X(\tau_k), \epsilon(\tau_k))I_{\{X(\tau_k)=k\}} + v(X(\tau_k), \epsilon(\tau_k))I_{\{X(\tau_k)<k\}}.$$

Note that

$$v(X(\tau_k), \epsilon(\tau_k))I_{\{X(\tau_k)<k\}} = (K - X(\tau_k))^+ I_{\{X(\tau_k)<k\}} \le (K - X(\tau_k))^+.$$

Moreover, note that $0 \le v(x, i) \le K$ and $e^{-r\tau_k}I_{\{X(\tau_k)=k\}} \to 0$, as $k \to \infty$, a.s. It follows that

$$E[e^{-r\tau_k}v(X(\tau_k), \epsilon(\tau_k))I_{\{X(\tau_k)=k\}}] \to 0.$$

Therefore, we have, as $k \to \infty$,

$$v(x,i) \le E[e^{-r\tau_k}v(X(\tau_k), \epsilon(\tau_k)] \le E[e^{-r\tau^*}(K - X(\tau^*))^+].$$

Combining this with (21), we have

$$v(x,i) = E[e^{-r\tau^*}(K - X(\tau^*))^+].$$

This completes the proof.

Remark 3.1. As mentioned earlier, when $\sigma_1 \neq \sigma_2$, $\epsilon(t)$ becomes observable from the quadratic variation of $X(t)$ by Ito's calculus (see McKean [14]) and yields the joint Markov structure of $(X(t), \epsilon(t))$. This is one of the key points for our analysis. Although the case $\sigma_1 = \sigma_2$ is of independent interest from the filtering perspective since $\epsilon(t)$ is no longer observable (see Wonham [22] for estimating the probability distribution of $\epsilon(t)$, Liptser and Shiryayev [13] for general filtering, and Zhang [26, 27] for state detection and hybrid filtering), the option pricing problem is exactly the McKean problem, since a Girsanov transformation will reduce the regime switching model to the Black–Scholes model.

Remark 3.2. When $\lambda_1\lambda_2 = 0$, the corresponding $\epsilon(t)$ reduces to a single jump process, and the value functions can be solved sequentially using our method.

Remark 3.3. The optimality proof in Theorem 3.1 indicates the uniqueness of the value functions and that of the corresponding x_i's. Moreover, the assumption $V_i(x) > (K - x)^+$ or the existence of x_1, x_2 would be redundant if we assume the C^1 smoothness at the boundary x_1, x_2.

Remark 3.4. The assumption $\mu_i \geq 0$ guarantees that $e^{-rt}v(X(t), \epsilon(t))$ is a supermartingale. This is not restrictive in general. Indeed, it is standard in risk-neutral option pricing to have $\mu_1 = \mu_2 = r \geq 0$, following a change of measure via the Girsanov transformation.

Remark 3.5. It is clear from our analysis that a closed-form solution is possible if and only if K, the number of states of $\epsilon(t)$, equals two, since in general an algebraic equation of order $2K$ needs to be solved.

4. Numerical simulation. In this section we perform numerical experiments to compare the analytical solutions with the TPBVDE solutions studied in Zhang [24], together with the numerical results derived from a dynamic programming (DP) approach.

To this end, we first briefly review both DP and TPBVDE methods.

4.1. Dynamic programming. The DP approach we adopt here is built on the discretization method of the regime switching model proposed by Guo [6].

For a fixed T, let us divide the interval $[0, T]$ into N subintervals such that $T = Nh$. Moreover, if we define

$$u_i = e^{\sigma_i\sqrt{h}}, \quad l_i = e^{-\lambda_i h}, \quad d = e^{-rh}, \tag{22}$$

$$p_i = \frac{\mu_i h + \sigma_i\sqrt{h} - 0.5\sigma_i^2 h}{2\sigma_i\sqrt{h}}, \quad p_i + q_i = 1, \tag{23}$$

then the discrete counterpart of the process $(X(t), \epsilon(t))$ becomes the two-dimensional Markov chain (X_n, ϵ_n) that satisfies the recurrence

$$(X_n, \epsilon_n) = \eta_n^{(\epsilon_n, \epsilon_{n-1})}(X_{n-1}, \epsilon_{n-1}), \tag{24}$$

where $\eta_n^{i,j}$ are independently and identically distributed (i.i.d.) random variables taking values u_j with probability $p_j(\chi_{i,1-j} + (-1)^{\chi_{i,1-j}}e^{-\lambda_j h})$ and $1/u_j$ with probability

$(1-p_j)(\chi_{i,1-j}+(-1)^{\chi_{i,1-j}}e^{-\lambda_j h})$, respectively, where $(i,j=1,2)$ and

$$\chi(i,j)=\begin{cases}1, & i=j=1,2,\\ 0, & i\neq j.\end{cases}$$

In other words, (X_n,ϵ_n) is a random walk taking values on the set $(u_1^m u_2^n,i)$ with $i=1,2$ and $m,n=0,\pm1,\pm2,\ldots$ such that X_n represents the stock price at time n and ϵ_n the state of the market at time n.

Furthermore, the optimal stopping problem in question becomes

$$\widetilde{V}_i(x)=\sup_{\tau\in\{1,2,\ldots,\}}E[d^n(K-X_n)^+|\epsilon_0=i,X_0=x]. \tag{25}$$

Given the Markov chain $X=((X_n,\epsilon_n),\mathcal{F}_n,P)$, the optimal stopping problem (25) can be derived via the following dynamic programming principle:

$$\begin{aligned}
W_0(x) &= (K-x)^+,\\
Z_0(x) &= (K-x)^+,\\
W_m(x) &= \max\Big\{W_{m-1}(x),dp_1l_1W_{m-1}(u_1x)+dl_1q_1W_{m-1}\left(\tfrac{x}{u_1}\right)\\
&\qquad +d(1-l_1)p_2Z_{m-1}(u_2x)+d(1-l_1)q_2Z_{m-1}\left(\tfrac{x}{u_2}\right)\Big\},\\
Z_m(x) &= \max\Big\{Z_{m-1}(x),dp_2l_2Z_{m-1}(u_2x)+dl_2q_2Z_{m-1}\left(\tfrac{x}{u_2}\right)\\
&\qquad +(1-l_2)dp_1W_{m-1}(u_1x)+(1-l_2)dq_1W_{m-1}\left(\tfrac{x}{u_1}\right)\Big\}.
\end{aligned}$$

It is clear that $W_m(x),Z_m(x)$ are nondecreasing sequences, and

$$\widetilde{V}_1(x)=\lim_{n\to\infty}W_n(x),$$

$$\widetilde{V}_2(x)=\lim_{n\to\infty}Z_n(x).$$

Evidently, $\widetilde{V}_1(x)$ and $\widetilde{V}_2(x)$ are bounded nonnegative decreasing functions, and $\widetilde{V}_1(x)\geq(K-x)^+$, $\widetilde{V}_2(x)\geq(K-x)^+$. They are also called the least excessive dominating functions.

If we define

$$x_1=\min\left\{x\geq0,\min_{i\in\{1,2\}}\widetilde{V}_i(x)=(K-x)^+\right\}$$

and

$$x_2=\min\left\{x\geq0,\max_{i\in\{1,2\}}\widetilde{V}_i(x)=(K-x)^+\right\},$$

then x_1,x_2 are the so-called free boundary for the stopping rule.

With proper smooth conditions, $\widetilde{V}_i(x)$ coincides with $V(x,i)$ and hence with $V^*(x,i)$. For more detailed discussions on the least excessive dominating function and its application in option pricing, interested readers are referred to Guo [7] and Shiryayev et al. [21].

4.2. The TPBVDE approach. The TPBVDE approach was proposed by Zhang [24] to derive certain selling rules of threshold type. The stopping rule is to stop whenever the underlying stock price reaches two predefined bounds, an upper bound B or a lower bound A:

$$\tau_0 = \inf\{t > 0 \mid X(t) \notin (A, B)\}.$$

This rule is suboptimal since it limits the holder's choice to a smaller class of stopping times. If one takes $A = x^*$ and $B = \infty$ in Case 3, then it leads to a preferable stopping rule of $\tau_0 = \tau^*$.

The basic idea is to first choose a region of (A, B) so that for any given $0 \le a < b$,

$$X(0)e^{-b} \le A \le X(0)e^{-a},$$

$$X(0)e^{a} \le B \le X(0)e^{b}.$$

Next, we choose A and B within this interval to maximize

$$E[e^{-r\tau}(K - X(\tau)^+].$$

With this given A and B, the value function can thus be derived via analysis of a TPBVDE. (See [24] for details.)

4.3. Numerics. This section will report the numerical comparison results. First, we take

$$r = 3, \quad \mu_1 = \mu_2 = 3, \quad K = 5,$$

$$\lambda_1 = \lambda_2 = 100, \quad \sigma_1 = 9, \quad \sigma_2 = 5,$$

and compare the closed-form solution with the numerical solutions from the DP and TPBVDE methods; for the latter, we use the lower bound $a = 0$ and upper bounds $b = 3, b = 10$. The numerical results are plotted in Figure 1 and labeled with $V^{\mathrm{e}}(x, i)$, $V^{\mathrm{DP}}(x, i)$, $V^{\mathrm{b}=3}(x, i)$, and $V^{\mathrm{b}=10}(x, i)$, accordingly.

After 4000 iterations, with $N = 100,000$ and $h = 0.0001$, we obtain the threshold levels $(x_1^*, x_2^*) = (0.454, 0.617)$ for the DP approach, in comparison to the $(x_1^*, x_2^*) = (0.441, 0.614)$ derived from the closed-form solution.

Figure 2 confirms $V_i(x) \ge (K - x)^+$ and illustrates the differences of these value functions. As is shown, the accuracy of the two-point value method improves with increases in the upper bound b. The DP approach approximates the exact solutions better than the TPBVDE method for $b = 3$, while the converse is true with $b = 10$. In addition, these differences equal zero on the intervals (x_1^*, ∞) and (x_2^*, ∞) for $\epsilon(0) = 1$ or 2, respectively.

Next, we check the monotonicity of these threshold levels with respect to σ_i and λ_i. First, we vary σ_1 and keep all other parameters fixed. The resulting (x_1^*, x_2^*) are listed in Table 1. Both threshold levels x_1^* and x_2^* decrease with decreasing σ_1. This shows that a larger σ_1 leads to a higher option premium and therefore a smaller threshold level.

We then vary λ_1. The result in Table 2 implies that both x_1^* and x_2^* increase with λ_1 increasing: this is because a larger λ_1 implies a shorter period for $\epsilon(t)$ staying at

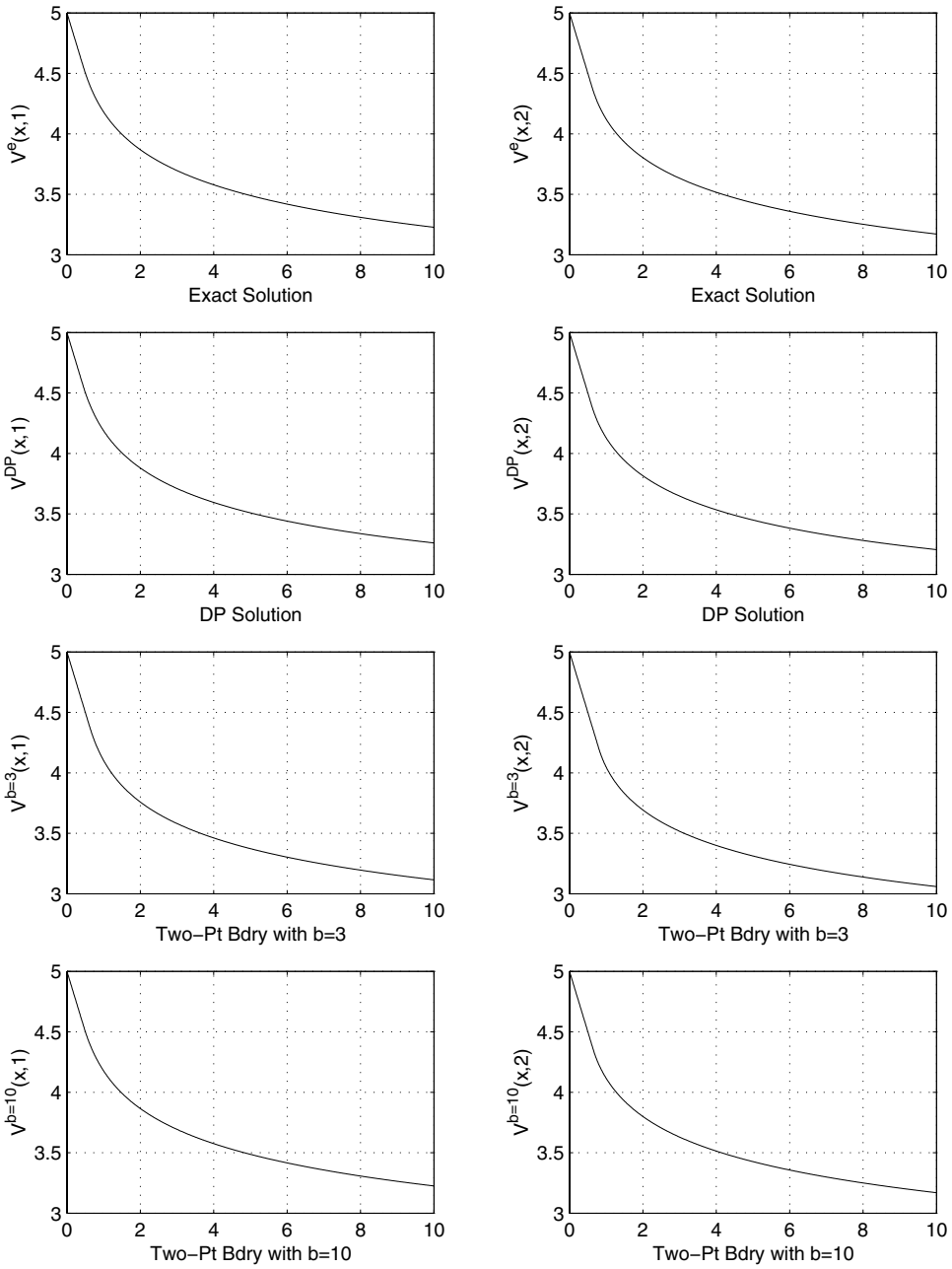


FIG. 1. *Value functions.*

$\epsilon(t) = 1$ and a smaller weight on $\sigma_1 = 9$ ($> \sigma_2 = 5$), which leads to smaller average volatility.

These monotonicity properties may be better explained using the average volatility $\overline{\sigma} = \sqrt{\nu_1\sigma_1^2 + \nu_2\sigma_2^2}$, where (ν_1, ν_2) is the stationary distribution corresponding to the generator of $\epsilon(t)$. The results in Tables 1 and 2 suggest that both x_1^* and x_2^*

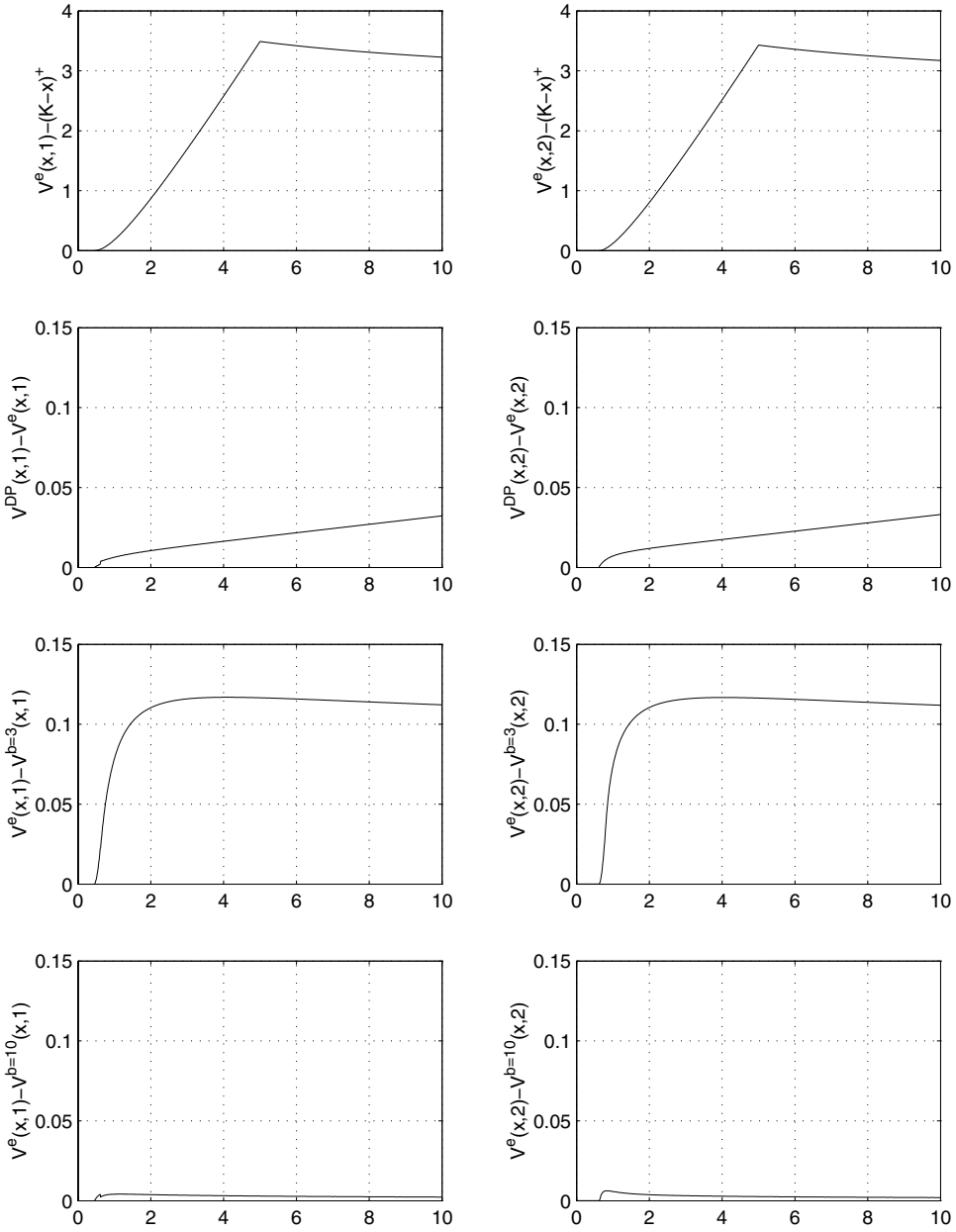


FIG. 2. *Differences between value functions.*

decrease with decreasing $\overline{\sigma}$.

Not surprisingly, the convergence rate of the DP approach depends on the choice of parameters. This in essence has to do with the specific discretization method of the underlying diffusion process. For example, with the same parameters specified above and with perturbations on the magnitude of r, we found that the smaller the r, the longer the computational time.

TABLE 1
Dependency on σ_1, given $\sigma_2 = 5$.

σ_1	7	8	9	10	11	12
Exact	(.646,.764)	(.531,.683)	(.441,.614)	(.369,.554)	(.312,.505)	(.266,.462)
DP	(.660,.773)	(.545,.687)	(.454,.617)	(.381,.557)	(.324,.506)	(.277,.465)

TABLE 2
Dependency on λ_1, given $\lambda_2 = 100$.

λ_1	80	90	100	110	120	130
Exact	(.425,.596)	(.433,.605)	(.441,.614)	(.448,.621)	(.456,.629)	(.463,.637)
DP	(.437,.599)	(.446,.607)	(.454,.617)	(.461,.624)	(.469,.632)	(.476,.640)

As far as total CPU usage is concerned, the DP approach took substantially longer time than the closed-form and the TPBVDE methods. For example, with a basic Linux 7.2 i386 system, it took a little more than 30 minutes for our DP solution to complete 4000 iterations, while it took just seconds for both the exact and TPBVDE methods.

5. Concluding remarks. In this paper we have derived a closed-form solution to the optimal stopping problem for pricing perpetual American put options in a regime switching model.

It remains to be seen whether there are alternative methods for deriving the solution. One obvious candidate is the first passage time technique, which was exploited in solving the McKean problem (McKean [14] and Karlin and Taylor [12]). However, despite the two promising features that (i) $(X(t), \epsilon(t))$ is jointly Markovian and (ii) the free boundaries are of threshold type, it seems hard to explicitly solve the integral equation system using results of the first passage time for regime switching models (derived in Guo [8]). The main obstacle seems to be the instantaneous jump due to the regime switching.

It is also of interest to extend our analysis to the case when T is finite. Needless to say, this case would be mathematically interesting and practically appealing. However, a closed-form solution for a finite time horizon problem with regime switching is difficult to obtain. Even the special case with no regime switching remains an open problem to date. Moreover, with all the structural insights gained from the infinite case, it is not even clear whether the boundary is monotonic; i.e., will $x_1 < x_2$ imply $x_1(T) < x_2(T)$? Assuming this monotonicity condition a priori, Buffington and Elliott [2] extended our analysis and obtained certain properties for the value functions of American put options with $T < \infty$.

Nevertheless, our hope is that the closed-form solutions in this paper will provide better understanding of and some insight into the nature of optimal stopping rules, and our approach can be useful for numerical approximations of long-term American options.

Acknowledgments. We thank the referees for a very careful reading of the manuscript and many constructive suggestions.

REFERENCES

[1] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Political Economy, 81 (1973), pp. 637–654.

[2] J. Buffington and R. J. Elliott, *American options with regime switching*, Int. J. Theor. Appl. Finance, 5 (2002), pp. 497–514.

[3] G. B. Di Masi, Yu. M. Kabanov, and W. J. Runggaldier, *Mean-variance hedging of options on stocks with Markov volatilities*, Theory Probab. Appl., 39 (1994), pp. 172–181.

[4] D. Duffie, *Dynamic Asset Pricing Theory*, 2nd ed., Princeton University Press, Princeton, NJ, 1996.

[5] X. Guo, *Inside Information and Stock Fluctuations*, Ph.D. Dissertation, Department of Mathematics, Rutgers University, Newark, NJ, 1999.

[6] X. Guo, *Inside information and option pricings*, Quant. Finance, 1 (2000), pp. 38–44.

[7] X. Guo, *An explicit solution to an optimal stopping problem with regime switching*, J. Appl. Probab., 38 (2001), pp. 464–481.

[8] X. Guo, *When the "bear" meets the "bull": A first passage time problem in a hidden Markov process*, Methodol. Comput. Appl. Probab., 3, (2001), pp. 134–143.

[9] S. D. Jacka, *Optimal stopping and the American put*, Math. Finance, 1 (1991), pp. 1–14.

[10] I. Karatzas, *On the pricing of American options*, Appl. Math. Optim., 17 (1988), pp. 37–60.

[11] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1998.

[12] S. Karlin and H. Taylor, *A First Course in Stochastic Processes*, Academic Press, New York, 1975.

[13] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes*, Springer-Verlag, New York, 1977.

[14] H. P. McKean, *A free boundary problem for the heat equation arising from a problem in mathematical economics*, Ind. Management Rev., 6 (1965), pp. 32–39.

[15] H. P. McKean, *Stochastic Integrals*, Academic Press, New York, 1969.

[16] B. Øksendal, *Stochastic Differential Equations*, 4th ed., Springer-Verlag, New York, 1995.

[17] E. Renault and N. Touzi, *Option hedging and implied volatilities in a stochastic volatility model*, Math. Finance, 6 (1996), pp. 279–302.

[18] D. Revuz and M. Yor, *Continuous Martingale and Brownian Motion*, Springer-Verlag, New York, 1991.

[19] H. Robbins, D. Sigmund, and Y. Chow, *Great Expectations: The Theory of Optimal Stopping*, Houghton Mifflin, Boston, 1971.

[20] P. A. Samuelson, *Mathematics of speculative price (with an appendix on continuous-time speculative processes by R. C. Merton)*, SIAM Rev., 15 (1973), pp. 1–42.

[21] A. N. Shiryaev, Yu. M. Kabanov, O. D. Kramkov, and A. V. Mel'Nikov, *Toward the theory of pricing of options of both European and American types* I*, Discrete time*, Theory Probab. Appl., 39 (1994), pp. 14–60.

[22] W. M. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering*, SIAM J. Control, 2 (1964), pp. 347–369.

[23] D. D. Yao, Q. Zhang, and X. Y. Zhou, *Option Pricing with Markov-Modulated Volatility*, preprint.

[24] Q. Zhang, *Stock trading: An optimal selling rule*, SIAM J. Control Optim., 40 (2001), pp. 64–87.

[25] Q. Zhang and G. Yin, *Nearly optimal asset allocation in hybrid stock-investment models*, J. Optim. Theory Appl., to appear.

[26] Q. Zhang, *Nonlinear filtering and control of a switching diffusion with small observation noise*, SIAM J. Control Optim., 36 (1998), pp. 1638–1668.

[27] Q. Zhang, *Hybrid filtering for linear systems with non-Gaussian disturbances*, IEEE Trans. Automat. Control, 45 (2000), pp. 50–61.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 2050–2075

MINIMAL ROTATIONALLY INVARIANT BASES FOR HYPERELASTICITY*

GREGORY H. MILLER†

Abstract. Rotationally invariant polynomial bases of the hyperelastic strain energy function are rederived using methods of group theory, invariant theory, and computational algebra. A set of minimal basis functions is given for each of the 11 Laue groups, with a complete set of rewriting syzygies. The ideal generated from this minimal basis agrees with the classic work of Smith and Rivlin [*Trans. Amer. Math. Soc.*, 88 (1958), pp. 175–193]. However, the structure of the invariant algebra described here calls for fewer terms, beginning with the fourth degree in strain, for most groups.

Key words. elasticity, hyperelasticity, symmetry, integrity basis, Gröbner bases

AMS subject classifications. 74A20, 74B20, 13P10, 13A50, 14L24, 68W30

DOI. 10.1137/S0036139903438776

1. Introduction.

In 1958 Smith and Rivlin [20] derived a set of so-called integrity bases: a finite set of homogeneous polynomial functions of the strain, unique to each of 11 sets of symmetry groups (the Laue groups) which govern the symmetry of the strain energy function. These invariants were derived using theorems for the invariants of permutation groups (e.g., Weyl [25]). By "basis" it is meant that any arbitrary symmetry-invariant polynomial may be rewritten as a polynomial in these basis functions. Since the number of invariant homogeneous polynomials is unbounded, the discovery of a finite basis makes the problem of hyperelastic constitutive modeling tractable (and, indeed, far simpler than constructing a symmetry-invariant function as an expansion in symmetry-correct fourth- and higher-order elastic constant tensors). The integrity bases are particularly important for modeling time-dependent large deformation solid mechanics. Examples of computational methods requiring properly invariant hyperelastic descriptions include [15, 16, 14].

Since the important classification by Smith and Rivlin, a number of significant advances have been made in computational tools for algebra, particularly the theory of Gröbner bases, which has opened powerful new approaches to the study of group invariants (e.g., [24]). In this paper the elastic integrity bases are rederived using these new algorithmic approaches. The main point of this paper is to reexamine the invariant structure of hyperelastic materials using these modern methods. It will be shown that the integrity bases of Smith and Rivlin are correct in the sense that their integrity bases are finite bases which generate the correct invariant polynomial ideals. However, for most symmetry groups a number of syzygies exist which interrelate the invariant basis polynomials, and therefore their bases are not minimal (syzygies and minimality are described in this context in, e.g., [22]). It will be shown that for most groups, beginning at degree 4 in the Cauchy tensor (equivalently, the Lagrangian

*Received by the editors December 12, 2003; accepted for publication (in revised form) May 19, 2004; published electronically September 2, 2004. This work was sponsored by the U.S. Department of Energy (DOE) Mathematical, Information, and Computing Sciences Division contracts DE-AC03-76SF00098 and DE-FG02-03ER25579.

http://www.siam.org/journals/siap/64-6/43877.html

†Department of Applied Science, University of California, 1 Shields Avenue, Davis, CA 95616 (grgmiller@ucdavis.edu) and Applied Numerical Algorithms Group, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720.

strain tensor), the Smith and Rivlin integrity bases imply the existence of unnecessary polynomial terms.

In section 2 properties of the strain energy function are reviewed. This section sets the thermodynamic context for subsequent more mathematical sections and identifies the Cauchy tensor as the key fundamental variable controlling hyperelasticity. In section 3 the group theoretical properties of the 32 crystallographic point groups are described as they relate to the Cauchy tensor. Section 4 extends group theory to describe the algebraic structure of polynomial invariants of each group. Algebraic algorithms are described briefly in section 5 and used to construct a complete set of invariants. Simplifying relations among these invariants, syzygies are described in the appendix. A complete set of "rewriting syzygies" is presented, with which one could cast an arbitrary invariant polynomial into a minimal form. Select syzygies are also presented which demonstrate the algebraic dependence of "secondary invariants" upon the "primary invariants." Concluding remarks are made in section 6.

2. The strain energy function. The fundamental kinematic variable that governs hyperelasticity is the deformation tensor

$$F_{\alpha\beta} = \frac{\partial x_\alpha}{\partial a_\beta}, \tag{2.1}$$

which describes the deformation of a spatial (Eulerian) frame x with respect to a material (Lagrangian) frame a. The internal energy $\mathcal{E}$ is some function of the nine components of F, entropy S, and possibly other material constitutive parameters: $\mathcal{E} = \mathcal{E}(F, S)$.

Since the internal energy is a scalar function, its value must be independent of the reference frame of the observer. Thus, an observer utilizing a reference frame $\hat{x}$ will interpret the laboratory reference frame x rotated through an arbitrary orthogonal rotation Q ($Q^{-1} = Q^T$) and translated by an arbitrary vector $\hat{x}_0$:

$$\hat{x} = \hat{x}_0 + Qx, \tag{2.2}$$

$$\hat{F} = QF. \tag{2.3}$$

The possibly time-dependent translation $\hat{x}_0$ is independent of the material reference frame $\{a\}$ and therefore does not affect the observer's deformation tensor $\hat{F}$. In the observer's frame, the internal energy would be $\hat{\mathcal{E}} = \hat{\mathcal{E}}(QF, S)$. And so, for the internal energy to be independent of the reference frame of the observer, the function must depend not on the components of F individually but upon some combination of them that removes the dependence on Q.

One way of removing the Q-dependence is to factor $\hat{F}$ into a matrix of pure stretches and a matrix of rotations. The so-called right-polar decomposition of F is

$$F = RU, \tag{2.4}$$

where R is a rotation ($R^{-1} = R^T$) and U is symmetric. This decomposition is unique, with

$$U^T U = F^T F = F^T Q^T Q F = \hat{F}^T \hat{F}. \tag{2.5}$$

Instead of solving (2.5) for the six independent components of U, one might use directly the six independent components of C—the Cauchy tensor (or "right Cauchy–Green tensor")

$$C = F^T F, \tag{2.6}$$

$$\mathcal{E} = \mathcal{E}(C, S). \tag{2.7}$$

These manipulations determine the functional dependence of the internal energy in such a way as to make the result independent of the reference frame of an observer. The Cauchy tensor remains, however, dependent upon the orientation of the material with respect to the material reference frame a. For crystals with no rotational symmetry, this result is adequate, and one may without loss of generality construct hyperelastic equations of state (2.7) that are consistent with all symmetry constraints.

There are 230 space groups that classify the symmetry of single crystals. These are based upon 32 crystallographic point groups, which derive from consideration of rotations and reflections (reflections may also be referred to as improper S_1 rotations; collectively such operations will be called simply "rotations"), and become 230 upon consideration of translations consistent with the rotational symmetry. To discuss rotational invariance it is sufficient to consider the point groups. Of these 32 point groups, only two (C_1 and C_i) are correctly modeled by (2.7) without additional considerations of symmetry. The remaining 30 point groups classify materials which are symmetric with respect to certain discrete symmetry operations $\bar{\pi}$ on the atomic coordinates in the material reference frame:

$$\check{a} = \bar{\pi}^{-1} a, \tag{2.8}$$

$$\check{F} = F\bar{\pi}, \tag{2.9}$$

$$\check{C} = \bar{\pi}^T C \bar{\pi}. \tag{2.10}$$

For the internal energy to be invariant with respect to each of these discrete rotational mappings, one must have

$$\mathcal{E} = \mathcal{E}(C, S) = \mathcal{E}(\bar{\pi}^T C \bar{\pi}, S) \quad \forall\ \bar{\pi} \in \bar{\Gamma}^{\bar{G}}, \tag{2.11}$$

where $\bar{\Gamma}^{\bar{G}}$ represents the set of rotation operations of the crystallographic point group $\bar{G}$ of the material (the symbols $\bar{\pi}$, $\bar{\Gamma}$, $\bar{G}$, etc. are used here to describe the group properties in the $\mathbb{R}^3$ coordinate space; the symbols π, Γ, G, etc. will denote the corresponding extension of these group properties to the $\mathbb{R}^6$ space of the unique Cauchy tensor elements).

3. Group theory. The crystallographic point groups may be described by a finite number of 3×3 matrices which rotate a vector, reflect it across a plane, or combinations thereof. The set $\bar{\Gamma}$ of these matrices $\bar{\pi}$ are a representation of a group algebra $\bar{G}$, which means (1) multiplication is associative, $(\bar{\pi}_\alpha \bar{\pi}_\beta)\bar{\pi}_\gamma = \bar{\pi}_\alpha(\bar{\pi}_\beta \bar{\pi}_\gamma)$ for each $\bar{\pi}_\alpha, \bar{\pi}_\beta, \bar{\pi}_\gamma \in \bar{\Gamma}$; (2) that for each $\bar{\pi}_\alpha, \bar{\pi}_\beta \in \bar{\Gamma}$, the product $\bar{\pi}_\alpha \bar{\pi}_\beta$ is also contained in $\bar{\Gamma}$; (3) there exists an identity $\bar{E} \in \bar{\Gamma}$ such that $\bar{E}\bar{\pi}_\alpha = \bar{\pi}_\alpha \bar{E} = \bar{\pi}_\alpha$ for each $\bar{\pi}_\alpha \in \bar{\Gamma}$; and (4) for each $\bar{\pi}_\alpha \in \bar{\Gamma}$, there exists an inverse $\bar{\pi}_\alpha^{-1}$ such that $\bar{\pi}_\alpha \bar{\pi}_\alpha^{-1} = \bar{E}$.

One may use the property (2.10) to construct a set of 6×6 matrix operators π_1

$$\pi_1 = \begin{pmatrix} \bar{\pi}_{11}^2 & \bar{\pi}_{21}^2 & \bar{\pi}_{31}^2 & 2\bar{\pi}_{31}\bar{\pi}_{21} & 2\bar{\pi}_{11}\bar{\pi}_{31} & 2\bar{\pi}_{11}\bar{\pi}_{21} \\ \bar{\pi}_{12}^2 & \bar{\pi}_{22}^2 & \bar{\pi}_{32}^2 & 2\bar{\pi}_{22}\bar{\pi}_{32} & 2\bar{\pi}_{32}\bar{\pi}_{12} & 2\bar{\pi}_{22}\bar{\pi}_{12} \\ \bar{\pi}_{13}^2 & \bar{\pi}_{23}^2 & \bar{\pi}_{33}^2 & 2\bar{\pi}_{33}\bar{\pi}_{23} & 2\bar{\pi}_{33}\bar{\pi}_{13} & 2\bar{\pi}_{23}\bar{\pi}_{13} \\ \bar{\pi}_{12}\bar{\pi}_{13} & \bar{\pi}_{22}\bar{\pi}_{23} & \bar{\pi}_{32}\bar{\pi}_{33} & \bar{\pi}_{22}\bar{\pi}_{33}+\bar{\pi}_{32}\bar{\pi}_{23} & \bar{\pi}_{12}\bar{\pi}_{33}+\bar{\pi}_{32}\bar{\pi}_{13} & \bar{\pi}_{12}\bar{\pi}_{23}+\bar{\pi}_{22}\bar{\pi}_{13} \\ \bar{\pi}_{11}\bar{\pi}_{13} & \bar{\pi}_{21}\bar{\pi}_{23} & \bar{\pi}_{31}\bar{\pi}_{33} & \bar{\pi}_{21}\bar{\pi}_{33}+\bar{\pi}_{31}\bar{\pi}_{23} & \bar{\pi}_{11}\bar{\pi}_{33}+\bar{\pi}_{31}\bar{\pi}_{13} & \bar{\pi}_{11}\bar{\pi}_{23}+\bar{\pi}_{21}\bar{\pi}_{13} \\ \bar{\pi}_{11}\bar{\pi}_{12} & \bar{\pi}_{21}\bar{\pi}_{22} & \bar{\pi}_{31}\bar{\pi}_{32} & \bar{\pi}_{21}\bar{\pi}_{32}+\bar{\pi}_{31}\bar{\pi}_{22} & \bar{\pi}_{11}\bar{\pi}_{32}+\bar{\pi}_{31}\bar{\pi}_{12} & \bar{\pi}_{11}\bar{\pi}_{22}+\bar{\pi}_{21}\bar{\pi}_{12} \end{pmatrix} \tag{3.1}$$

that transform the six-dimensional vector $\eta = (C_{11}, C_{22}, C_{33}, C_{23}(= C_{32}), C_{13}(= C_{31}), C_{12}(= C_{21}))^T$, or in Voigt notation $(C_1, C_2, C_3, C_4, C_5, C_6)^T$, according to

$$\check{\eta} = \pi_1 \eta. \tag{3.2}$$

In the language of Murnaghan [18, Ch. 3], the matrices π_1 are *symmetrized* Kronecker products of the transformation matrices $\bar{\pi}^T$. The d-form matrices π_d introduced below are symmetrized Kronecker d-powers of the transformations π_1.

Note that the set Γ of matrices π_1, formed from the elements $\bar{\pi} \in \bar{\Gamma}$ of group $\bar{G}$, define a group algebra G that may be different from $\bar{G}$ (e.g., [8]). In particular, the transformation matrices (3.1) effectively introduce inversion symmetry where none may have existed in the original group. Thus, as with Laue diffraction, the 32 crystallographic point groups reduce immediately to the 11 Laue groups.

A linear combination κ of elements of the Cauchy tensor is invariant to all symmetry operations if for each $\pi_1 \in \Gamma$ one has $\kappa = \pi_1 \kappa$; thus κ must be an eigenvector of each matrix π_1 with eigenvalue 1. Or

$$\kappa = \mathcal{P}_R \kappa, \tag{3.3}$$

with

$$\mathcal{P}_R = \frac{1}{|\Gamma|} \sum_{\pi_1 \in \Gamma} \pi_1 \tag{3.4}$$

and $|\Gamma|$ the cardinality of the group. $\mathcal{P}_R$ is the Reynolds operator, a special case of the more general symmetry projection operator which projects a vector onto an irreducible representation of the group (e.g., [6, Ch. 6]). The Reynolds operator projects a vector onto the unique totally symmetric representation. $\mathcal{P}_R$ is a projection, ${\mathcal{P}_R}^2 = \mathcal{P}_R$, by virtue of the property of groups that $\pi_{1\alpha}\Gamma = \Gamma$ for all $\pi_{1\alpha} \in \Gamma$. Consequently, the eigenvalues of $\mathcal{P}_R$ are all either 0 or 1. And, therefore, the number N_1 of linearly independent degree-1 invariant vectors is given by the number of unity eigenvalues of $\mathcal{P}_R$, which is equal to the trace of $\mathcal{P}_R$:

$$N_1 = \text{trace}(\mathcal{P}_R) = \frac{1}{|\Gamma|} \sum_{\pi_1 \in \Gamma} \text{trace}(\pi_1). \tag{3.5}$$

To evaluate this equation for any group, one tabulates the symmetry operations by type (Table 3.1 displays the operations of each group and their assumed orientation with respect to the assumed orthogonal Lagrangian coordinate system a). The numbers of symmetry operations π_1, by type and group Γ, are given in Table 3.2. The traces may be calculated from the eigenvalues listed in Table 3.3.

Invariants of higher degree lie in the $\binom{6+d-1}{d}$-dimensional space formed by the unique combinations of degree-d monomials (e.g., a basis for degree-2 monomials is given by the 21 homogeneous terms $C_i C_{j \geq i}$ in a process analogous to that described by (3.1)). From the matrices π_1, so-called d-form matrices π_d may be constructed easily to represent the action of the symmetry operations on the degree-d terms. The number N_d of linearly independent degree-d symmetry-invariant terms are constructed as in the degree-1 case with

$$\mathcal{P}_{R,d} = \frac{1}{|\Gamma|} \sum_{\pi_d \in \Gamma} \pi_d, \tag{3.6}$$

TABLE 3.1

Settings for those crystallographic point groups with planes and axes (Wulff stereographic projections). Bold lines are mirror planes. Open and closed circles are general positions, above and below plane $z = 0$, respectively. Closed symbols with n-fold symmetry are rotation axes, and mixed open-closed symbols with n-fold symmetry are improper rotations.

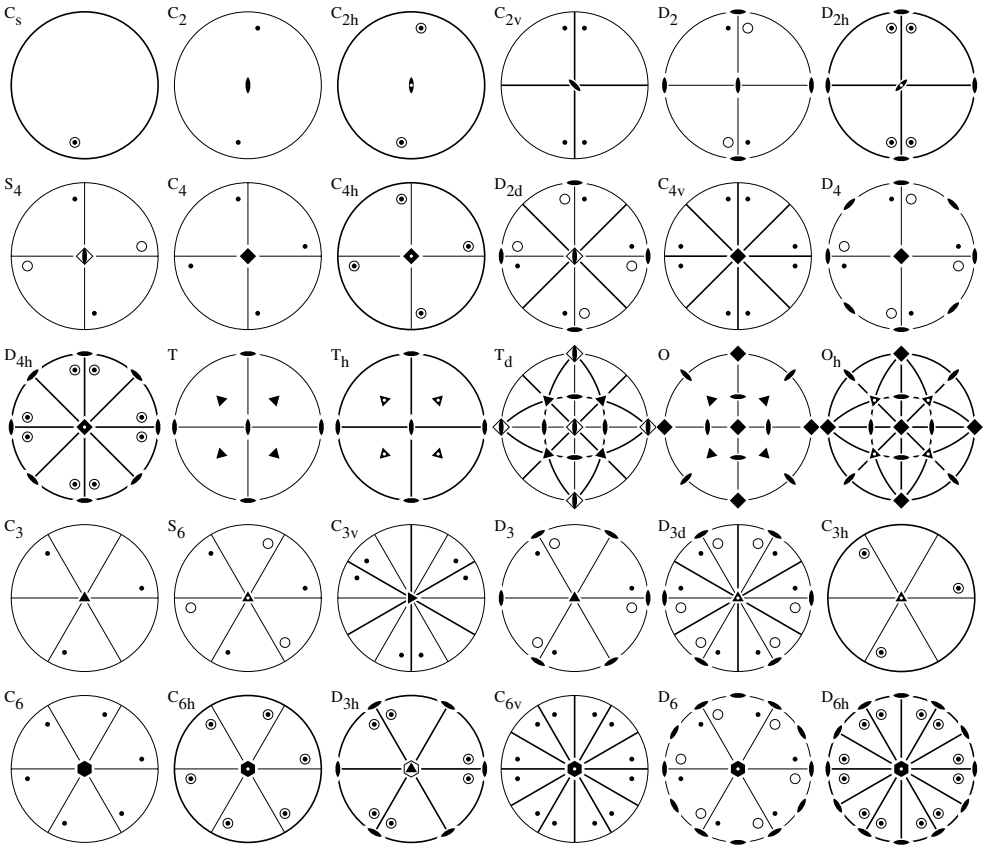


TABLE 3.2

Number of distinct occurrences of operations π in the crystallographic point groups, in the $\mathbb{R}^6$ space of the Cauchy tensor. The symbols used here are Schoenflies notation: E is the identity, I is inversion on all three orthogonal axes, C_n is an n-fold rotation, S_n is an improper n-fold rotation, and $\sigma = S_1$ is a mirror reflection.

Group Γ	$\lvert\Gamma\rvert$	E, I	C_2, σ	C_3, S_6	C_4, S_4	C_6, S_3
C_1, C_i	1	1				
C_s, C_2, C_{2h}	2	1	1			
C_{2v}, D_2, D_{2h}	4	1	3			
S_4, C_4, C_{4h}	4	1	1		2	
$D_{2d}, C_{4v}, D_4, D_{4h}$	8	1	5		2	
T, T_h	12	1	3	8		
T_d, O, O_h	24	1	9	8	6	
C_3, S_6	3	1		2		
C_{3v}, D_3, D_{3d}	6	1	3	2		
C_{3h}, C_6, C_{6h}	6	1	1	2		2
$D_{3h}, C_{6v}, D_6, D_{6h}$	12	1	7	2		2

TABLE 3.3
Eigenvalues of the point group operators π in the space $\mathbb{R}^6$ of the Cauchy tensor.

Operator π_1		Eigenvalues					
E	I	1	1	1	1	1	1
C_2	σ	1	1	1	1	-1	-1
C_3	S_6	1	1	$e^{\frac{2\pi i}{3}}$	$e^{\frac{2\pi i}{3}}$	$e^{-\frac{2\pi i}{3}}$	$e^{-\frac{2\pi i}{3}}$
C_4	S_4	1	1	-1	-1	$e^{\frac{\pi i}{2}}$	$e^{-\frac{\pi i}{2}}$
C_6	S_3	1	1	$e^{\frac{\pi i}{3}}$	$e^{-\frac{\pi i}{3}}$	$e^{\frac{2\pi i}{3}}$	$e^{-\frac{2\pi i}{3}}$

$$N_d = \operatorname{trace}(\mathcal{P}_{R,d}) = \frac{1}{|\Gamma|}\sum_{\pi_d\in\Gamma}\operatorname{trace}(\pi_d). \tag{3.7}$$

In practice it is not necessary to actually create the matrices π_d. The eigenvalues of the d-form matrices π_d are $\lambda_1^{d_1}\cdots\lambda_6^{d_6}$, with λ_i representing the ith eigenvalue of π_1, and with the exponents d_i subject to $d_1+\cdots+d_6=d$. Therefore,

$$\operatorname{trace}(\pi_d) = \sum_{d_1+\cdots+d_6=d}\lambda_1^{d_1}\cdots\lambda_6^{d_6}. \tag{3.8}$$

For completeness, one has also the scalar degree-0 term: the number "1." This polynomial invariant of degree 0 is generated by $\pi_0 = 1$, whence $N_0 = 1$.

Via the projections $\mathcal{P}_{R,d}$ group theory provides a method for the construction of all linearly independent degree-d symmetry-invariant homogeneous polynomials. The number of such polynomials is unbounded, however, since for any degree d the number of terms is at least as large as $\binom{N_1+d-1}{d}$—the number of distinct degree-d polynomials formed from by multiplying together different combinations of degree-1 polynomials.

A finite polynomial basis, a set of invariant polynomials from which all others may be constructed, exists. The number of terms in this basis and some properties of it are provided by theorems of invariant theory described below. The construction of the actual invariant polynomial bases is accomplished with the Gröbner basis methods of computational algebra, described in the subsequent section.

4. Invariant theory. The Hilbert series $\Phi(z)$ (also known as Poincaré's series) is a polynomial in the dummy variable z where the coefficient of z^n is the number of polynomial invariants of degree n. Spencer's generating functions [21], used in the study of invariants in continuum mechanics, are particular instances of the Hilbert series constructed by different means than those employed here. Following Sturmfels [24, Theorem 2.2.1] after [17],

$$\begin{aligned}
\Phi(z) &= \sum_{d=0}^{\infty} N_d z^d = \frac{1}{|\Gamma|}\sum_{\pi_1\in\Gamma}\sum_{d=0}^{\infty}\sum_{d_1+\cdots+d_6=d}\lambda_1^{d_1}\cdots\lambda_6^{d_6}z^d,\\
&= \frac{1}{|\Gamma|}\sum_{\pi_1\in\Gamma}\sum_{d_1,\ldots,d_6=0}^{\infty}\lambda_1^{d_1}\cdots\lambda_6^{d_6}z^{d_1+\cdots+d_6}\\
&= \frac{1}{|\Gamma|}\sum_{\pi_1\in\Gamma}\prod_{n=1}^{6}\left(1+(\lambda_n z)+(\lambda_n z)^2+\cdots\right)\\
&= \frac{1}{|\Gamma|}\sum_{\pi_1\in\Gamma}\prod_{n=1}^{6}\frac{1}{(1-\lambda_n z)} = \frac{1}{|\Gamma|}\sum_{\pi_1\in\Gamma}\frac{1}{\det(I-\pi_1 z)},
\end{aligned} \tag{4.1}$$

where $\lambda_\alpha = \lambda_\alpha(\pi_1)$, and where one assumes that $|z|<1$.

TABLE 4.1
Contributions of point group operators to the Hilbert series.

Operation π_1	Hilbert series contribution
E, I	$\dfrac{1}{(1-z)^6}$
C_2, σ	$\dfrac{1}{(1-z)^4(1+z)^2} = \dfrac{1}{(1-z)^2(1-z^2)^2}$
C_3, S_6	$\dfrac{1}{(1-z)^2(1+z+z^2)^2} = \dfrac{1}{(1-z^3)^2}$
C_4, S_4	$\dfrac{1}{(1-z^2)^2(1+z^2)} = \dfrac{1}{(1-z^2)(1-z^4)}$
C_6, S_3	$\dfrac{1}{(1-z)^2(1-z+z^2)(1+z+z^2)} = \dfrac{(1+z)}{(1-z)(1-z^6)}$

By means of this result, it is apparent that the Hilbert series for a given crystallographic point group may be algebraically constructed by summing factors $1/\det(I - \pi_1 z)$ corresponding to the individual operators π_1 that occur in the point group. These factors are summarized in Table 4.1.

One may then construct the Hilbert series according to the formula

$$\Phi_\Gamma(z) = \frac{1}{|\Gamma|}\left[\frac{N(E,I)}{(1-z)^6} + \frac{N(C_2,\sigma)}{(1-z)^2(1-z^2)^2} + \frac{N(C_3,S_6)}{(1-z^3)^2} + \frac{N(C_4,S_4)}{(1-z^2)(1-z^4)} + \frac{N(C_6,S_3)(1+z)}{(1-z)(1-z^6)}\right], \tag{4.2}$$

where $N(\pi, \dots)$ is the number of occurrences of symmetry operators $\pi, \dots \in \Gamma$, tabulated for the crystallographic point groups in Table 3.2.

Finite groups have the Cohen–Macaulay property of commutative algebra (e.g., [23]), which has significance for this project in that it implies that certain important properties of the invariant group algebra may be deduced by appropriate factorizations of the Hilbert series, called "Molien functions" or "Hironaka decompositions":

$$\Phi(z) = \frac{\sum_{i=0}^{t-1} z^{e_i}}{\prod_{j=1}^{n}(1-z^{d_j})}. \tag{4.3}$$

The interpretation of these factorizations is that there exist n *primary invariants* θ which comprise a "homogeneous system of parameters" (HSOP), with degrees $d_i = \deg(\theta_i)$, and where n is the number of variables (six for the Cauchy tensor elements). These are represented in the denominator of the Molien function. Since a factor $1/(1-z^d)$ contributes (multiplicatively) $1+z^d+z^{2d}+z^{3d}+\cdots$ to the Hilbert series, these factors are unrestricted in a polynomial representation. The numerator represents the t *secondary invariants* ϕ, with degrees $e_j = \deg(\phi_j)$ and cardinality

$$t = \frac{1}{|\Gamma|}\prod_{i=1}^{n} d_i, \tag{4.4}$$

including the degree-0 term "1." Since factors z^e in the numerator contribute only z^e (multiplicatively) to the Hilbert series, the implication is that these factors occur at most once in a polynomial representation.

TABLE 4.2

Molien factorizations of the Hilbert series of crystallographic point groups in the $\mathbb{R}^6$ space of the Cauchy tensor.

Groups Γ	Molien function
triclinic $C_1,\ C_i$	$\dfrac{1}{(1-z)^6}$
monoclinic $C_s,\ C_2,\ C_{2h}$	$\dfrac{1+z^2}{(1-z)^4(1-z^2)^2}$
orthorhombic $C_{2v},\ D_2,\ D_{2h}$	$\dfrac{1+z^3}{(1-z)^3(1-z^2)^3}$
tetragonal $S_4,\ C_4,\ C_{4h}$	$\dfrac{1+z^2+4z^3+z^4+z^6}{(1-z)^2(1-z^2)^3(1-z^4)}$
tetragonal $D_{2d},\ C_{4v},\ D_4,\ D_{4h}$	$\dfrac{1+2z^3+z^6}{(1-z)^2(1-z^2)^3(1-z^4)}$
cubic $T,\ T_h$	$\dfrac{1+3z^3+2z^4+2z^5+3z^6+z^9}{(1-z)(1-z^2)^2(1-z^3)^2(1-z^4)}$
cubic $T_d,\ O,\ O_h$	$\dfrac{1+z^3+z^4+z^5+z^6+z^9}{(1-z)(1-z^2)^2(1-z^3)^2(1-z^4)}$
trigonal $C_3,\ S_6$	$\dfrac{1+2z^2+6z^3+2z^4+z^6}{(1-z)^2(1-z^2)^2(1-z^3)^2}$
trigonal $C_{3v},\ D_3,\ D_{3d}$	$\dfrac{1+z^2+2z^3+z^4+z^6}{(1-z)^2(1-z^2)^2(1-z^3)^2}$
hexagonal $C_{3h},\ C_6,\ C_{6h}$	$\dfrac{1+3z^3+2z^4+2z^5+3z^6+z^9}{(1-z)^2(1-z^2)^2(1-z^3)(1-z^6)}$
hexagonal $D_{3h},\ C_{6v},\ D_6,\ D_{6h}$	$\dfrac{1+z^3+z^4+z^5+z^6+z^9}{(1-z)^2(1-z^2)^2(1-z^3)(1-z^6)}$

A consequence of the Molien function is that all symmetry-invariant polynomial functions P of the Cauchy tensor may be expressed in the form

$$P(\{\theta\},\{\phi\}) = \sum_{\alpha=0}^{t-1} \phi_\alpha P_\alpha(\{\theta\}), \tag{4.5}$$

where $P_\alpha(\{\theta\})$ is an arbitrary polynomial in the primary invariants, and where each secondary invariant ϕ_α occurs at most once. For all groups henceforth let $\phi_0 = 1$, and consider only nontrivial secondary invariants.

Molien factorizations of the crystallographic point groups, constructed using (4.2) for the invariants of the Cauchy tensor terms, are given in Table 4.2. These functions are fully reduced in that there is no common algebraic factor to both numerator and denominator, and in this sense the implied size of the invariant set is minimal. These factorizations are not unique. For example, the factorization displayed for group C_{2v} implies primary invariants of degree $1,1,1,2,2,2$ and one nontrivial secondary invariant of degree 3. However, multiplication of numerator and denominator by $(1+z)$ gives the function

$$\frac{1+z+z^2+z^4}{(1-z)^2(1-z^2)^4}, \tag{4.6}$$

implying primary invariants of degree $1,1,2,2,2,2$ and three nontrivial secondary invariants of degree $1,2,4$. If the form given in Table 4.2 exists, then the alternative form (4.6) is not minimal. The question of existence must be settled by constructing the algebraically independent basis functions of the group and partitioning them into primary and secondary invariants (e.g., [19, p. 101]). It will be shown that HSOPs of the degrees indicated by the reduced functions in Table 4.2 exist.

Since secondary ϕ_α is an invariant, so is ϕ_α^2. However, ϕ_α^2 is not represented in the Hilbert series. Thus, the invariant ϕ_α^2 must be expressible by some polynomial of the form (4.5). This implies the existence of syzygies—polynomial equalities that relate the secondary and primary invariants. A set of syzygies may be found to serve as "rewriting rules" for systematic conversion of a general polynomial $P(\{\theta\},\{\phi\})$ into the minimal form given by the right-hand side of (4.5).

An HSOP is a minimal set of algebraically independent polynomials, with cardinality 6, equal to the number of independent variables in C. There cannot be more algebraically independent homogeneous polynomials, and so any additional polynomial (in particular, the secondary invariants) must possess an algebraic dependence upon the primary invariants. The algebraic relations expressing a particular secondary invariant in terms of the primary ones are also expressible as syzygies.

The six primary invariants of each crystal point group are therefore algebraic (vs. polynomial) invariants; all polynomial invariants are expressible as algebraic functions of the primary invariants. This settles a conundrum regarding the number of degrees of freedom. The elastic Cauchy tensor has six degrees of freedom, and there are six algebraic invariants. The additional apparent degrees of freedom represented by the number of secondary invariants (aside from the trivial one, 1, of degree zero) are a consequence of assuming a polynomial form for the invariant energy function.

5. Computational algebra: Gröbner bases. In $\mathbb{R}^3$, the space of material coordinates a, and in $\mathbb{R}^6$, the space of Cauchy tensor components, some matrix operations of the groups (e.g., corresponding to rotations through $2\pi/3$) contain factors of $\sqrt{3}/2$. However, polynomials generated through the Reynolds operator contain only integer coefficients, so it is sufficient to study the properties of $\mathbb{Q}[C]$, the ring over rational numbers $\mathbb{Q}$ of polynomials in the Cauchy tensor elements C. Let $F=\{f_i | f_i \in \mathbb{Q}[C],\ f_i = \mathcal{P}_R f_i\}$ be some set of invariant polynomials. The ideal generated by F, $I(F)$ is the set of all polynomials $p_1 f_1 + p_2 f_2 + \cdots$, $p_i \in \mathbb{Q}[C]$, that are dependent on elements of F; i.e., $g \in I(F)$ and $h \in \mathbb{Q}[C]$ implies that $gh \in I(F)$, and $g, f \in I(F)$ implies that $g+f \in I(F)$. The objective is to construct the smallest basis F consisting of homogeneous invariant polynomials, with degrees consistent with the Molien function of the group, such that $I(F)$ is equal to the complete invariant ideal $I(\mathbb{Q}[C]^G)$ (see, e.g., [23, 4] and [3] for algebra concepts, and the latter also for Gröbner bases).

Algorithms designed to address this problem require the capability of deciding whether some polynomial f is in the ideal $I(F)$. The solution is to construct a special basis $\mathcal{GB}(F)$, the Gröbner basis, with $I(\mathcal{GB}(F)) = I(F)$. That is, $\mathcal{GB}$ is an alternative basis that generates the same ideal as F. The key property of a Gröbner basis is that for any $f \in I(F)$, $f \to_{\mathcal{GB}} 0$: f is reducible to zero by successive steps of a Euclidean reduction algorithm. The reduction property is linked to a notion of term order; a unique reduced Gröbner is specified by the basis F, and a specification of the term order $\succ$.

Group theory shows how, using the Reynolds operator, invariant polynomials may be generated, and from the Molien function one has an idea of what the degrees of primary and secondary invariants may be. Given a set of six homogeneous polynomial functions with degrees that are compatible with their being primary invariants, the first task is to show whether or not they are an HSOP. An algorithm for this task is given by Sturmfels [24, algorithm 2.5.3]. First, one uses the Reynolds operator[1] to generate a set of homogeneous polynomial invariants $\theta'(C)$ of the Cauchy tensor elements. Next, construct a polynomial basis set $F = \{\theta'_1 - y_1, \dots, \theta'_6 - y_6\}$ in the variables C and new slack variables y, with lexicographic order $C_1 > \cdots > C_6 > y_1 > \cdots > y_6$. This is an "elimination order" that systematically eliminates terms in C from the head of each polynomial in the basis during construction of the Gröbner basis (the head term is the greatest with respect to the specified order—lexicographic in this case). Generate the Gröbner basis $\mathcal{GB}(F)$. Let $\mathcal{GB}' = \mathcal{GB}(F) \cap \mathbb{Q}[y]$ be the set of polynomials found in $\mathcal{GB}(F)$ containing only variables y. If $\mathcal{GB}' = \emptyset$, then $\{\theta'\}$ are algebraically independent; they may be chosen to comprise the HSOP of the group. If $\mathcal{GB}' \neq \emptyset$, then the functions contained in $\mathcal{GB}'$ represent polynomial equations $P(y_1, y_2, \dots, y_6) = 0$ which represent syzygies among the variables y hence amongst the functions θ'. This property will be exploited to determine syzygies.

The second task is the determination of secondary invariant polynomials [24, algorithm 2.5.14]. Begin with $F = \{\theta\}$, the set of primary invariants, and let $\phi = \emptyset$ be the set of discovered secondary invariants. Compute $\mathcal{GB} = \mathcal{GB}(F)$ with respect to any valid term order. For each degree indicated in the numerator of the Molien function, use the Reynolds operator to construct a set of linearly independent homogeneous invariants. Those candidate polynomials ϕ' that reduce to zero with $\mathcal{GB}$ have a polynomial dependence on $\{\theta\}$ and are not valid secondary invariants. Those ϕ' that do not reduce to zero are secondary invariants; $\phi := \phi \cup \{\phi'\}$.

To deduce rewriting syzygies, i.e., syzygies of the form $\phi_i \phi_j = p_0 + \sum_k p_k \phi_k$, $p_i \in \mathbb{Q}[\theta]$, another algorithm based on Gröbner bases has been proposed [24, algorithm 2.5.6]. One computes the Gröbner basis of $F = \{\theta_1 - y_1, \dots, \theta_6 - y_6, \phi_1 - z_1 \dots\}$ in the variables C and slack variables y and z. Sturmfels recommends the variable order $C_1 > \cdots > C_6 > y_1 > \cdots > y_6 > z_1 > \cdots$ and suggests the following term order $\succ$. Term $C^\alpha y^\beta z^\gamma \succ C^{\alpha'} y^{\beta'} z^{\gamma'}$ if $C^\alpha > C^{\alpha'}$ in the purely lexicographic order, or if $C^\alpha = C^{\alpha'}$ and $y^\beta > y^{\beta'}$ in the degree lexicographic order, or if $C^\alpha = C^{\alpha'}$ and $y^\beta = y^{\beta'}$ and $z^\gamma > z^{\gamma'}$ in the purely lexicographic order. The resulting Gröbner basis will contain the desired syzygies.

To compute syzygies relating one secondary invariant ϕ_i to the primary invariants, essentially the same procedure is employed. The Gröbner basis of $F = \{\theta_1 - y_1, \dots, \theta_6 - y_6, \phi_i - z\}$ is computed in the variables $C_1 > \cdots > C_6 > y_1 > \cdots > y_6 > z$ with an order that eliminates the variables C. Good success was found using a matrix order (e.g., [9]) that first selects for graded degree (using the degrees in C as weights), then selects for degree in the variables C, and then enforces reverse lexicographical ordering on the C and y blocks. As suggested by Bayer and Stillman [2], the reverse lexicographical refinement was substantially more efficient than purely lexicographical order.

The construction of Gröbner bases is given by a simple algorithm by Buchberger [5], but the simplicity of the algorithm belies the complexity of the computational

[1] Note that the Reynolds operator does depend on the "setting," or orientation, of the symmetry axes given in Table 3.1. To this point in the manuscript, only the setting-independent eigenvalues of the operations have been used.

task. The maximum degree computed in a Gröbner basis may be as large as doubly exponential in the number of variables used [13]; and integer or rational coefficients have been reported to contain as many as $\mathcal{O}(10^5)$ significant decimal figures with basis functions containing $\mathcal{O}(1)$ coefficients [1]. Thus, poor algorithmic choices (and there are many choices one is free to make) render even simple basis calculations impossible. To control these issues directly, implementations of Buchberger's algorithm and the F_4 variant of this algorithm by Faugére [7] were constructed in C++ using GMP [12] to represent and manipulate arbitrary precision integers. Superfluous pairs were eliminated using the method of Gebauer and Möller [10], and selection strategies used the "sugar" phantom degree order method of [11]. The F_4 algorithm has been reported to be on the order of 10 times faster than the equivalent Buchberger method. Our implementation of F_4 modifies the selection criterion as follows. Let deg_W be a W-graded degree, chosen so all polynomials are W-homogeneous (e.g, weights w_i correspond to the degree of a variable when expressed in the common basis of C elements). An F_4 row echelon calculation containing polynomials of different deg_W may be immediately block diagonalized according to deg_W. Including polynomials of different deg_W in a row echelon calculation does not affect the correctness of the method, but in practice it is found that selecting only those pairs whose deg_W are equal and as small as possible improves efficiency.

The results of these algorithms applied to the 11 Laue groups are presented below. The following subsections present the computed invariant bases, with elements distinguished as being primary or secondary invariants. In all cases the minimal factorizations displayed in Table 4.2 are realized. In the appendix, a complete set of rewriting syzygies is presented. Application of these equations may transform any polynomial $P(\{\theta\},\{\phi\})$ into the minimal form given by (4.5). These are offered in proof of the simplification implied by the Molien factorizations. Also presented in the appendix is a representative example of an algebraic dependence syzygy, a polynomial of the form $P(\phi_\alpha,\theta_1,\ldots,\theta_6)$ which demonstrates the algebraic dependence of the secondary invariants. Several such algebraic dependence syzygies also appear in the set of rewriting syzygies. Note that the computation of these algebraic dependence syzygies is difficult, and several such syzygies have thus far defied computation. With the algorithms used, the relevant Gröbner basis calculation may consume all available core memory (8Gb) in the span of a few days.

5.1. Triclinic groups C_1 and C_i. The group C_1 contains no symmetry operations aside from the identity E. The group C_i contains only the identity and a center of inversion. With respect to the action of these groups on the Cauchy tensor components, the groups are therefore identical. Since no Cauchy tensor components are mixed by the action of these groups, there are no nontrivial Reynolds projections. The basis for these groups consists of the Cauchy tensor components, all primary invariants.

$$\theta_1 = C_6, \tag{5.1}$$

$$\theta_2 = C_5, \tag{5.2}$$

$$\theta_3 = C_4, \tag{5.3}$$

$$\theta_4 = C_3, \tag{5.4}$$

$$\theta_5 = C_2, \tag{5.5}$$

$$\theta_6 = C_1. \tag{5.6}$$

5.2. Monoclinic groups C_s, C_2, C_{2h}. A single secondary invariant exists for this group. An invariant basis is

$$\theta_1 = C_4, \tag{5.7}$$
$$\theta_2 = C_3, \tag{5.8}$$
$$\theta_3 = C_2, \tag{5.9}$$
$$\theta_4 = C_1, \tag{5.10}$$
$$\theta_5 = C_6^2, \tag{5.11}$$
$$\theta_6 = C_5^2, \tag{5.12}$$
$$\phi_1 = C_5 C_6. \tag{5.13}$$

5.3. Orthorhombic groups C_{2v}, D_2, D_{2h}. A single secondary invariant of degree 3 exists:

$$\theta_1 = C_3, \tag{5.14}$$
$$\theta_2 = C_2, \tag{5.15}$$
$$\theta_3 = C_1, \tag{5.16}$$
$$\theta_4 = C_6^2, \tag{5.17}$$
$$\theta_5 = C_5^2, \tag{5.18}$$
$$\theta_6 = C_4^2, \tag{5.19}$$
$$\phi_1 = C_4 C_5 C_6. \tag{5.20}$$

5.4. Tetragonal groups S_4, C_4, C_{4h}. An invariant basis obeying the Molien factorization of Table 4.2 is

$$\theta_1 = C_3, \tag{5.21}$$
$$\theta_2 = C_1 + C_2, \tag{5.22}$$
$$\theta_3 = C_6^2, \tag{5.23}$$
$$\theta_4 = C_4^2 + C_5^2, \tag{5.24}$$
$$\theta_5 = C_1^2 + C_2^2, \tag{5.25}$$
$$\theta_6 = C_4^4 + C_5^4, \tag{5.26}$$
$$\phi_1 = (C_1 - C_2) C_6, \tag{5.27}$$
$$\phi_2 = (C_4^2 - C_5^2) C_6, \tag{5.28}$$
$$\phi_3 = C_4 C_5 C_6, \tag{5.29}$$
$$\phi_4 = C_1 C_4^2 + C_2 C_5^2, \tag{5.30}$$
$$\phi_5 = (C_1 - C_2) C_4 C_5, \tag{5.31}$$
$$\phi_6 = C_4 C_5 (C_4^2 - C_5^2), \tag{5.32}$$
$$\phi_7 = \phi_3 \phi_4. \tag{5.33}$$

5.5. Tetragonal groups D_{2d}, C_{4v}, D_4, D_{4h}. The invariant relations for these groups are also relatively simple:

$$\theta_1 = C_3, \tag{5.34}$$
$$\theta_2 = C_1 + C_2, \tag{5.35}$$
$$\theta_3 = C_6^2, \tag{5.36}$$

$$\theta_4 = C_4^2 + C_5^2, \tag{5.37}$$
$$\theta_5 = C_1^2 + C_2^2, \tag{5.38}$$
$$\theta_6 = C_4^4 + C_5^4, \tag{5.39}$$
$$\phi_1 = C_4 C_5 C_6, \tag{5.40}$$
$$\phi_2 = C_1 C_4^2 + C_2 C_5^2, \tag{5.41}$$
$$\phi_3 = \phi_1 \phi_2. \tag{5.42}$$

5.6. Cubic groups T, T_h and groups T_d, O, O_h. Group T_d is subset of group T; they share the same primary invariants and several secondary invariants. Those secondary invariants found in group T but absent from T_d are denoted by an asterisk:

$$\theta_1 = C_1 + C_2 + C_3, \tag{5.43}$$
$$\theta_2 = C_4^2 + C_5^2 + C_6^2, \tag{5.44}$$
$$\theta_3 = C_1^2 + C_2^2 + C_3^2, \tag{5.45}$$
$$\theta_4 = C_4 C_5 C_6, \tag{5.46}$$
$$\theta_5 = C_1^3 + C_2^3 + C_3^3, \tag{5.47}$$
$$\theta_6 = C_4^4 + C_5^4 + C_6^4, \tag{5.48}$$
$$\phi_1 = C_1 C_4^2 + C_2 C_5^2 + C_3 C_6^2, \tag{5.49}$$
$$\phi_2 = \phi_1^2, \tag{5.50}$$
$$\phi_3 = \phi_1^3, \tag{5.51}$$
$$(*) \quad \phi_4 = C_1 C_6^2 + C_2 C_4^2 + C_3 C_5^2, \tag{5.52}$$
$$(*) \quad \phi_5 = \phi_4^2, \tag{5.53}$$
$$(*) \quad \phi_6 = C_1^2 C_3 + C_1 C_2^2 + C_2 C_3^2, \tag{5.54}$$
$$\phi_7 = C_1^2 C_4^2 + C_2^2 C_5^2 + C_3^2 C_6^2, \tag{5.55}$$
$$(*) \quad \phi_8 = C_1^2 C_6^2 + C_2^2 C_4^2 + C_3^2 C_5^2, \tag{5.56}$$
$$\phi_9 = C_1 C_4^4 + C_2 C_5^4 + C_3 C_6^4, \tag{5.57}$$
$$(*) \quad \phi_{10} = C_1 C_4^2 C_6^2 + C_2 C_4^2 C_5^2 + C_3 C_5^2 C_6^2, \tag{5.58}$$
$$(*) \quad \phi_{11} = C_4^4 C_6^2 + C_4^2 C_5^4 + C_5^2 C_6^4. \tag{5.59}$$

It is interesting to note that the groups T and T_d share the same primary invariants. Consider (A.62), an algebraic dependence syzygy for ϕ_4, which occurs in T but not in T_d. The coefficients of ϕ_4^m, $m \in (0, 6)$, in (A.62) are expressed in terms of θ, and therefore the coefficients are invariant with respect to both T and T_d. However, the roots of this syzygy are not invariant. In T the roots ϕ of (A.62) describe an orbit of size 6 under the action of the reflection symmetry operations found in T but not in T_d.

5.7. Trigonal groups C_3, S_6. An invariant basis is

$$\theta_1 = C_3, \tag{5.60}$$
$$\theta_2 = C_1 + C_2, \tag{5.61}$$
$$\theta_3 = (C_1 - C_2)^2 + C_6^2, \tag{5.62}$$
$$\theta_4 = C_4^2 + C_5^2, \tag{5.63}$$
$$\theta_5 = C_6^3 - 3C_6(C_1 - C_2)^2, \tag{5.64}$$

$$\theta_6 = C_5(C_5^2 - 3C_4^2), \tag{5.65}$$
$$\phi_1 = C_4(C_1 - C_2) + C_5C_6, \tag{5.66}$$
$$\phi_2 = \phi_1^2, \tag{5.67}$$
$$\phi_3 = \phi_1^3, \tag{5.68}$$
$$\phi_4 = C_5(C_1 - C_2) - C_4C_6, \tag{5.69}$$
$$\phi_5 = C_5(C_1 - C_2)^2 + 2C_4C_6(C_1 - C_2) - C_5C_6^2, \tag{5.70}$$
$$\phi_6 = 2C_4C_5(C_1 - C_2) + C_6(C_4^2 - C_5^2), \tag{5.71}$$
$$\phi_7 = C_4(C_1 - C_2)^2 - 2C_5C_6(C_1 - C_2) - C_4C_6^2, \tag{5.72}$$
$$\phi_8 = (C_1 - C_2)(C_4^2 - C_5^2) - 2C_4C_5C_6, \tag{5.73}$$
$$\phi_9 = C_4(C_4^2 - 3C_5^2), \tag{5.74}$$
$$\phi_{10} = (C_1 - C_2)^2(3C_1 + C_2) - C_6^2(C_1 - 5C_2), \tag{5.75}$$
$$\phi_{11} = \phi_1\phi_4. \tag{5.76}$$

5.8. Trigonal groups C_{3v}, D_3, D_{3d}. An invariant basis is

$$\theta_1 = C_3, \tag{5.77}$$
$$\theta_2 = C_1 + C_2, \tag{5.78}$$
$$\theta_3 = (C_1 - C_2)^2 + C_6^2, \tag{5.79}$$
$$\theta_4 = C_4^2 + C_5^2, \tag{5.80}$$
$$\theta_5 = C_4(C_4^2 - 3C_5^2), \tag{5.81}$$
$$\theta_6 = (C_1 - C_2)^2(3C_1 + C_2) - C_6^2(C_1 - 5C_2), \tag{5.82}$$
$$\phi_1 = C_4(C_1 - C_2) + C_5C_6, \tag{5.83}$$
$$\phi_2 = \phi_1^2, \tag{5.84}$$
$$\phi_3 = \phi_1^3, \tag{5.85}$$
$$\phi_4 = C_4(C_1 - C_2)^2 - 2C_5C_6(C_1 - C_2) - C_4C_6^2, \tag{5.86}$$
$$\phi_5 = (C_1 - C_2)(C_4^2 - C_5^2) - 2C_4C_5C_6. \tag{5.87}$$

5.9. Hexagonal groups C_{3h}, C_6, C_{6h}. An invariant basis is

$$\theta_1 = C_3, \tag{5.88}$$
$$\theta_2 = C_1 + C_2, \tag{5.89}$$
$$\theta_3 = (C_1 - C_2)^2 + C_6^2, \tag{5.90}$$
$$\theta_4 = C_4^2 + C_5^2, \tag{5.91}$$
$$\theta_5 = C_6^3 - 3C_6(C_1 - C_2)^2, \tag{5.92}$$
$$\theta_6 = 9C_4^6 + 45C_4^4C_5^2 + 15C_4^2C_5^4 + 11C_5^6, \tag{5.93}$$
$$\phi_1 = 2C_4C_5(C_1 - C_2) + C_6(C_4^2 - C_5^2), \tag{5.94}$$
$$\phi_2 = \phi_1^2, \tag{5.95}$$
$$\phi_3 = \phi_1^3, \tag{5.96}$$
$$\phi_4 = (C_1 - C_2)(C_4^2 - C_5^2) - 2C_4C_5C_6, \tag{5.97}$$
$$\phi_5 = (C_1 - C_2)^2(3C_1 + C_2) - C_6^2(C_1 - 5C_2), \tag{5.98}$$
$$\phi_6 = (3C_4^2 + C_5^2)(C_1 - C_2)^2 + 4C_4C_5C_6(C_1 - C_2) + C_6^2(C_4^2 + 3C_5^2), \tag{5.99}$$
$$\phi_7 = (C_1C_5 - C_2C_5 - C_4C_6)(C_1C_4 - C_2C_4 + C_5C_6), \tag{5.100}$$

$$\phi_8 = 4C_4C_5(C_1 - C_2)(3C_4^2 + C_5^2) + C_6(3C_4^4 + 6C_4^2C_5^2 - 5C_5^4), \tag{5.101}$$

$$\phi_9 = 8C_4C_5^3C_6 - (C_1 - C_2)(3C_4^4 - 6C_4^2C_5^2 - C_5^4), \tag{5.102}$$

$$\begin{aligned}\phi_{10} = {} & 4C_4C_5^3(C_1 - C_2)^2 - 4C_4C_5^3C_6^2 - (C_1 - C_2) \\ & \times (3C_4^4C_6 - 6C_4^2C_5^2C_6 - C_5^4C_6),\end{aligned} \tag{5.103}$$

$$\phi_{11} = -C_4C_5(C_4^2 - 3C_5^2)(3C_4^2 - C_5^2). \tag{5.104}$$

5.10. Hexagonal groups D_{3h}, C_{6v}, D_6, D_{6h}. An invariant basis for these groups is

$$\theta_1 = C_3, \tag{5.105}$$

$$\theta_2 = C_1 + C_2, \tag{5.106}$$

$$\theta_3 = (C_1 - C_2)^2 + C_6^2, \tag{5.107}$$

$$\theta_4 = C_4^2 + C_5^2, \tag{5.108}$$

$$\theta_5 = (C_1 - C_2)^2(3C_1 + C_2) - C_6^2(C_1 - 5C_2), \tag{5.109}$$

$$\theta_6 = 9C_4^6 + 45C_4^4C_5^2 + 15C_4^2C_5^4 + 11C_5^6, \tag{5.110}$$

$$\phi_1 = (C_1 - C_2)(C_4^2 - C_5^2) - 2C_4C_5C_6, \tag{5.111}$$

$$\phi_2 = \phi_1^2, \tag{5.112}$$

$$\phi_3 = \phi_1^3, \tag{5.113}$$

$$\begin{aligned}\phi_4 = {} & (C_1 - C_2)^2(3C_4^2 + C_5^2) + 4C_4C_5C_6 \\ & \times (C_1 - C_2) + C_6^2(C_4^2 + 3C_5^2),\end{aligned} \tag{5.114}$$

$$\phi_5 = 8C_4C_5^3C_6 - (C_1 - C_2)(3C_4^4 - 6C_4^2C_5^2 - C_5^4). \tag{5.115}$$

6. Conclusions. The invariant bases presented above agree with those presented by Smith and Rivlin [20] and are identical in the sense that they generate the same ideal. In many cases the particular form of the invariants differs. This has no significance and is merely an artifact of the particular methods used. For example, in the group T the invariants K presented by Smith and Rivlin are related to the invariants θ and ϕ in (5.43)–(5.59) via $\theta_1 = K_1$, $\theta_2 = K_4$, $\theta_3 = K_1^2 - 2K_2$, $\theta_4 = K_0$, $\theta_5 = K_1^3 - 3K_1K_2 + 3K_3$, $\theta_6 = K_4^2 - 2K_5$, $\phi_1 = K_1K_4 - K_7 - K_8$, $\phi_4 = K_8$, $\phi_6 = K_1K_2 - K_9 - 3K_3$, $\phi_7 = K_1^2K_4 - K_1K_7 - K_1K_8 + K_{12} - K_2K_4$, $\phi_8 = K_1K_8 - K_2K_4 + K_{13}$, $\phi_9 = K_1K_4^2 - K_4K_8 - K_4K_7 - K_1K_5 + K_{11}$, $\phi_{10} = K_{14}$, and $\phi_{11} = K_4K_5 - K_{10} - 3K_6$ (with $K_0 = \sqrt{K_6} = C_4C_5C_6$). By writing the K's in terms of θ's and ϕ's, it is apparent on inspection that K_1, K_4, K_2, K_0, K_3, and K_5 form an HSOP and a set of primary invariants of minimal degree indicated in Table 4.2. Likewise, K_7, K_8, K_9, K_{12}, K_{13}, K_{11}, K_{14}, and K_{10} are valid secondary invariants. To make a complete set of secondary invariants, one could include both K_7^2 and K_8^2 of degree 6 and one of K_7^3 or K_8^3 of degree 9.

The following truncated series displays the difference between the Hilbert series implied by the Smith and Rivlin integrity bases and the invariants deduced above in their Molien form. The Smith and Rivlin results differ beginning with fourth-degree (in C) polynomials.

$$C_2: \quad z^4+4z^5+13z^6+32z^7+71z^8+140z^9+259z^{10}+448z^{11}+742z^{12}+\cdots, \tag{6.1}$$

$$C_{2v}: \quad z^6+3z^7+9z^8+20z^9+42z^{10}+78z^{11}+139z^{12}+\cdots, \tag{6.2}$$

$$S_4: \quad 2z^4+8z^5+32z^6+80z^7+194z^8+404z^9+808z^{10}+1488z^{11}+2663z^{12}+\cdots, \tag{6.3}$$

$$D_{2d}: \quad 2z^6+4z^7+12z^8+24z^9+50z^{10}+88z^{11}+157z^{12}+\cdots, \tag{6.4}$$

$$T: \quad 4z^6+10z^7+27z^8+63z^9+126z^{10}+239z^{11}+439z^{12}+\cdots, \tag{6.5}$$

$$T_d: \quad z^7+3z^8+6z^9+14z^{10}+26z^{11}+47z^{12}+\cdots, \tag{6.6}$$

$$C_3: \quad z^4+14z^5+53z^6+136z^7+341z^8+750z^9+1485z^{10}+2856z^{11}+5206z^{12}+\cdots, \tag{6.7}$$

$$C_{3v}: \quad 2z^5+7z^6+18z^7+43z^8+90z^9+170z^{10}+308z^{11}+528z^{12}+\cdots, \tag{6.8}$$

$$C_{3h}: \quad 4z^6+14z^7+41z^8+100z^9+212z^{10}+414z^{11}+767z^{12}+\cdots, \tag{6.9}$$

$$D_{3h}: \quad z^7+4z^8+10z^9+23z^{10}+45z^{11}+83z^{12}+\cdots. \tag{6.10}$$

Appendix. Syzygies.

The invariant bases for the triclinic groups C_1 and C_i contain no secondary invariants and hence no syzygies.

A.1. Monoclinic groups C_s, C_2, C_{2h}. This syzygy is a rewriting expression and also displays the algebraic dependence of the secondary invariant upon the HSOP:

$$\phi_1^2 = \theta_5\theta_6. \tag{A.1}$$

A.2. Orthorhombic groups C_{2v}, D_2, D_{2h}. An obvious syzygy exists:

$$\phi_1^2 = \theta_4\theta_5\theta_6. \tag{A.2}$$

This is a rewriting expression and displays the algebraic dependence.

A.3. Tetragonal groups S_4, C_4, C_{4h}. Again, the algebraic dependence syzygies are included in the set of rewriting syzygies. Rewriting syzygies for $\phi_7\phi_\alpha$ are omitted since they may be simply constructed by rewriting $\phi_3(\phi_4\phi_\alpha)$ or $\phi_4(\phi_3\phi_\alpha)$.

$$\phi_1^2 = -\theta_3\left[\theta_2^2 - 2\theta_5\right], \tag{A.3}$$

$$\phi_1\phi_2 = -\theta_2\theta_3\theta_4 + 2\theta_3\phi_4, \tag{A.4}$$

$$\phi_1\phi_3 = \theta_3\phi_5, \tag{A.5}$$

$$\phi_1\phi_4 = \frac{1}{2}\theta_2\theta_4\phi_1 - \frac{1}{2}\left[\theta_2^2 - 2\theta_5\right]\phi_2, \tag{A.6}$$

$$\phi_1\phi_5 = -\left[\theta_2^2 - 2\theta_5\right]\phi_3, \tag{A.7}$$

$$\phi_1\phi_6 = -\theta_2\theta_4\phi_3 + 2\phi_7, \tag{A.8}$$

$$\phi_2^2 = -\theta_3\left[\theta_4^2 - 2\theta_6\right], \tag{A.9}$$

$$\phi_2\phi_3 = \theta_3\phi_6, \tag{A.10}$$

$$\phi_2\phi_4 = -\frac{1}{2}\left[\theta_4^2 - 2\theta_6\right]\phi_1 + \frac{1}{2}\theta_2\theta_4\phi_2, \tag{A.11}$$

$$\begin{aligned}\phi_2\phi_5 &= \phi_1\phi_6\\ &= -\theta_2\theta_4\phi_3 + 2\phi_7,\end{aligned} \tag{A.12}$$

$$\phi_2\phi_6 = -\left[\theta_4^2 - 2\theta_6\right]\phi_3, \tag{A.13}$$

$$\phi_3^2 = \frac{1}{2}\theta_3\left[\theta_4^2 - \theta_6\right], \tag{A.14}$$

$$\phi_3\phi_4 = \phi_7, \tag{A.15}$$

$$\phi_3\phi_5 = \frac{1}{2}\left[\theta_4^2 - \theta_6\right]\phi_1, \tag{A.16}$$

$$\phi_3\phi_6 = \frac{1}{2}\left[\theta_4^2 - \theta_6\right]\phi_2, \tag{A.17}$$

$$\phi_4^2 = -\frac{1}{2}\left[\theta_2^2\theta_6 + \theta_4^2\theta_5 - 2\theta_5\theta_6\right] + \theta_2\theta_4\phi_4, \tag{A.18}$$

$$\phi_4\phi_5 = \frac{1}{2}\theta_2\theta_4\phi_5 - \frac{1}{2}\left[\theta_2^2 - 2\theta_5\right]\phi_6, \tag{A.19}$$

$$\phi_4\phi_6 = -\frac{1}{2}\left[\theta_4^2 - 2\theta_6\right]\phi_5 + \frac{1}{2}\theta_2\theta_4\phi_6, \tag{A.20}$$

$$\phi_5^2 = -\frac{1}{2}\left[\theta_2^2\theta_4^2 - \theta_2^2\theta_6 - 2\theta_4^2\theta_5 + 2\theta_5\theta_6\right], \tag{A.21}$$

$$\phi_5\phi_6 = -\frac{1}{2}\theta_2\theta_4\left[\theta_4^2 - \theta_6\right] + \left[\theta_4^2 - \theta_6\right]\phi_4, \tag{A.22}$$

$$\phi_6^2 = -\frac{1}{2}\left[\theta_4^4 - 3\theta_4^2\theta_6 + 2\theta_6^2\right]. \tag{A.23}$$

A.4. Tetragonal groups D_{2d}, C_{4v}, D_4, D_{4h}. Algebraic dependence syzygies coincide with the rewriting syzygies for these groups:

$$\phi_1^2 = \frac{1}{2}\theta_3\left[\theta_4^2 - \theta_6\right], \tag{A.24}$$

$$\phi_2^2 = -\frac{1}{2}\left[\theta_2^2\theta_6 + \theta_4^2\theta_5 - 2\theta_5\theta_6\right] + \theta_2\theta_4\phi_2. \tag{A.25}$$

A.5. Cubic groups T, T_h and groups T_d, O, O_h. The rewriting syzygies for this group are complicated and do not contain all algebraic dependence syzygies. Note that the rewriting syzygies for invariants of group T_d are expressed in terms of primaries θ and only those secondary invariants of group T_d.

$$\begin{aligned}\phi_1^4 = {}& \frac{1}{36}\left[2\theta_1^4\theta_2^4 - 7\theta_1^4\theta_2^2\theta_6 - 12\theta_1^4\theta_2\theta_4^2 + 3\theta_1^4\theta_6^2 - 10\theta_1^2\theta_2^4\theta_3\right.\\ &+ 28\theta_1^2\theta_2^2\theta_3\theta_6 + 54\theta_1^2\theta_2\theta_3\theta_4^2 - 12\theta_1^2\theta_3\theta_6^2 + 14\theta_1\theta_2^4\theta_5\\ &- 36\theta_1\theta_2^2\theta_5\theta_6 - 108\theta_1\theta_2\theta_4^2\theta_5 + 18\theta_1\theta_5\theta_6^2 - 6\theta_2^4\theta_3^2 + 15\theta_2^2\theta_3^2\theta_6\\ &\left.+ 54\theta_2\theta_3^2\theta_4^2 - 9\theta_3^2\theta_6^2\right] - \frac{1}{18}\left[2\theta_1^3\theta_2^3 - 12\theta_1^3\theta_2\theta_6 - 18\theta_1^3\theta_4^2\right.\\ &- 21\theta_1\theta_2^3\theta_3 + 51\theta_1\theta_2\theta_3\theta_6 + 108\theta_1\theta_3\theta_4^2 + 21\theta_2^3\theta_5 - 45\theta_2\theta_5\theta_6\\ &\left.- 162\theta_4^2\theta_5\right]\phi_1 - \frac{1}{36}\left[19\theta_1^2\theta_2^2 + 15\theta_1^2\theta_6 + 15\theta_2^2\theta_3 - 45\theta_3\theta_6\right]\phi_2\\ &+ \frac{4}{3}\theta_1\theta_2\phi_3 - \frac{1}{36}\left[\theta_1^2 - 3\theta_3\right]\left[5\theta_2^3 - 9\theta_2\theta_6 - 54\theta_4^2\right]\phi_7\\ &+ \frac{1}{18}\left[2\theta_1^3 - 9\theta_1\theta_3 + 9\theta_5\right]\left[\theta_2^2 - 3\theta_6\right]\phi_9,\end{aligned} \tag{A.26}$$

$$\phi_1\phi_4 = -\frac{1}{4}\left[\theta_1^2\theta_2^2 + \theta_1^2\theta_6 + \theta_2^2\theta_3 - 3\theta_3\theta_6\right] + \theta_1\theta_2\phi_1 - \phi_2 + \theta_1\theta_2\phi_4 - \phi_5, \tag{A.27}$$

$$\begin{aligned}\phi_1\phi_6 = {}& -\frac{1}{6}\theta_2\left[\theta_1^2\theta_3 + 3\theta_3^2 - 4\theta_1\theta_5\right] + \frac{1}{3}\left[2\theta_1\theta_3 - 3\theta_5\right]\phi_1 + \frac{1}{3}\left[\theta_1\theta_3 - 3\theta_5\right]\phi_4\\ &+ \frac{1}{3}\theta_1\theta_2\phi_6 - \frac{1}{6}\left[\theta_1^2 - 3\theta_3\right]\phi_7 - \frac{1}{3}\left[\theta_1^2 - 3\theta_3\right]\phi_8,\end{aligned} \tag{A.28}$$

$$\begin{aligned}\phi_1\phi_7 = {}& \frac{1}{6}\left[\theta_1^3\theta_6 + \theta_1\theta_2^2\theta_3 - 4\theta_1\theta_3\theta_6 - \theta_2^2\theta_5 + 3\theta_5\theta_6\right] - \frac{1}{3}\theta_1^2\theta_2\phi_1\\ &+ \frac{2}{3}\theta_1\phi_2 + \frac{1}{3}\theta_1\theta_2\phi_7 - \frac{1}{6}\left[\theta_1^2 - 3\theta_3\right]\phi_9,\end{aligned} \tag{A.29}$$

$$\begin{aligned}\phi_1\phi_8 = {}& -\frac{1}{12}\theta_1\left[\theta_1^2\theta_2^2 + 3\theta_1^2\theta_6 + 3\theta_2^2\theta_3 - 7\theta_3\theta_6\right] + \frac{1}{6}\theta_2\left[3\theta_1^2 + \theta_3\right]\phi_1\\ &- \frac{2}{3}\theta_1\phi_2 + \frac{1}{6}\theta_2\left[3\theta_1^2 - \theta_3\right]\phi_4 - \frac{2}{3}\theta_1\phi_5 - \frac{1}{6}\left[\theta_2^2 - 3\theta_6\right]\phi_6\\ &+ \frac{1}{3}\theta_1\theta_2\phi_8 - \frac{1}{6}\left[\theta_1^2 - 3\theta_3\right]\phi_{10},\end{aligned} \tag{A.30}$$

$$\begin{aligned}\phi_1\phi_9 &= \frac{1}{12}\big[\theta_1^2\theta_2^3 - \theta_1^2\theta_2\theta_6 - 6\theta_1^2\theta_4^2 - \theta_2^3\theta_3 + \theta_2\theta_3\theta_6 + 18\theta_3\theta_4^2\big]\\ &\quad - \frac{1}{3}\theta_1\theta_2^2\phi_1 + \frac{2}{3}\theta_2\phi_2 - \frac{1}{6}\left[\theta_2^2 - 3\theta_6\right]\phi_7 + \frac{1}{3}\theta_1\theta_2\phi_9, \qquad \text{(A.31)}\\ \phi_1\phi_{10} &= -\frac{1}{12}\theta_2\big[\theta_1^2 + \theta_3\big]\big[\theta_2^2 - \theta_6\big] - \frac{1}{12}\theta_2\big[\theta_1^2\theta_2^2 - \theta_1^2\theta_6 + \theta_2^2\theta_3 - \theta_3\theta_6\big]\\ &\quad + \frac{1}{6}\theta_1\big[\theta_2^2 - \theta_6\big]\phi_1 + \frac{1}{3}\theta_1\big[\theta_2^2 - \theta_6\big]\phi_4 - \frac{1}{3}\theta_2\phi_5\\ &\quad - \frac{1}{6}\big[\theta_2^2 - 3\theta_6\big]\phi_8 + \frac{1}{3}\theta_1\theta_2\phi_{10} - \frac{1}{6}\big[\theta_1^2 - 3\theta_3\big]\phi_{11}, \qquad \text{(A.32)}\\ \phi_1\phi_{11} &= -\theta_1\theta_2\theta_4^2 + \frac{1}{3}\theta_2\big[\theta_2^2 - 2\theta_6\big]\phi_1 - \frac{1}{6}\big[\theta_2^3 - \theta_2\theta_6 - 18\theta_4^2\big]\phi_4\\ &\quad - \frac{1}{6}\big[\theta_2^2 - 3\theta_6\big]\phi_9 - \frac{1}{3}\big[\theta_2^2 - 3\theta_6\big]\phi_{10} + \frac{1}{3}\theta_1\theta_2\phi_{11}, \qquad \text{(A.33)}\\ \phi_4^3 &= \frac{1}{4}\big[\theta_1^2\theta_2^2 + \theta_1^2\theta_6 + \theta_2^2\theta_3 - 3\theta_3\theta_6\big]\phi_1 - \theta_1\theta_2\phi_2 + \phi_3\\ &\quad - \frac{1}{4}\big[\theta_1^2\theta_2^2 + \theta_1^2\theta_6 + \theta_2^2\theta_3 - 3\theta_3\theta_6\big]\phi_4 + \theta_1\theta_2\phi_5, \qquad \text{(A.34)}\\ \phi_4\phi_6 &= -\frac{1}{6}\theta_2\big[\theta_1^2\theta_3 - 3\theta_3^2 + 2\theta_1\theta_5\big] - \frac{1}{3}\big[\theta_1\theta_3 - 3\theta_5\big]\phi_1 + \frac{1}{3}\theta_1\theta_3\phi_4\\ &\quad + \frac{1}{3}\theta_1\theta_2\phi_6 + \frac{1}{3}\big[\theta_1^2 - 3\theta_3\big]\phi_7 + \frac{1}{6}\big[\theta_1^2 - 3\theta_3\big]\phi_8, \qquad \text{(A.35)}\\ \phi_4\phi_7 &= -\frac{1}{12}\big[\theta_1^3\theta_2^2 + 3\theta_1^3\theta_6 + 5\theta_1\theta_2^2\theta_3 - 13\theta_1\theta_3\theta_6 - 2\theta_2^2\theta_5 + 6\theta_5\theta_6\big]\\ &\quad + \frac{1}{6}\theta_2\big[3\theta_1^2 - \theta_3\big]\phi_1 - \frac{2}{3}\theta_1\phi_2 + \frac{1}{6}\theta_2\big[3\theta_1^2 + \theta_3\big]\phi_4 - \frac{2}{3}\theta_1\phi_5\\ &\quad + \frac{1}{6}\big[\theta_2^2 - 3\theta_6\big]\phi_6 + \frac{1}{3}\theta_1\theta_2\phi_7 - \frac{1}{6}\big[\theta_1^2 - 3\theta_3\big]\phi_{10}, \qquad \text{(A.36)}\\ \phi_4\phi_8 &= \frac{1}{12}\big[\theta_1^3\theta_2^2 + \theta_1^3\theta_6 - \theta_1\theta_2^2\theta_3 - 5\theta_1\theta_3\theta_6 - 2\theta_2^2\theta_5 + 6\theta_5\theta_6\big]\\ &\quad - \frac{1}{6}\theta_2\big[\theta_1^2 - 3\theta_3\big]\phi_1 - \frac{1}{2}\theta_2\big[\theta_1^2 - \theta_3\big]\phi_4 + \frac{2}{3}\theta_1\phi_5\\ &\quad + \frac{1}{3}\theta_1\theta_2\phi_8 + \frac{1}{6}\big[\theta_1^2 - 3\theta_3\big]\phi_9 + \frac{1}{6}\big[\theta_1^2 - 3\theta_3\big]\phi_{10}, \qquad \text{(A.37)}\\ \phi_4\phi_9 &= -\frac{1}{6}\theta_2\big[3\theta_1^2\theta_6 + 2\theta_2^2\theta_3 - 5\theta_3\theta_6\big] + \frac{1}{3}\theta_1\big[\theta_2^2 + \theta_6\big]\phi_1 - \frac{2}{3}\theta_2\phi_2\\ &\quad + \frac{1}{3}\theta_1\big[\theta_2^2 + 2\theta_6\big]\phi_4 - \frac{2}{3}\theta_2\phi_5 + \frac{1}{6}\big[\theta_2^2 - 3\theta_6\big]\phi_7\\ &\quad + \frac{1}{6}\big[\theta_2^2 - 3\theta_6\big]\phi_8 + \frac{1}{3}\theta_1\theta_2\phi_9 - \frac{1}{6}\big[\theta_1^2 - 3\theta_3\big]\phi_{11}, \qquad \text{(A.38)}\\ \phi_4\phi_{10} &= -\frac{1}{6}\big[\theta_1^2\theta_2^3 - 3\theta_1^2\theta_4^2 - \theta_1^2\theta_2\theta_6 - \theta_2^3\theta_3 + \theta_2\theta_3\theta_6 + 9\theta_3\theta_4^2\big]\\ &\quad + \frac{1}{3}\theta_1\big[\theta_2^2 - \theta_6\big]\phi_1 - \frac{1}{3}\theta_2\phi_2 + \frac{1}{6}\theta_1\big[\theta_2^2 - \theta_6\big]\phi_4 - \frac{1}{6}\big[\theta_2^2 - 3\theta_6\big]\phi_7\\ &\quad + \frac{1}{3}\theta_1\theta_2\phi_{10} + \frac{1}{6}\big[\theta_1^2 - 3\theta_3\big]\phi_{11}, \qquad \text{(A.39)}\\ \phi_4\phi_{11} &= -\frac{1}{4}\theta_1\big[\theta_2^4 - 2\theta_2^2\theta_6 - 8\theta_2\theta_4^2 + \theta_6^2\big] + \frac{1}{3}\big[\theta_2^3 - 2\theta_2\theta_6 - 9\theta_4^2\big]\phi_1\\ &\quad + \frac{1}{3}\big[\theta_2^3 - \theta_2\theta_6 - 9\theta_4^2\big]\phi_4 - \frac{1}{6}\big[\theta_2^2 - 3\theta_6\big]\phi_9\\ &\quad + \frac{1}{6}\big[\theta_2^2 - 3\theta_6\ \big]\phi_{10} + \frac{1}{3}\theta_1\theta_2\phi_{11}, \qquad \text{(A.40)}\\ \phi_6^2 &= -\frac{1}{24}\big[\theta_1^6 - 9\theta_1^4\theta_3 + 8\theta_1^3\theta_5 + 27\theta_1^2\theta_3^2 - 48\theta_1\theta_3\theta_5 - 3\theta_3^3 + 24\theta_5^2\big]\\ &\quad + \left[\theta_1\theta_3 - \theta_5\right]\phi_6, \qquad \text{(A.41)}\end{aligned}$$

$$\begin{aligned}\phi_6\phi_7 = {}& \frac{1}{12}\theta_2\big[\theta_1^5 - 6\theta_1^3\theta_3 + 8\theta_1^2\theta_5 + \theta_1\theta_3^2 - 4\theta_3\theta_5\big] - \frac{1}{12}\big[\theta_1^4 - 6\theta_1^2\theta_3 + \theta_3^2 + 8\theta_1\theta_5\big]\phi_1\\ & - \frac{1}{6}\big[\theta_1^4 - 6\theta_1^2\theta_3 + 8\theta_1\theta_5 + \theta_3^2\big]\phi_4 + \frac{1}{3}\theta_2\theta_3\phi_6\\ & + \frac{1}{3}\theta_1\theta_3\phi_7 - \frac{1}{3}\big[\theta_1\theta_3 - 3\theta_5\big]\phi_8,\end{aligned} \tag{A.42}$$

$$\begin{aligned}\phi_6\phi_8 = {}& -\frac{1}{12}\theta_2\big[\theta_1^2 - \theta_3\big]\big[\theta_1^3 - 5\theta_1\theta_3 + 8\theta_5\big] + \frac{1}{6}\big[\theta_1^4 + \theta_3^2 - 6\theta_1^2\theta_3 + 8\theta_1\theta_5\big]\phi_1\\ & + \frac{1}{12}\big[\theta_1^4 - 6\theta_1^2\theta_3 + 8\theta_1\theta_5 + \theta_3^2\big]\phi_4 + \frac{1}{3}\theta_2\theta_3\phi_6\\ & + \frac{1}{3}\big[\theta_1\theta_3 - 3\theta_5\big]\phi_7 + \frac{1}{3}\big[2\theta_1\theta_3 - 3\theta_5\big]\phi_8,\end{aligned} \tag{A.43}$$

$$\begin{aligned}\phi_6\phi_9 = {}& \frac{1}{36}\big[\theta_1^4\theta_2^2 - \theta_1^4\theta_6 - 6\theta_1^2\theta_2^2\theta_3 + 18\theta_1\theta_2^2\theta_5 + 6\theta_1\theta_5\theta_6 - 9\theta_2^2\theta_3^2 - 9\theta_3^2\theta_6\big]\\ & - \frac{1}{9}\theta_2\big[\theta_1^3 - 6\theta_1\theta_3 + 9\theta_5\big]\phi_1 - \frac{1}{18}\big[\theta_1^2 - 3\theta_3\big]\phi_2\\ & + \frac{1}{3}\theta_2\big[\theta_1\theta_3 - 3\theta_5\big]\phi_4 - \frac{1}{9}\big[\theta_1^2 - 3\theta_3\big]\phi_5 + \frac{1}{3}\theta_1\theta_6\phi_6\\ & - \frac{1}{9}\theta_2\big[\theta_1^2 - 3\theta_3\big]\phi_7 - \frac{2}{9}\theta_2\big[\theta_1^2 - 3\theta_3\big]\phi_8 + \frac{1}{9}\theta_1^3\phi_9\\ & + \frac{1}{9}\big[2\theta_1^3 - 9\theta_1\theta_3 + 9\theta_5\big]\phi_{10},\end{aligned} \tag{A.44}$$

$$\begin{aligned}\phi_6\phi_{10} = {}& -\frac{1}{18}\theta_1\big[\theta_1^3 - 3\theta_1\theta_3 + 3\theta_5\big]\big[\theta_2^2 - \theta_6\big] + \frac{1}{9}\theta_2\big[2\theta_1^3 - 9\theta_1\theta_3 + 9\theta_5\big]\phi_1\\ & - \frac{1}{18}\big[\theta_1^2 - 3\theta_3\big]\phi_2 + \frac{1}{18}\big[\theta_1^2 - 3\theta_3\big]\phi_5 + \frac{1}{6}\theta_1\big[\theta_2^2 - \theta_6\big]\phi_6\\ & + \frac{1}{18}\theta_2\big[\theta_1^2 - 3\theta_3\big]\phi_7 - \frac{1}{18}\theta_2\big[\theta_1^2 - 3\theta_3\big]\phi_8\\ & - \frac{1}{9}\big[2\theta_1^3 - 9\theta_1\theta_3 + 9\theta_5\big]\phi_9 - \frac{1}{9}\big[\theta_1^3 - 9\theta_1\theta_3 + 9\theta_5\big]\phi_{10},\end{aligned} \tag{A.45}$$

$$\begin{aligned}\phi_6\phi_{11} = {}& \frac{1}{24}\big[\theta_1^3\theta_2^3 - 3\theta_1^3\theta_2\theta_6 - 12\theta_1^3\theta_4^2 - 9\theta_1\theta_2^3\theta_3 + 15\theta_1\theta_2\theta_3\theta_6 + 72\theta_1\theta_3\theta_4^2 + 8\theta_2^3\theta_5\\ & \qquad - 12\theta_2\theta_5\theta_6 - 72\theta_4^2\theta_5\big] + \frac{1}{8}\big[\theta_1^2\theta_2^2 + \theta_1^2\theta_6 + \theta_2^2\theta_3 - 3\theta_3\theta_6\big]\phi_1\\ & - \frac{1}{2}\theta_1\theta_2\phi_2 + \frac{1}{2}\phi_3 + \frac{1}{4}\big[\theta_2^3 - \theta_2\theta_6 - 6\theta_4^2\big]\phi_6\\ & + \frac{1}{2}\big[\theta_1\theta_3 - \theta_5\big]\phi_{11},\end{aligned} \tag{A.46}$$

$$\begin{aligned}\phi_7^2 = {}& \frac{1}{6}\left[\theta_1^4\theta_6 + \theta_1^2\theta_2^2\theta_3 - 4\theta_1^2\theta_3\theta_6 + 2\theta_1\theta_5\theta_6 - \theta_2^2\theta_3^2 + \theta_3^2\theta_6\right]\\ & - \frac{1}{3}\theta_2\left[\theta_1^3 - \theta_1\theta_3 + 2\theta_5\right]\phi_1 + \frac{1}{6}\left[3\theta_1^2 - \theta_3\right]\phi_2 + \frac{2}{3}\theta_2\theta_3\phi_7\\ & - \frac{1}{3}\big[\theta_1\theta_3 - 3\theta_5\big]\phi_9,\end{aligned} \tag{A.47}$$

$$\begin{aligned}\phi_7\phi_8 = {}& -\frac{1}{24}\big[\theta_1^4\theta_2^2 + 5\theta_1^4\theta_6 + 6\theta_1^2\theta_2^2\theta_3 - 18\theta_1^2\theta_3\theta_6 - 4\theta_1\theta_2^2\theta_5\\ & + 4\theta_1\theta_5\theta_6 + \theta_2^2\theta_3^2 + 5\theta_3^2\theta_6\big] + \frac{1}{3}\theta_2\big[\theta_1^3 - \theta_5\big]\phi_1 - \frac{1}{6}\big[3\theta_1^2 - \theta_3\big]\phi_2\\ & + \frac{1}{3}\theta_2\big[\theta_1^3 - \theta_5\big]\phi_4 - \frac{1}{6}\big[3\theta_1^2 - \theta_3\big]\phi_5 + \frac{1}{3}\theta_2\theta_3\phi_7\\ & + \frac{1}{3}\theta_2\theta_3\phi_8 - \frac{1}{3}\big[\theta_1\theta_3 - 3\theta_5\big]\phi_{10},\end{aligned} \tag{A.48}$$

$$\begin{aligned}\phi_7\phi_9 =& -\frac{1}{18}\big[\theta_1^3\theta_2^3 - 5\theta_1^3\theta_2\theta_6 - 6\theta_1^3\theta_4^2 - 8\theta_1\theta_2^3\theta_3 + 20\theta_1\theta_2\theta_3\theta_6 + 36\theta_1\theta_3\theta_4^2 - 15\theta_2\theta_5\theta_6\\ &+ 7\theta_2^3\theta_5 - 54\theta_4^2\theta_5\big] - \frac{1}{9}\big[2\theta_1^2\theta_2^2 + 3\theta_1^2\theta_6 + 3\theta_2^2\theta_3 - 3\theta_3\theta_6\big]\phi_1\\ &+ \frac{7}{9}\theta_1\theta_2\phi_2 - \frac{1}{3}\phi_3 - \frac{1}{9}\theta_1\left[\theta_2^2 - 6\theta_6\right]\phi_7 - \frac{1}{9}\theta_2\left[\theta_1^2 - 6\theta_3\right]\phi_9, \qquad \text{(A.49)}\\ \phi_7\phi_{10} =& -\frac{1}{72}\big[3\theta_1^3\theta_2^3 - \theta_1^3\theta_2\theta_6 + 12\theta_1^3\theta_4^2 + \theta_1\theta_2^3\theta_3 - 7\theta_1\theta_2\theta_3\theta_6 + 4\theta_2^3\theta_5\big]\\ &+ \frac{1}{72}\big[7\theta_1^2\theta_2^2 - 9\theta_2^2\theta_3 + 3\theta_1^2\theta_6 - 9\theta_3\theta_6\big]\phi_1 - \frac{1}{6}\theta_1\theta_2\phi_2 + \frac{1}{6}\phi_3\\ &+ \frac{1}{12}\big[3\theta_1^2 - \theta_3\big]\big[\theta_2^2 - \theta_6\big]\phi_4 - \frac{2}{9}\theta_1\theta_2\phi_5 - \frac{1}{12}\big[\theta_2^3 - \theta_2\theta_6 - 18\theta_4^2\big]\phi_6\\ &+ \frac{1}{6}\theta_1\big[\theta_2^2 - \theta_6\big]\phi_7 - \frac{1}{9}\theta_1\big[\theta_2^2 - 3\theta_6\big]\phi_8 - \frac{1}{18}\theta_2\big[\theta_1^2 - 3\theta_3\big]\phi_9\\ &- \frac{1}{18}\theta_2\big[\theta_1^2 - 9\theta_3\big]\phi_{10} - \frac{1}{6}\big[\theta_1\theta_3 - 3\theta_5\big]\phi_{11}, \qquad \text{(A.50)}\\ \phi_7\phi_{11} =& -\frac{1}{36}\big[\theta_1^2\theta_2^4 - 4\theta_1^2\theta_2^2\theta_6 + 3\theta_1^2\theta_6^2 - 3\theta_2^4\theta_3 + 12\theta_2^2\theta_3\theta_6 + 36\theta_2\theta_3\theta_4^2 - 9\theta_3\theta_6^2\big]\\ &+ \frac{1}{9}\theta_1\theta_2\big[\theta_2^2 - 3\theta_6\big]\phi_1 + \frac{1}{18}\big[\theta_2^2 - 3\theta_6\big]\phi_2 + \frac{1}{9}\big[\theta_2^2 - 3\theta_6\big]\phi_5\\ &+ \frac{1}{9}\theta_2^3\phi_7 - \frac{1}{18}\big[5\theta_2^3 - 9\theta_2\theta_6 - 54\theta_4^2\big]\phi_8\\ &- \frac{1}{9}\theta_1\big[\theta_2^2 - 3\theta_6\big]\phi_9 - \frac{2}{9}\theta_1\big[\theta_2^2 - 3\theta_6\big]\phi_{10} + \frac{1}{3}\theta_2\theta_3\phi_{11}, \qquad \text{(A.51)}\\ \phi_8^2 =& \frac{1}{6}\big[\theta_1^4\theta_6 + 2\theta_1^2\theta_2^2\theta_3 - 5\theta_1^2\theta_3\theta_6 - 3\theta_1\theta_2^2\theta_5 + 5\theta_1\theta_5\theta_6 - \theta_2^2\theta_3^2 + \theta_3^2\theta_6\big]\\ &- \frac{1}{3}\theta_2\big[\theta_1\theta_3 - 3\theta_5\big]\phi_1 - \frac{1}{3}\theta_2\big[\theta_1^3 - \theta_5\big]\phi_4 + \frac{1}{6}\big[3\theta_1^2 - \theta_3\big]\phi_5\\ &+ \frac{2}{3}\theta_2\theta_3\phi_8 + \frac{1}{3}\big[\theta_1\theta_3 - 3\theta_5\big]\phi_9 + \frac{1}{3}\big[\theta_1\theta_3 - 3\theta_5\big]\phi_{10}, \qquad \text{(A.52)}\\ \phi_8\phi_9 =& \frac{1}{72}\big[\theta_1^3\theta_2^3 - 12\theta_1^3\theta_4^2 - 27\theta_1^3\theta_2\theta_6 - 13\theta_1\theta_2^3\theta_3 + 43\theta_1\theta_2\theta_3\theta_6 - 4\theta_2^3\theta_5\big]\\ &+ \frac{1}{24}\big[7\theta_1^2\theta_2^2 - \theta_2^2\theta_3 + 7\theta_1^2\theta_6 - 5\theta_3\theta_6\big]\phi_1 - \frac{11}{18}\theta_1\theta_2\phi_2\\ &+ \frac{1}{6}\phi_3 + \frac{1}{12}\big[3\theta_1^2 - \theta_3\big]\big[\theta_2^2 + \theta_6\big]\phi_4 - \frac{4}{9}\theta_1\theta_2\phi_5\\ &- \frac{1}{12}\big[3\theta_2^3 - 7\theta_2\theta_6 - 18\theta_4^2\big]\phi_6 + \frac{1}{9}\theta_1\big[\theta_2^2 - 3\theta_6\big]\phi_7 + \frac{1}{9}\theta_1\theta_2^2\phi_8\\ &+ \frac{1}{3}\theta_2\theta_3\phi_9 - \frac{1}{9}\theta_2\big[\theta_1^2 - 3\theta_3\big]\phi_{10} - \frac{1}{6}\big[\theta_1\theta_3 - 3\theta_5\big]\phi_{11}, \qquad \text{(A.53)}\\ \phi_8\phi_{10} =& -\frac{1}{72}\big[\theta_1^3\theta_2^3 + \theta_1^3\theta_2\theta_6 + 12\theta_1^3\theta_4^2 + 19\theta_1\theta_2^3\theta_3 - 25\theta_1\theta_2\theta_3\theta_6\\ &- 144\theta_1\theta_3\theta_4^2 - 20\theta_2^3\theta_5 + 24\theta_2\theta_5\theta_6 + 216\theta_4^2\theta_5\big]\\ &+ \frac{1}{72}\big[17\theta_1^2\theta_2^2 - 15\theta_1^2\theta_6 + 9\theta_2^2\theta_3 - 3\theta_3\theta_6\big]\phi_1 - \frac{7}{18}\theta_1\theta_2\phi_2\\ &+ \frac{1}{6}\phi_3 + \frac{1}{12}\big[\theta_2^3 - \theta_2\theta_6 - 18\theta_4^2\big]\phi_6 - \frac{1}{9}\theta_1\big[\theta_2^2 - 3\theta_6\big]\phi_7\\ &+ \frac{1}{6}\theta_1\big[\theta_2^2 - \theta_6\big]\phi_8 + \frac{1}{18}\theta_2\big[\theta_1^2 - 3\theta_3\big]\phi_9 + \frac{1}{3}\theta_2\theta_3\phi_{10}\\ &+ \frac{1}{6}\big[\theta_1\theta_3 - 3\theta_5\big]\phi_{11}, \qquad \text{(A.54)}\end{aligned}$$

$$\begin{aligned}\phi_8\phi_{11} =& -\frac{1}{36}\big[\theta_1^2\theta_2^4 - 4\theta_1^2\theta_2^2\theta_6 + 3\theta_1^2\theta_6^2 + 9\theta_2^4\theta_3 - 18\theta_2^2\theta_3\theta_6 - 72\theta_2\theta_3\theta_4^2 + 9\theta_3\theta_6^2\big]\\ &+\frac{1}{9}\theta_1\theta_2\big[\theta_2^2 - 3\theta_6\big]\phi_1 - \frac{1}{9}\big[\theta_2^2 - 3\theta_6\big]\phi_2 - \frac{1}{18}\big[\theta_2^2 - 3\theta_6\big]\phi_5\\ &+\frac{1}{18}\big[5\theta_2^3 - 9\theta_2\theta_6 - 54\theta_4^2\big]\phi_7 + \frac{1}{18}\big[7\theta_2^3 - 9\theta_2\theta_6 - 54\theta_4^2\big]\phi_8\\ &-\frac{1}{9}\theta_1\big[\theta_2^2 - 3\theta_6\big]\phi_9 + \frac{1}{9}\theta_1\big[\theta_2^2 - 3\theta_6\big]\phi_{10} + \frac{1}{3}\theta_2\theta_3\phi_{11},\end{aligned} \tag{A.55}$$

$$\begin{aligned}\phi_9^2 =& \frac{1}{6}\left[\theta_1^2\theta_2^2\theta_6 - \theta_1^2\theta_6^2 - \theta_2^2\theta_3\theta_6 + 6\theta_2\theta_3\theta_4^2 + \theta_3\theta_6^2\right]\\ &-\frac{2}{3}\theta_1\big[\theta_2\theta_6 + 3\theta_4^2\big]\phi_1 + \frac{1}{6}\left[3\theta_2^2 - \theta_6\right]\phi_2 - \frac{1}{6}\left[3\theta_2^3 - 7\theta_2\theta_6 - 18\theta_4^2\right]\phi_7\\ &+\frac{2}{3}\theta_1\theta_6\phi_9,\end{aligned} \tag{A.56}$$

$$\begin{aligned}\phi_9\phi_{10} =& -\frac{1}{24}\big[\theta_1^2\theta_2^4 - 12\theta_1^2\theta_2\theta_4^2 - \theta_1^2\theta_6^2 + \theta_2^4\theta_3 + 12\theta_2\theta_3\theta_4^2 - \theta_3\theta_6^2\big]\\ &+\frac{1}{3}\theta_1\big[\theta_2^3 - \theta_2\theta_6 - 6\theta_4^2]\phi_4 - \frac{1}{6}\big[\theta_2^2 + \theta_6\big]\phi_5 - \frac{1}{3}\big[\theta_2^3 - 2\theta_2\theta_6 - 9\theta_4^2\big]\phi_8\\ &+\frac{1}{6}\theta_1\big[\theta_2^2 - \theta_6\big]\phi_9 + \frac{1}{3}\theta_1\theta_6\phi_{10} - \frac{1}{6}\theta_2\big[\theta_1^2 - 3\theta_3\big]\phi_{11},\end{aligned} \tag{A.57}$$

$$\begin{aligned}\phi_9\phi_{11} =& -\frac{1}{2}\theta_1\theta_4^2(\theta_2^2 - \theta_6) + \frac{1}{12}\big[3\theta_2^4 - 8\theta_2^2\theta_6 - 12\theta_2\theta_4^2 + \theta_6^2\big]\phi_1\\ &-\frac{1}{6}\big[\theta_2^2\theta_6 - 6\theta_2\theta_4^2 - \theta_6^2\big]\phi_4 + \frac{1}{3}\theta_2\theta_6\phi_9\\ &-\frac{1}{6}\big[3\theta_2^3 - 7\theta_2\theta_6 - 18\theta_4^2\big]\phi_{10} + \frac{1}{3}\theta_1\theta_6\phi_{11},\end{aligned} \tag{A.58}$$

$$\begin{aligned}\phi_{10}^2 =& -\frac{1}{12}\big[12\theta_1^2\theta_2\theta_4^2 + \theta_2^4\theta_3 - 2\theta_2^2\theta_3\theta_6 - 12\theta_2\theta_3\theta_4^2 + \theta_3\theta_6^2\big] + 2\theta_1\theta_4^2\phi_1\\ &-\frac{1}{6}\big[\theta_2^2 - \theta_6\big]\phi_2 + 2\theta_1\theta_4^2\phi_4 - \frac{1}{6}\big[\theta_2^2 - \theta_6\big]\phi_5\\ &+\frac{1}{6}\big[\theta_2^3 - \theta_2\theta_6 - 18\theta_4^2\big]\phi_7 + \frac{1}{6}\big[\theta_2^3 - \theta_2\theta_6 - 18\theta_4^2\big]\phi_8\\ &+\frac{1}{3}\theta_1\big[\theta_2^2 - \theta_6\big]\phi_{10},\end{aligned} \tag{A.59}$$

$$\begin{aligned}\phi_{10}\phi_{11} =& -\frac{1}{2}\theta_1\theta_4^2\big[\theta_2^2 - \theta_6\big] - \frac{1}{12}\big[\theta_2^4 - 24\theta_2\theta_4^2 - \theta_6^2\big]\phi_1\\ &-\frac{1}{12}\big[\theta_2^4 - 2\theta_2^2\theta_6 - 12\theta_2\theta_4^2 + \theta_6^2\big]\phi_4 + \frac{1}{6}\big[\theta_2^3 - \theta_2\theta_6 - 18\theta_4^2\big]\phi_9\\ &+\frac{1}{3}\big[\theta_2^3 - \theta_2\theta_6 - 9\theta_4^2\big]\phi_{10} + \frac{1}{6}\theta_1\big[\theta_2^2 - \theta_6\big]\phi_{11},\end{aligned} \tag{A.60}$$

$$\begin{aligned}\phi_{11}^2 =& -\frac{1}{8}\big[\theta_2^6 - 3\theta_2^4\theta_6 - 16\theta_2^3\theta_4^2 + 3\theta_2^2\theta_6^2 + 24\theta_2\theta_4^2\theta_6 + 72\theta_4^4 - \theta_6^3\big]\\ &+\frac{1}{2}\big[\theta_2^3 - \theta_2\theta_6 - 6\theta_4^2\big]\phi_{11}.\end{aligned} \tag{A.61}$$

Algebraic dependence syzygies for ϕ_6 and ϕ_{11} are given by the rewriting syzygies. For the other secondary invariants, the algebraic syzygies are higher-order polynomial expressions. A representative algebraic dependence syzygy for ϕ_4 is

$$
\begin{aligned}
0 = {} & \big[\theta_1^6\theta_2^6 - 12\theta_1^6\theta_2^3\theta_4^2 - 3\theta_1^6\theta_2^2\theta_6^2 - 36\theta_1^6\theta_2\theta_4^2\theta_6 + 36\theta_1^6\theta_4^4 - 6\theta_1^6\theta_6^3 - 3\theta_1^4\theta_2^6\theta_3 \\
& - 12\theta_1^4\theta_2^4\theta_3\theta_6 + 36\theta_1^4\theta_2^3\theta_3\theta_4^2 + 9\theta_1^4\theta_2^2\theta_3\theta_6^2 + 252\theta_1^4\theta_2\theta_3\theta_4^2\theta_6 - 324\theta_1^4\theta_3\theta_4^4 \\
& + 54\theta_1^4\theta_3\theta_6^3 + 4\theta_1^3\theta_2^6\theta_5 - 24\theta_1^3\theta_2^3\theta_4^2\theta_5 + 12\theta_1^3\theta_2^2\theta_5\theta_6^2 - 216\theta_1^3\theta_2\theta_4^2\theta_5\theta_6 \\
& - 48\theta_1^3\theta_5\theta_6^3 - 9\theta_1^2\theta_2^6\theta_3^2 + 36\theta_1^2\theta_2^4\theta_3^2\theta_6 + 36\theta_1^2\theta_2^3\theta_3^2\theta_4^2 + 27\theta_1^2\theta_2^2\theta_3^2\theta_6^2 \\
& - 324\theta_1^2\theta_2\theta_3^2\theta_4^2\theta_6 + 972\theta_1^2\theta_3^2\theta_4^4 - 126\theta_1^2\theta_3^2\theta_6^3 + 12\theta_1\theta_2^6\theta_3\theta_5 - 24\theta_1\theta_2^4\theta_3\theta_5\theta_6 \\
& - 216\theta_1\theta_2^3\theta_3\theta_4^2\theta_5 - 108\theta_1\theta_2^2\theta_3\theta_5\theta_6^2 + 648\theta_1\theta_2\theta_3\theta_4^2\theta_5\theta_6 + 216\theta_1\theta_3\theta_5\theta_6^3 \\
& - 9\theta_2^6\theta_3^2 + 4\theta_2^6\theta_5^2 + 36\theta_2^4\theta_3^2\theta_6 - 36\theta_2^4\theta_5^2\theta_6 + 180\theta_2^3\theta_3^2\theta_4^2 \\
& - 45\theta_2^2\theta_3^2\theta_6^2 + 108\theta_2^2\theta_5^2\theta_6^2 - 324\theta_2\theta_3^2\theta_4^2\theta_6 - 972\theta_3^2\theta_4^4 + 18\theta_3^2\theta_6^3 \\
& - 108\theta_5^2\theta_6^3\big] - \big[12\theta_1^5\theta_2^5 - 24\theta_1^5\theta_2^3\theta_6 - 144\theta_1^5\theta_2^2\theta_4^2 - 36\theta_1^5\theta_2\theta_6^2 - 144\theta_1^5\theta_4^2\theta_6 \\
& - 60\theta_1^3\theta_2^5\theta_3 + 504\theta_1^3\theta_2^2\theta_3\theta_4^2 + 252\theta_1^3\theta_2\theta_3\theta_6^2 + 1080\theta_1^3\theta_3\theta_4^2\theta_6 + 60\theta_1^2\theta_2^5\theta_5 \\
& - 48\theta_1^2\theta_2^3\theta_5\theta_6 - 648\theta_1^2\theta_2^2\theta_4^2\theta_5 - 108\theta_1^2\theta_2\theta_5\theta_6^2 - 648\theta_1^2\theta_4^2\theta_5\theta_6 - 72\theta_1\theta_2^5\theta_3^2 \\
& + 360\theta_1\theta_2^3\theta_3^2\theta_6 + 648\theta_1\theta_2^2\theta_3^2\theta_4^2 - 432\theta_1\theta_2\theta_3^2\theta_6^2 - 1944\theta_1\theta_3^2\theta_4^2\theta_6 + 60\theta_2^5\theta_3\theta_5 \\
& - 288\theta_2^3\theta_3\theta_5\theta_6 - 648\theta_2^2\theta_3\theta_4^2\theta_5 + 324\theta_2\theta_3\theta_5\theta_6^2 + 1944\theta_3\theta_4^2\theta_5\theta_6\big]\phi_4 \\
& + 6\big[5\theta_1^4\theta_2^4 - 30\theta_1^4\theta_2^2\theta_6 - 96\theta_1^4\theta_2\theta_4^2 - 3\theta_1^4\theta_6^2 - 54\theta_1^2\theta_2^4\theta_3 + 108\theta_1^2\theta_2^2\theta_3\theta_6 \\
& + 432\theta_1^2\theta_2\theta_3\theta_4^2 + 18\theta_1^2\theta_3\theta_6^2 + 40\theta_1\theta_2^4\theta_5 - 72\theta_1\theta_2^2\theta_5\theta_6 - 432\theta_1\theta_2\theta_4^2\theta_5 \\
& - 3\theta_2^4\theta_3^2 + 18\theta_2^2\theta_3^2\theta_6 - 27\theta_3^2\theta_6^2\big]\phi_4^2 + 48\big[2\theta_1^3\theta_2^3 + 6\theta_1^3\theta_2\theta_6 + 12\theta_1^3\theta_4^2 \\
& + 9\theta_1\theta_2^3\theta_3 - 21\theta_1\theta_2\theta_3\theta_6 - 54\theta_1\theta_3\theta_4^2 - 5\theta_2^3\theta_5 + 9\theta_2\theta_5\theta_6 + 54\theta_4^2\theta_5\big]\phi_4^3 \\
& - 144\big[3\theta_1^2\theta_2^2 + \theta_1^2\theta_6 + \theta_2^2\theta_3 - 3\theta_3\theta_6\big]\phi_4^4 + 576\theta_1\theta_2\phi_4^5 - 288\phi_4^6. \qquad \text{(A.62)}
\end{aligned}
$$

A.6. Trigonal groups C_3, S_6. The rewriting syzygies are

$$\phi_1^4 = -\frac{1}{4}\theta_3^2\theta_4^2 + \frac{1}{2}\theta_5\theta_6\phi_1 + \frac{5}{4}\theta_3\theta_4\phi_2 + \frac{1}{4}\theta_3\theta_6\phi_5 + \frac{1}{4}\theta_4\theta_5\phi_6, \qquad \text{(A.63)}$$

$$\phi_1\phi_4 = \phi_{11}, \qquad \text{(A.64)}$$

$$\phi_1\phi_5 = -\frac{1}{2}\theta_4\theta_5 + \frac{1}{2}\theta_3\phi_6, \qquad \text{(A.65)}$$

$$\phi_1\phi_6 = -\frac{1}{2}\theta_3\theta_6 + \frac{1}{2}\theta_4\phi_5, \qquad \text{(A.66)}$$

$$\phi_1\phi_7 = -\theta_2\theta_3\theta_4 + \frac{1}{2}\theta_3\phi_8 + \frac{1}{2}\theta_4\phi_{10}, \qquad \text{(A.67)}$$

$$\phi_1\phi_8 = \frac{1}{2}\theta_4\phi_7 + \frac{1}{2}\theta_3\phi_9, \qquad \text{(A.68)}$$

$$\phi_1\phi_9 = \theta_6\phi_4 + \theta_4\phi_8, \qquad \text{(A.69)}$$

$$\phi_1\phi_{10} = 2\theta_2\theta_3\phi_1 - \theta_5\phi_4 + \theta_3\phi_7, \qquad \text{(A.70)}$$

$$\phi_4^2 = \theta_3\theta_4 - \phi_2, \qquad \text{(A.71)}$$

$$\phi_4\phi_5 = -\theta_2\theta_3\theta_4 - \frac{1}{2}\theta_3\phi_8 + \frac{1}{2}\theta_4\phi_{10}, \qquad \text{(A.72)}$$

$$\phi_4\phi_6 = \frac{1}{2}\theta_4\phi_7 - \frac{1}{2}\theta_3\phi_9, \qquad \text{(A.73)}$$

$$\phi_4\phi_7 = +\frac{1}{2}\theta_4\theta_5 + \frac{1}{2}\theta_3\phi_6, \qquad \text{(A.74)}$$

$$\phi_4\phi_8 = -\frac{1}{2}\theta_3\theta_6 - \frac{1}{2}\theta_4\phi_5, \qquad \text{(A.75)}$$

$$\phi_4\phi_9 = -\theta_6\phi_1 - \theta_4\phi_6, \qquad \text{(A.76)}$$

$$\phi_4\phi_{10} = \theta_5\phi_1 + 2\theta_2\theta_3\phi_4 + \theta_3\phi_5, \qquad \text{(A.77)}$$

$$\phi_5^2 = \theta_3^2\theta_4 - \theta_3\phi_2 - \theta_5\phi_6, \tag{A.78}$$
$$\phi_5\phi_6 = \theta_5\theta_6 + 2\theta_3\theta_4\phi_1 - 2\phi_3, \tag{A.79}$$
$$\phi_5\phi_7 = -\theta_5\phi_8 + \theta_3\phi_{11}, \tag{A.80}$$
$$\phi_5\phi_8 = \theta_2\theta_3\theta_6 - \frac{1}{2}\theta_3\theta_4\phi_4 - \frac{1}{2}\theta_5\phi_9 - \frac{1}{2}\theta_6\phi_{10}, \tag{A.81}$$
$$\phi_5\phi_9 = -\theta_6\phi_7 - 2\theta_4\phi_{11}, \tag{A.82}$$
$$\phi_5\phi_{10} = \theta_3^2\phi_4 + 2\theta_2\theta_3\phi_5 - \theta_5\phi_7, \tag{A.83}$$
$$\phi_6^2 = \theta_3\theta_4^2 - \theta_4\phi_2 - \theta_6\phi_5, \tag{A.84}$$
$$\phi_6\phi_7 = \theta_2\theta_3\theta_6 + \frac{1}{2}\theta_3\theta_4\phi_4 - \frac{1}{2}\theta_5\phi_9 - \frac{1}{2}\theta_6\phi_{10}, \tag{A.85}$$
$$\phi_6\phi_8 = -\theta_6\phi_7 - \theta_4\phi_{11}, \tag{A.86}$$
$$\phi_6\phi_9 = -\theta_4^2\phi_4 - \theta_6\phi_8, \tag{A.87}$$
$$\phi_6\phi_{10} = 2\theta_2\theta_3\phi_6 - \theta_5\phi_8 + 2\theta_3\phi_{11}, \tag{A.88}$$
$$\phi_7^2 = \theta_3\phi_2 + \theta_5\phi_6, \tag{A.89}$$
$$\phi_7\phi_8 = -\theta_5\theta_6 - \theta_3\theta_4\phi_1 + 2\phi_3, \tag{A.90}$$
$$\phi_7\phi_9 = -\theta_3\theta_4^2 + 2\theta_4\phi_2 + \theta_6\phi_5, \tag{A.91}$$
$$\phi_7\phi_{10} = \theta_3^2\phi_1 + 2\theta_2\theta_3\phi_7 + \theta_5\phi_5, \tag{A.92}$$
$$\phi_8^2 = \theta_4\phi_2 + 1\theta_6\phi_5, \tag{A.93}$$
$$\phi_8\phi_9 = \theta_4^2\phi_1 + \theta_6\phi_6, \tag{A.94}$$
$$\phi_8\phi_{10} = -\theta_3^2\theta_4 + 2\theta_3\phi_2 + \theta_5\phi_6 + 2\theta_2\theta_3\phi_8, \tag{A.95}$$
$$\phi_9^2 = \big[\theta_4^3 - \theta_6^2\big], \tag{A.96}$$
$$\phi_9\phi_{10} = -\theta_5\theta_6 - 3\theta_3\theta_4\phi_1 + 4\phi_3 + 2\theta_2\theta_3\phi_9, \tag{A.97}$$
$$\phi_{10}^2 = -\big[4\theta_2^2\theta_3^2 - \theta_3^3 + \theta_5^2\big] + 4\theta_2\theta_3\phi_{10}. \tag{A.98}$$

A.7. Trigonal groups C_{3v}, D_3, D_{3d}. The rewriting syzygies for these groups are

$$\phi_1^4 = -\frac{1}{4}\theta_3^2\theta_4^2 - \frac{1}{2}\theta_5\big[2\theta_2\theta_3 - \theta_6\big]\phi_1 + \frac{5}{4}\theta_3\theta_4\phi_2 - \frac{1}{4}\theta_3\theta_5\phi_4 + \frac{1}{4}\theta_4\big[2\theta_2\theta_3 - \theta_6\big]\phi_5, \tag{A.99}$$
$$\phi_1\phi_4 = -\frac{1}{2}\theta_4\big[2\theta_2\theta_3 - \theta_6\big] + \frac{1}{2}\theta_3\phi_5, \tag{A.100}$$
$$\phi_1\phi_5 = \frac{1}{2}\theta_3\theta_5 + \frac{1}{2}\theta_4\phi_4, \tag{A.101}$$
$$\phi_4^2 = \theta_3^2\theta_4 - \theta_3\phi_2 + (\theta_6 - 2\theta_2\theta_3)\phi_5, \tag{A.102}$$
$$\phi_4\phi_5 = -\theta_5\big[2\theta_2\theta_3 - \theta_6\big] + 2\theta_3\theta_4\phi_1 - 2\phi_3, \tag{A.103}$$
$$\phi_5^2 = \theta_3\theta_4^2 - \theta_4\phi_2 + \theta_5\phi_4. \tag{A.104}$$

A.8. Hexagonal groups C_{3h}, C_6, C_{6h}. The rewriting syzygies are

$$\begin{aligned}\phi_1^4 = {} & -\frac{1}{4}\theta_3^2\theta_4\big[21\theta_4^3 - 2\theta_6\big] + \frac{1}{2}\theta_5\big[8\theta_4^3 - \theta_6\big]\phi_1 \\ & + \frac{5}{4}\theta_3\theta_4^2\phi_2 + \frac{1}{4}\theta_3\big[10\theta_4^3 - \theta_6\big]\phi_6 + \frac{1}{4}\theta_4^2\theta_5\phi_8,\end{aligned} \tag{A.105}$$
$$\phi_1\phi_4 = 2\theta_4\phi_7 - \phi_{10}, \tag{A.106}$$
$$\phi_1\phi_5 = 2\theta_2\theta_3\phi_1 - \theta_5\phi_4 + 2\theta_3\phi_7, \tag{A.107}$$
$$\phi_1\phi_6 = -\frac{1}{2}\theta_4^2\theta_5 + \frac{1}{2}\theta_3\phi_8, \tag{A.108}$$

$$\phi_1\phi_7 = -\frac{1}{2}\theta_2\theta_3\theta_4^2 + \frac{1}{2}\theta_3\theta_4\phi_4 + \frac{1}{4}\theta_4^2\phi_5 + \frac{1}{4}\theta_3\phi_9, \tag{A.109}$$

$$\phi_1\phi_8 = -\frac{1}{2}\theta_3\big[12\theta_4^3 - \theta_6\big] + 4\theta_4\phi_2 + \frac{1}{2}\theta_4^2\phi_6, \tag{A.110}$$

$$\phi_1\phi_9 = -3\theta_4^2\phi_7 + 2\theta_4\phi_{10} + \theta_3\phi_{11}, \tag{A.111}$$

$$\begin{aligned}\phi_1\phi_{10} = {}& -\frac{1}{2}\theta_2\theta_3\big[12\theta_4^3 - \theta_6\big] + \frac{3}{4}\theta_3\theta_4^2\phi_4 \\ & + \frac{1}{4}\big[12\theta_4^3 - \theta_6\big]\phi_5 + \frac{1}{2}\theta_3\theta_4\phi_9 - \frac{1}{2}\theta_5\phi_{11},\end{aligned} \tag{A.112}$$

$$\phi_1\phi_{11} = \frac{1}{2}\big[12\theta_4^3 - \theta_6\big]\phi_4 + \frac{1}{2}\theta_4^2\phi_9, \tag{A.113}$$

$$\phi_4^2 = \theta_3\theta_4^2 - \phi_2, \tag{A.114}$$

$$\phi_4\phi_5 = -2\theta_3^2\theta_4 + \theta_5\phi_1 + 2\theta_2\theta_3\phi_4 + \theta_3\phi_6, \tag{A.115}$$

$$\phi_4\phi_6 = -\theta_2\theta_3\theta_4^2 + \theta_3\theta_4\phi_4 + \frac{1}{2}\theta_4^2\phi_5 - \frac{1}{2}\theta_3\phi_9, \tag{A.116}$$

$$\phi_4\phi_7 = \frac{1}{4}\theta_4^2\theta_5 - \theta_3\theta_4\phi_1 + \frac{1}{4}\theta_3\phi_8, \tag{A.117}$$

$$\phi_4\phi_8 = 9\theta_4^2\phi_7 - 4\theta_4\phi_{10} - \theta_3\phi_{11}, \tag{A.118}$$

$$\phi_4\phi_9 = -\frac{1}{2}\theta_3\big[12\theta_4^3 - \theta_6\big] + 2\theta_4\phi_2 - \frac{1}{2}\theta_4^2\phi_6, \tag{A.119}$$

$$\phi_4\phi_{10} = \frac{1}{2}\theta_4^3\theta_5 - 3\theta_3\theta_4^2\phi_1 + \phi_3 + \frac{1}{2}\theta_3\theta_4\phi_8, \tag{A.120}$$

$$\phi_4\phi_{11} = -\frac{1}{2}\big[6\theta_4^3 - \theta_6\big]\phi_1 - \frac{1}{2}\theta_4^2\phi_8, \tag{A.121}$$

$$\phi_5^2 = -\big[4\theta_2^2\theta_3^2 - \theta_3^3 + \theta_5^2\big] + 4\theta_2\theta_3\phi_5, \tag{A.122}$$

$$\phi_5\phi_6 = -4\theta_2\theta_3^2\theta_4 + \theta_3^2\phi_4 + 2\theta_3\theta_4\phi_5 + 2\theta_2\theta_3\phi_6 - 2\theta_5\phi_7, \tag{A.123}$$

$$\phi_5\phi_7 = -\theta_3\theta_4\theta_5 + \frac{1}{2}\theta_3^2\phi_1 + \frac{1}{2}\theta_5\phi_6 + 2\theta_2\theta_3\phi_7, \tag{A.124}$$

$$\phi_5\phi_8 = -6\theta_4\theta_5\phi_4 + 12\theta_3\theta_4\phi_7 + 2\theta_2\theta_3\phi_8 - \theta_5\phi_9 - 2\theta_3\phi_{10}, \tag{A.125}$$

$$\phi_5\phi_9 = 3\theta_3^2\theta_4^2 - 6\theta_4\theta_5\phi_1 + 2\theta_3\phi_2 - 2\theta_3\theta_4\phi_6 + \theta_5\phi_8 + 2\theta_2\theta_3\phi_9, \tag{A.126}$$

$$\phi_5\phi_{10} = -\frac{3}{2}\theta_3\theta_4^2\theta_5 + 3\theta_3^2\theta_4\phi_1 - \theta_5\phi_2 + \theta_4\theta_5\phi_6 - \frac{1}{2}\theta_3^2\phi_8 + 2\theta_2\theta_3\phi_{10}, \tag{A.127}$$

$$\phi_5\phi_{11} = -\frac{1}{2}\theta_5\big[10\theta_4^3 - \theta_6\big] - \frac{3}{2}\theta_3\theta_4^2\phi_1 + 2\phi_3 + 2\theta_2\theta_3\phi_{11}, \tag{A.128}$$

$$\phi_6^2 = -3\theta_3^2\theta_4^2 + 4\theta_4\theta_5\phi_1 - \theta_3\phi_2 + 4\theta_3\theta_4\phi_6 - \theta_5\phi_8, \tag{A.129}$$

$$\phi_6\phi_7 = -\theta_4\theta_5\phi_4 + 3\theta_3\theta_4\phi_7 - \frac{1}{2}\theta_5\phi_9 - \frac{1}{2}\theta_3\phi_{10}, \tag{A.130}$$

$$\phi_6\phi_8 = \theta_5\big[8\theta_4^3 - \theta_6\big] - 6\theta_3\theta_4^2\phi_1 - 2\phi_3 + 4\theta_3\theta_4\phi_8, \tag{A.131}$$

$$\phi_6\phi_9 = \theta_2\theta_3\big[12\theta_4^3 - \theta_6\big] + \frac{3}{2}\theta_3\theta_4^2\phi_4 - \frac{1}{2}\big[12\theta_4^3 - \theta_6\big]\phi_5 + 3\theta_3\theta_4\phi_9 - \theta_5\phi_{11}, \tag{A.132}$$

$$\phi_6\phi_{10} = -\frac{3}{2}\theta_4^2\theta_5\phi_4 + \frac{3}{2}\theta_3\theta_4^2\phi_7 - \theta_4\theta_5\phi_9 + \theta_3\theta_4\phi_{10} + \frac{1}{2}\theta_3^2\phi_{11}, \tag{A.133}$$

$$\phi_6\phi_{11} = -\big[12\theta_4^3 - \theta_6\big]\phi_7 + \theta_4^2\phi_{10} + 2\theta_3\theta_4\phi_{11}, \tag{A.134}$$

$$\phi_7^2 = -\theta_4\theta_5\phi_1 + \frac{1}{4}\theta_3\phi_2 + \frac{1}{4}\theta_5\phi_8, \tag{A.135}$$

$$\phi_7\phi_8 = \frac{1}{2}\theta_2\theta_3\big[6\theta_4^3 - \theta_6\big] + \frac{9}{4}\theta_3\theta_4^2\phi_4 - \frac{1}{4}\big[6\theta_4^3 - \theta_6\big]\phi_5 + \theta_3\theta_4\phi_9 - \frac{1}{2}\theta_5\phi_{11}, \tag{A.136}$$

$$\phi_7\phi_9 = -\frac{1}{2}\theta_5\big[11\theta_4^3 - \theta_6\big] + \frac{3}{2}\theta_3\theta_4^2\phi_1 + \phi_3 - \frac{1}{2}\theta_3\theta_4\phi_8, \tag{A.137}$$

$$\phi_7\phi_{10} = \frac{1}{8}\theta_3^2\big[12\theta_4^3 - \theta_6\big] - \frac{9}{4}\theta_4^2\theta_5\phi_1 + \frac{1}{2}\theta_3\theta_4\phi_2 - \frac{1}{8}\theta_3\theta_4^2\phi_6 + \frac{1}{2}\theta_4\theta_5\phi_8, \tag{A.138}$$

$$\phi_7\phi_{11} = -\frac{1}{4}\theta_3\theta_4\big[21\theta_4^3 - 2\theta_6\big] + \frac{1}{2}\theta_4^2\phi_2 + \frac{1}{4}\big[10\theta_4^3 - \theta_6\big]\phi_6, \tag{A.139}$$

$$\phi_8^2 = -\theta_3\theta_4\big[27\theta_4^3 - 2\theta_6\big] + 15\theta_4^2\phi_2 - \big[6\theta_4^3 - \theta_6\big]\phi_6, \tag{A.140}$$

$$\phi_8\phi_9 = -2\big[18\theta_4^3 - \theta_6\big]\phi_7 + 9\theta_4^2\phi_{10} + 6\theta_3\theta_4\phi_{11}, \tag{A.141}$$

$$\begin{aligned}\phi_8\phi_{10} = {}& -\frac{1}{2}\theta_2\theta_3\theta_4\big[27\theta_4^3 - 2\theta_6\big] + \frac{1}{2}\theta_3\big[18\theta_4^3 - \theta_6\big]\phi_4 \\ & + \frac{1}{4}\theta_4\big[27\theta_4^3 - 2\theta_6\big]\phi_5 + \frac{9}{4}\theta_3\theta_4^2\phi_9 - 3\theta_4\theta_5\phi_{11},\end{aligned} \tag{A.142}$$

$$\phi_8\phi_{11} = \frac{1}{2}\theta_4\big[27\theta_4^3 - 2\theta_6\big]\phi_4 - \frac{1}{2}\big[6\theta_4^3 - \theta_6\big]\phi_9, \tag{A.143}$$

$$\phi_9^2 = -3\theta_4^2\phi_2 + \big[12\theta_4^3 - \theta_6\big]\phi_6, \tag{A.144}$$

$$\phi_9\phi_{10} = -\frac{1}{4}\theta_4\theta_5\big[45\theta_4^3 - 4\theta_6\big] + \frac{1}{2}\theta_3\big[18\theta_4^3 - \theta_6\big]\phi_1 - \frac{3}{4}\theta_3\theta_4^2\phi_8, \tag{A.145}$$

$$\phi_9\phi_{11} = -\frac{1}{2}\theta_4\big[27\theta_4^3 - 2\theta_6\big]\phi_1 + \frac{1}{2}\big[12\theta_4^3 - \theta_6\big]\phi_8, \tag{A.146}$$

$$\begin{aligned}\phi_{10}^2 = {}& \frac{1}{4}\theta_3^2\theta_4\big[45\theta_4^3 - 4\theta_6\big] - \frac{1}{2}\theta_5\big[18\theta_4^3 - \theta_6\big]\phi_1 \\ & + \frac{3}{4}\theta_3\theta_4^2\phi_2 - \frac{1}{4}\theta_3\big[12\theta_4^3 - \theta_6\big]\phi_6 + \frac{3}{4}\theta_4^2\theta_5\phi_8,\end{aligned} \tag{A.147}$$

$$\phi_{10}\phi_{11} = -\frac{1}{4}\theta_3\theta_4^2\big[54\theta_4^3 - 5\theta_6\big] + \frac{1}{2}\big[12\theta_4^3 - \theta_6\big]\phi_2 + \frac{1}{4}\theta_4\big[21\theta_4^3 - 2\theta_6\big]\phi_6, \tag{A.148}$$

$$\phi_{11}^2 = -\frac{1}{4}\big[99\theta_4^6 - 20\theta_4^3\theta_6 + \theta_6^2\big]. \tag{A.149}$$

A.9. Hexagonal groups $\boldsymbol{D_{3h}}$, $\boldsymbol{C_{6v}}$, $\boldsymbol{D_6}$, $\boldsymbol{D_{6h}}$. The rewriting syzygies are

$$\begin{aligned}\phi_1^4 = {}& \frac{1}{4}\big[19\theta_3^2\theta_4^4 - 2\theta_3^2\theta_4\theta_6\big] - \frac{1}{2}\big[22\theta_2\theta_3\theta_4^3 - 11\theta_4^3\theta_5 - 2\theta_2\theta_3\theta_6 + \theta_5\theta_6\big]\phi_1 \\ & + \frac{5}{4}\theta_3\theta_4^2\phi_2 - \frac{1}{4}\theta_3\big[10\theta_4^3 - \theta_6\big]\phi_4 - \frac{1}{4}\theta_4^2\big[2\theta_2\theta_3 - \theta_5\big]\phi_5,\end{aligned} \tag{A.150}$$

$$\phi_1\phi_4 = \left[-\theta_2\theta_3\theta_4^2 + \frac{1}{2}\theta_4^2\theta_5\right] + \theta_3\theta_4\phi_1 - \frac{1}{2}\theta_3\phi_5, \tag{A.151}$$

$$\phi_1\phi_5 = \left[-4\theta_3\theta_4^3 + \frac{1}{2}\theta_3\theta_6\right] - 2\theta_4\phi_2 - \frac{1}{2}\theta_4^2\phi_4, \tag{A.152}$$

$$\begin{aligned}\phi_4^2 = {}& -3\theta_3^2\theta_4^2 + 2\theta_4\big[2\theta_2\theta_3 - \theta_5\big]\phi_1 \\ & - \theta_3\phi_2 + 4\theta_3\theta_4\phi_4 + \big[2\theta_2\theta_3 - \theta_5\big]\phi_5,\end{aligned} \tag{A.153}$$

$$\begin{aligned}\phi_4\phi_5 = {}& \big[22\theta_2\theta_3\theta_4^3 - 2\theta_2\theta_3\theta_6 - 11\theta_4^3\theta_5 + \theta_5\theta_6\big] \\ & + 2\phi_3 + 3\theta_3\theta_4\phi_5,\end{aligned} \tag{A.154}$$

$$\phi_5^2 = -3\theta_3\theta_4^4 + 3\theta_4^2\phi_2 + \big[12\theta_4^3 - \theta_6\big]\phi_4. \tag{A.155}$$

Acknowledgments. I thank B. Sturmfels and M. Rashid for helpful comments.

REFERENCES

[1] E. A. Arnold, *Modular algorithms for computing Gröbner bases*, J. Symbolic Comput., 35 (2003), pp. 403–419.

[2] D. Bayer and M. Stillman, *A theorem on refining division orders by the reverse lexicographical order*, Duke Math. J., 55 (1987), pp. 321–328.

[3] T. Becker and V. Weispfenning, *Gröbner Bases: A Computational Approach to Commutative Algebra*, Graduate Texts in Math. 141, Springer-Verlag, New York, 1993.

[4] G. Birkhoff and S. MacLane, *A Survey of Modern Algebra*, 5th ed., A. K. Peters, Natick, MA, 1997.

[5] B. Buchberger, *Gröbner bases: An algorithmic method in polynomial ideal theory*, in Multidimensional Systems Theory, N. K. Bose, ed., D. Reidel, Boston, 1985, pp. 184–232.

[6] F. A. Cotton, *Chemical Applications of Group Theory*, Wiley-Interscience, New York, 1990.

[7] J.-C. Faugére, *A new efficient algorithm for computing Gröbner bases* (F_4), J. Pure Appl. Algebra, 139 (1999), pp. 61–88.

[8] S. Forte and M. Vianello, *Symmetry classes for elasticity tensors*, J. Elasticity, 43 (1996), pp. 81–108.

[9] K. Gatermann, *Computer Algebra Methods for Equivariant Dynamical Systems*, Springer-Verlag, Berlin, 2000.

[10] R. Gebauer and H. M. Möller, *On an installation of Buchberger's algorithm*, J. Symbolic Comput., 6 (1988), pp. 275–286.

[11] A. Giovini, T. Mora, G. Niesi, L. Robbiano, and C. Traverso, *"One sugar cube, please" or Selection strategies in the Buchberger algorithm*, in Proceedings of the International Symposium on Algorithms and Computation, ACM, New York, 1991, pp. 49–54.

[12] T. Granlund, *GNU MP: The GNU Multiple Precision Arithmetic Library*, edition 4.1.2, Tech. report, Swox AB, Stockholm, Sweden, 2002.

[13] E. Mayr and A. Meyer, *The complexity of the word problem for commutative semigroups and polynomial ideals*, Adv. Math., 46 (1982), pp. 305–329.

[14] G. H. Miller, *An iterative Riemann solver for systems of hyperbolic conservation laws, with application to hyperelastic solid mechanics*, J. Comput. Phys., 193 (2003), pp. 198–225.

[15] G. H. Miller and P. Colella, *A high-order Eulerian Godunov method for elastic-plastic flow in solids*, J. Comput. Phys., 167 (2001), pp. 131–176.

[16] G. H. Miller and P. Colella, *A conservative three-dimensional Eulerian method for coupled fluid-solid shock capturing*, J. Comput. Phys., 183 (2002), pp. 26–82.

[17] T. Molien, *Über die invarianten der linearen substitutionsgruppe*, Sitzungsber. Königl. Preuss. Akad. Wiss., (1897), pp. 1152–1156.

[18] F. D. Murnaghan, *The Theory of Group Representations*, Dover, New York, 1963.

[19] N. Sloane, *Error-correcting codes and invariant theory: New applications of a nineteenth-century technique*, Amer. Math. Monthly, 84 (1977), pp. 82–107.

[20] G. F. Smith and R. S. Rivlin, *The strain-energy function for anisotropic elastic materials*, Trans. Amer. Math. Soc., 88 (1958), pp. 175–193.

[21] A. J. M. Spencer, *On generating functions for the number of invariants of orthogonal tensors*, Mathematika, 17 (1970), pp. 275–286.

[22] A. J. M. Spencer, *Theory of invariants*, in Continuum Physics, Vol. 1, A. Eringen, ed., Academic Press, New York, 1971, pp. 239–353.

[23] R. Stanley, *Invariants of finite groups and their applications to combinatorics*, Bull. Amer. Math. Soc. (N.S.), 1 (1979), pp. 475–511.

[24] B. Sturmfels, *Algorithms in Invariant Theory*, Springer-Verlag, New York, 1993.

[25] H. Weyl, *The Classical Groups, Their Invariants and Representations*, Princeton University Press, Princeton, NJ, 1946.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 2076–2095

ANALYTICAL SOLUTIONS OF A GROWTH MODEL FOR A MELT REGION INDUCED BY A FOCUSED LASER BEAM*

ANTOINE SAUCIER†, JEAN-YVES DEGORCE‡, AND MICHEL MEUNIER‡

Abstract. We consider processes in which a focused laser beam is used to induce the melting of silicium. The first goal of this paper is to propose a simple three-dimensional (3D) model of this melting process. Our model is partly based on an energy balance equation. This model leads to a nontrivial ODE describing the evolution in time of the dimension of the melt region. The second goal of this paper is to obtain approximate analytical solutions of this ODE. After using basic solution methods, we propose an original geometrical method to derive asymptotic solutions for time $\to \infty$. These solutions turn out to be the most useful for the description of this process.

Key words. focused laser beam, melting of material, three-dimensional (3D) modeling, ODE, analytic solution, asymptotic solution, geometrical method

AMS subject classifications. 34A05, 34A26, 34E10, 65L05, 74N20

DOI. 10.1137/S0036139902413015

1. Introduction.

Focusing an energetic pulsed photon beam on a material usually leads to a localized heating, possibly followed by atomic vaporization and even by an ejection of materials, a process called ablation [5]. All these mechanisms contribute to the dissipation of the laser beam energy into the materials. The time and spatial distribution of these dissipation phenomena depend on the localized heat source parameters and on the materials' properties. In this paper, we assume that the beam does not cause any ablation and that all the incoming energy is dissipated into heat, which leads to a local increase of temperature and to a melt region. The object of this study is the time evolution of the melt region size.

In general, these heating and melting effects constitute a three-dimensional (3D) heat flow problem usually solved numerically [4]. An analytical solution, even approximate, is very interesting because it allows analyzing the influence of the various physical parameters involved. Actually, simplifications to a one-dimensional (1D) heat flow problem have been proposed by many authors [9], [7], [6] for the case of a large beam dimension when compared to the heat diffusion length. However, for a long pulsed focused beam with beam dimension comparable to or smaller than the melt depth, the lateral heat flow is on the same order of magnitude as the perpendicular component. It follows that the 1D approximation is no longer valid. Nonlinear boundary conditions arising from a moving solid-liquid interface make exact analytical solutions of the 3D heat flow equation very difficult.

In this paper, we present a simplified 3D model based on an energy balance equation. This model was first introduced in [3] with a brief justification and with an emphasis on the comparison with experimental results. In this paper, our first goal is to present a *complete derivation* of this model (section 2), with an emphasis on the

*Received by the editors August 8, 2002; accepted for publication (in revised form) March 14, 2004; published electronically September 2, 2004.

http://www.siam.org/journals/siap/64-6/41301.html

†Department of Applied Mathematics and Industrial Engineering, École Polytechnique de Montréal, C.P. 6079, succ. centre-ville, Montréal, Québec, Canada, H3C-3A7 (Antoine.Saucier@polymtl.ca).

‡Laser Processing Laboratory, Department of Engineering Physics, École Polytechnique de Montréal, C.P. 6079, succ. centre-ville, Montréal, Québec, Canada, H3C-3A7 (jean-yves.degorce@polymtl.ca, meunier@email.phys.polymtl.ca).

nature of the approximations used. Our model leads to an ODE that describes the evolution in time of the dimension of the melt region for a material irradiated by a focused laser beam. The second goal of this paper is to obtain *approximate analytical solutions* of this ODE. In section 3, we analyze this ODE in detail using classical analytical and numerical methods. In sections 4 and 5, we derive asymptotic solutions for $t \to \infty$, using a possibly original geometrical method. Finally, in section 6, we compare the accuracy of the approximate solutions.

2. The model. Our model is based on four main hypotheses that we present and justify in the following.

2.1. Hypothesis A: The focused laser beam can be treated as a point heat source. In this paper, we assume that the laser beam is orthogonal to the flat material surface. A focused laser beam is most often characterized by a Gaussian curve of width r_0. The light intensity varies according to $I(r) = I_0 \ \exp(-(r/r_0)^2)$, where r is the distance from the beam center in a direction perpendicular to the beam. The heat diffusion characteristic length scale for a material of heat diffusivity D is usually defined by $d(t) = \sqrt{D\ t}$. $d(t)$ is an estimate of the heat front penetration depth at time t, assuming that the laser is turned on at $t = 0$.

A 1D model of the heating process is obtained if $r_0 >> d$. In this limit, the beam radius can be regarded as infinite. For an isotropic material, the resulting isothermals are planes which are perpendicular to the laser beam. If the beam is perpendicular to the flat material surface, then the isothermals are planes which are parallel to this surface.

A 3D model of the heating process is obtained if $r_0 << d(t)$. Photons entering a material are absorbed progressively. It follows that the light intensity within the material decreases according to an exponential law (Beers's law) characterized by a penetration depth ℓ (the inverse of the absorptivity). We develop our model of the heating process in the *point source approximation framework*, where both r_0 and ℓ are much smaller than the diffusion length, i.e.,

$$\begin{cases} r_0 << \sqrt{D\ t_{\rm p}} & \text{and} \\ \ell << \sqrt{D\ t_{\rm p}}, \end{cases} \tag{2.1}$$

where $t_{\rm p}$ is the pulse width and $d(t_{\rm p}) = \sqrt{D\ t_{\rm p}}$ is the diffusion length. We emphasize that the point source approximation (2.1) implies that our model is not expected to be valid or accurate for $t \approx 0$.

2.2. Hypothesis B: Heat losses at the surface of the melt domain are negligible during the whole melting process. Heat losses occur through two interfaces: the flat upper surface of the melt domain and the liquid-solid interface. In this section, we compare the magnitudes of these heat losses.

On one hand, it is a well-known experimental observation that the flat surface of the liquid domain has approximately the shape of a disk. On the other hand, the liquid-solid interface is a surface which is attached to the circumference of this disk. During the melting process, this surface is symmetric with respect to an axis going through the point heat source in the direction perpendicular to the solid-air plane. This symmetry implies that the area $\mathcal{A}$ of the liquid-solid interface can be expressed solely as a function of the disk radius r, i.e., $\mathcal{A} = \mathcal{A}(r)$. If we assume that the liquid-solid interface is not flat, then its area $\mathcal{A}(r)$ will be at least as large as

the area of the disk, i.e.,

$$\frac{\pi\ r^2}{\mathcal{A}(r)} \leq 1 \tag{2.2}$$

for $r > 0$.

Let us consider the ratio of the heat diffusion losses through these two surfaces. This ratio is $R = \frac{J_{\text{liquid/air}}}{J_{\text{liquid/solid}}}$, where $J_{\text{liquid/air}}$ and $J_{\text{liquid/solid}}$ are the heat fluxes through each interface. We use

$$\begin{cases} J_{\text{liquid/air}} &= \kappa_{\text{air}}\ \pi\ r^2 \quad \| \nabla T_{\text{air}} \|, \\ J_{\text{liquid/solid}} &= \kappa_{\text{solid}}\ \mathcal{A}(r) \quad \| \nabla T_{\text{solid}} \|, \end{cases} \tag{2.3}$$

where $(\kappa_{\text{air}}, \kappa_{\text{solid}})$ are the heat conductivities of the air and of the solid, respectively, and $(\| \nabla T_{\text{air}} \|, \| \nabla T_{\text{solid}} \|)$ are the magnitudes of the temperature gradients at the liquid surface in the air and in the solid, respectively. Using (2.3), the ratio R takes the form

$$R = \frac{\pi\ r^2}{\mathcal{A}(r)}\ \frac{\kappa_{\text{air}}}{\kappa_{\text{solid}}}\ \frac{\| \nabla T_{\text{air}} \|}{\| \nabla T_{\text{solid}} \|}. \tag{2.4}$$

If the solid and the air are at the same temperature initially, then we expect the temperature gradients at both interfaces to have similar magnitudes during the whole melting process, i.e., $\|\nabla T_{\text{air}}\| \approx \|\nabla T_{\text{solid}}\|$, and therefore $R \approx \frac{\pi\ r^2}{\mathcal{A}(r)}\ \frac{\kappa_{\text{air}}}{\kappa_{\text{solid}}}$. The inequality (2.2) then implies that R has an upper bound:

$$R \leq \frac{\kappa_{\text{air}}}{\kappa_{\text{solid}}}. \tag{2.5}$$

In general, the heat conductivity of gases is typically 100 times smaller than for solids, and therefore (2.5) implies that $R < 1/100$. Consequently, it seems reasonable to neglect heat losses in the air during the melting process. This kind of approximation has been discussed in the literature by Wood and Geist [8], who also took into account convection in the air and radiations.

2.3. Hypothesis C: The melt domain is hemispherical. We shall see that this hypothesis is essentially a consequence of the hypotheses A and B. We assume that the following three conditions are satisfied: the solid material is isotropic; the heat source is a point source (hypothesis A); the surface of the melt domain is effectively a thermal insulator (hypothesis B). It follows from these hypotheses that the temperature distribution has a *spherical symmetry*, i.e., that $T = T(r)$, where r is the distance from the point source.

The existence of a spherical symmetry can be understood by comparing our problem with another similar problem. Consider a point heat source within an *infinite* isotropic solid material, instead of a semi-infinite material. In this case, the temperature has obviously a spherical symmetry. Moreover, this symmetry implies that the heat flux going through an arbitrary plane containing the point heat source vanishes at each point of this plane. It follows that the problem with a spherical symmetry has exactly the same boundary condition (i.e., zero heat flux along a plane containing the source) as our problem in a semi-infinite material. For this reason, we expect the two problems to have the same symmetry.

Convective flow driven either by buoyancy or surface tension does not have enough time to develop for durations shorter than 1 μs, which is the laser pulse width in our application [1]. It follows that convection does not break the spherical symmetry.

Spherical symmetry implies that the isothermals are hemispherical. The liquid-solid interface, which is an isothermal, is therefore hemispherical.

2.4. Hypothesis D: Everything happens as if the heat flux at the surface of the melt region was transported instantaneously to the solid-liquid interface. In this section, we estimate the heat flux which crosses the liquid-solid interface. During a fixed time interval τ, the laser releases a constant quantity of energy which is absorbed at the top surface of the melt domain. This energy is used to heat the melt fluid, to melt a hemispherical shell of solid, and to heat the solid via diffusion. The thermal energy is transferred to the liquid-solid interface via conduction in the melt phase (convection being negligible).

In the framework of the Stefan problem [2], the energy density balance during a time lapse dt is evaluated at the moving liquid-solid interface (Figure 2.1):

$$j_{\text{in}}\ dt = j_{\text{out}}\ dt + L\ dr, \tag{2.6}$$

where j_{in} is the heat flux that reaches the interface inside the melt fluid, j_{out} is the heat flux diffused into the solid at the liquid-solid interface, and $L\ dr$ is the heat flux used to melt a region of solid of depth dr.

We integrate the energy balance equation (2.6) over the hemispherical shell of radius r and divide by dt to obtain the heat transfer rate balance

$$\int_{\text{interface}} j_{\text{in}}\ dS = j_{\text{out}}\ 2\pi r^2 + L\ \frac{dr}{dt}\ 2\pi r^2. \tag{2.7}$$

According to hypothesis B, we neglect heat losses in the air. It follows that the heat transfer rate $\int_{\text{interface}} j_{\text{in}}\ dS$ is the power provided by the laser (that we denote by P) minus the power used to heat the melt fluid (that we denote by $\frac{dE_{\text{h}}}{dt}$):

$$\int_{\text{interface}} j_{\text{in}}\ dS = P - \frac{dE_{\text{h}}}{dt}. \tag{2.8}$$

Substituting (2.8) into (2.7) yields

$$P = j_{\text{out}}\ 2\pi r^2 + L\ \frac{dr}{dt}\ 2\pi r^2 + \frac{dE_{\text{h}}}{dt}. \tag{2.9}$$

We approximate j_{out} by the linearization

$$j_{\text{out}} = -\kappa_{\text{s}}\ \left(\frac{\partial T}{\partial r}\right)_{r=r_{\text{m}}} \approx \frac{\kappa_{\text{s}}\ \Delta_{\text{s}}T}{\xi\sqrt{D\ t}}, \tag{2.10}$$

where κ_{s} is heat conductivity of the solid phase, r_{m} is the radius of the melt region, $\Delta_{\text{s}}T \equiv T_{\text{m}} - T_{\text{s}}$, T_{m} and T_{s} are the silicium fusion temperature and the solid silicium temperature far from the melt region, respectively (T_{s} equals the room temperature T_{room}), D is the thermal diffusivity in the solid at the fusion temperature T_{m}, $\sqrt{D\ t}$ is the heat diffusion characteristic length scale of the solid, and ξ is a geometry dependent constant usually fixed to 1.

The energy used to heat a hemisphere of melt solid satisfies

$$\frac{dE_h}{dt} = c_\ell\ \frac{2}{3}\ \pi\ r^3\ \frac{dT_\ell}{dt}, \tag{2.11}$$

where c_ℓ is the liquid silicium specific heat and T_ℓ is the mean temperature of the liquid silicium.

Substituting (2.10) and (2.11) into (2.9) leads to the energy transfer rate balance equation

$$P = \frac{\kappa_{\mathrm{s}}\ \Delta_{\mathrm{s}}T}{\xi\sqrt{D\ t}}\ 2\pi r^2 + L\ \frac{dr}{dt}\ 2\pi r^2 + c_\ell\ \frac{2}{3}\ \pi\ r^3\ \frac{dT_\ell}{dt}. \tag{2.12}$$

We will now compare the magnitude of the three terms

$$\begin{cases} P_{\mathrm{diffusion}} = \frac{\kappa_{\mathrm{s}}\ \Delta_{\mathrm{s}}T}{\xi\sqrt{D\ t}}\ 2\ \pi\ r^2, \\ P_{\mathrm{melting}} = L\ \frac{dr}{dt}\ 2\ \pi\ r^2, \\ P_{\mathrm{liquid\ heating}} = c_\ell\ \frac{2}{3}\ \pi\ r^3\ \frac{dT_\ell}{dt}. \end{cases} \tag{2.13}$$

In the second and third equations of (2.13), the instantaneous rates $\frac{dr}{dt}$ and $\frac{dT_\ell}{dt}$ are unknown a priori. However, the average rates can be estimated. For a time lapse τ, the average rates are defined by $\frac{dr}{dt}|_{\mathrm{mean}} = r/\tau$ and $\frac{dT_\ell}{dt}|_{\mathrm{mean}} = \frac{\Delta_\ell T}{\tau}$. In this problem, the laser beam power is constant and the size of the melt pool grows with time. Because the volume to heat increases as time passes, we expect both dr/dt and dT_ℓ/dt to decrease with time, i.e., to have negative second derivatives. This implies that *the average rates are larger than the instantaneous rates*. Substituting the average rates into (2.13) yields

$$\begin{cases} P_{\mathrm{diffusion}} = \frac{\kappa_{\mathrm{s}}\ \Delta_{\mathrm{s}}T}{\xi\sqrt{D\ t}}\ 2\ \pi\ r^2, \\ P_{\mathrm{melting}} \le L\ \frac{r}{\tau}\ 2\ \pi\ r^2, \\ P_{\mathrm{liquid\ heating}} \le c_\ell\ \frac{2}{3}\ \pi\ r^3\ \frac{\Delta_\ell T}{\tau}. \end{cases} \tag{2.14}$$

We use parameter values which are close to the ones observed experimentally for a focused laser beam on silicium, i.e., $r = 1\ \mu$m, $D = 0.1\ \mathrm{cm}^2/\mathrm{s}$, $\kappa_{\mathrm{s}} = 0.3\ \mathrm{W}/(\mathrm{cm}\ ^\circ\mathrm{K})$, $\Delta_\ell T = T_{\mathrm{vapor(Si)}} - T_{\mathrm{melt(Si)}} = 900\ ^\circ$K (by using the vaporization temperature of silicium, we overestimate $P_{\mathrm{liquid\ heating}}$), $\Delta_{\mathrm{s}}T = T_{\mathrm{melt(Si)}} - T_{\mathrm{room}} = 1400\ ^\circ$K, $c_{\mathrm{l}} = 0.91\ \mathrm{J}/(\mathrm{g}\ ^\circ\mathrm{K})$, $\tau = 1\ \mu$s, and $L = 4129\ \mathrm{J}/(\mathrm{cm}^3)$. Equation (2.14) leads to

$$\begin{cases} P_{\mathrm{diffusion}} = 0.083\ \mathrm{W}, \\ P_{\mathrm{melting}} \le 0.025\ \mathrm{W}, \\ P_{\mathrm{liquid\ heating}} \le 0.0017\ \mathrm{W}. \end{cases} \tag{2.15}$$

For short times, i.e., $t \le t_{\mathrm{p}}$, we expect both $r(t)$ and $T_\ell(t)$ to increase rapidly with time and consequently the instantaneous rates dr/dt and dT_ℓ/dt should be close to their average values. It follows that we can use the upper bounds in the second and third lines of (2.15) as estimates of P_{melting} and $P_{\mathrm{liquid\ heating}}$, which leads to $P_{\mathrm{melting}}/P_{\mathrm{liquid\ heating}} \approx 15$. $P_{\mathrm{liquid\ heating}}$ is therefore the smallest contribution to the energy balance, which is dominated by $P_{\mathrm{diffusion}}$ and P_{melting}. Combining the latter two contributions, we get

$$\frac{P_{\mathrm{liquid\ heating}}}{P_{\mathrm{diffusion}} + P_{\mathrm{melting}}} < \frac{2}{100}. \tag{2.16}$$

The energy stored by the heating fluid is therefore quite small compared to the energy transported by conduction and the energy used to melt the solid. In the following, we make the hypothesis that $P_{\mathrm{liquid\ heating}}$ can be neglected. It follows that the energy transferred by the laser during consecutive equal length time intervals can be regarded

as an *incompressible train* of equal size energy grains. Since the energy used to heat the fluid is negligible, then for each grain of energy entering the fluid at the upper surface, there is another grain of energy (emitted earlier) that exits the fluid at the liquid-solid interface. In other words, everything happens *as if* the energy entering the liquid at the top surface was transported instantaneously at the liquid-solid interface.

2.5. Derivation of the ODE based on the model hypotheses. According to hypothesis C, the melt region can be described by a hemisphere of radius r, as shown in Figure 2.1. We will therefore focus on the description of $r(t)$ as a function of the time t. According to the point heat source hypothesis A, the condition of validity of our model is

$$r(t) >> r_0, \tag{2.17}$$

where r_0 is the beam radius.

According to hypothesis D, we neglect the third term of (2.12) (on the right-hand side) to obtain

$$P = \frac{\kappa_s \, \Delta_s T}{\xi \sqrt{D \, t}} \, 2\pi r^2 + L \, 2\pi r^2 \, \frac{dr}{dt}. \tag{2.18}$$

Introducing the dimensionless quantities

$$x \equiv \, 2 \, \pi \, \frac{r}{r_0}, \quad \tau = 4 \, \pi^2 \, \frac{D \, t}{r_0^2}, \quad p = \frac{P}{D \, L \, r_0}, \tag{2.19}$$

we can rewrite (2.18) in the equivalent form

$$\frac{dx}{d\tau} = \frac{p}{x^2} - \frac{A}{\tau^{1/2}}, \tag{2.20}$$

where we introduce the dimensionless material-properties-only constant

$$A \equiv \frac{\kappa_s \, \Delta_s T}{\xi \, D \, L}. \tag{2.21}$$

For most materials, $A \approx 1$. With typical values for D (0.1 cm^2/s), L (4129 J/cm^3), $\Delta_s T$ (1400 K), r_0 (10^{-4} cm), and $P = 1$ W, we get $\tau \approx 4 \times 10^8 \, t$ and $p \approx 25$. Using the first quantity in (2.19), the constraint (2.17) implies that $x >> 2\pi$.

2.6. Initial value. The laser beam is turned on at $\tau = 0$ and is kept on afterward. The size of the melt region is zero at $\tau = 0$, and therefore it seems natural to use the initial value $x(0) \equiv x_0 = 0$. However, the differential equation (2.20) happens to be *singular* at $\tau = 0$ and $x = 0$. These singularities deserve a few comments.

First, we should stress that according to hypothesis A (point source approximation), we do not expect our model to be valid for $r = 0$. Indeed, the assumption (2.1) implies that $r(t) >> r_0$ and $t >> \frac{r_0^2}{D}$. Let us nevertheless consider our model in the limit $r \to 0$ and $t \to 0$.

The origin of the singularity at $t = 0$ is the term $\frac{\kappa_s \, \Delta_s T}{\xi \sqrt{D \, t}} \, 2\pi r^2$ in (2.18). The parameter $\Delta_s T \equiv T_m - T_{room}$ is fixed in our model, whereas in reality $\Delta_s T = 0$ for $t = 0$. Indeed, the medium does not melt instantaneously and therefore the temperature at the laser beam impact point increases rapidly from its initial value T_{room} to reach the melt value T_m. As expected, our equation does not correctly model this part of the heating process, which causes the singularity at time zero.

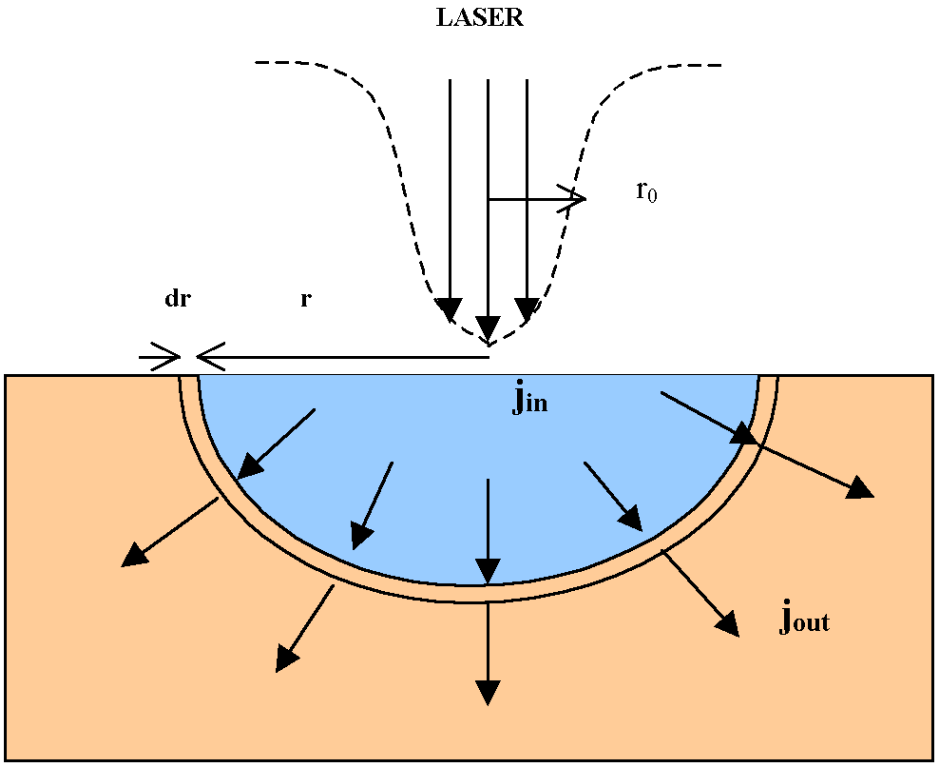


FIG. 2.1. *Schematics showing the hemispherical melt region of a semiconductor irradiated by a focused beam.*

The singularity at $r = 0$ is more physical because we expect $\frac{dr}{dt}$ to be very large for $t \approx 0$. Indeed, for $t \approx 0$ the finite power from the laser beam is transferred to a tiny hemisphere, which naturally causes the material to melt rapidly and consequently $\frac{dr}{dt}$ to be very large. In numerical simulations, we have to consider initial values such that $r_0 > 0$, even if it is not consistent with the physics of the problem. As a compromise, we will consider solutions with $r_0 > 0$ in the limit $r_0 \to 0^+$.

2.7. A preliminary simplification of the ODE. The change of variables

$$R = p^{-1}\, x \quad , \quad \theta = p^{-2}\, \tau \tag{2.22}$$

transforms (2.20) into

$$\frac{dR}{d\theta} = f(R,\theta) \equiv \frac{1}{R^2} - \frac{A}{\sqrt{\theta}}, \tag{2.23}$$

which contains only one parameter (i.e., $A \geq 0$), instead of two (i.e., A and p). In the following, we will study the ODE (2.23) with the initial condition $R(0) \equiv R_0 > 0$. The restriction $x >> 2\pi$ implies that $R >> 0.25$ (i.e., $R >> 2\pi/p$ with $p = 25$). According to (2.22), the solutions of (2.20) and (2.23) are directly related by $x(\tau, x_0) = p\ R(p^{-2}\tau, R_0)$. If $R_0 = x_0 \approx 0$, then

$$x(\tau) = p\ R(p^{-2}\tau). \tag{2.24}$$

In the following, we will assess the accuracy of approximate analytical solutions by comparison with numerical solutions, for which we use $A = 0.75$, which is the value of A corresponding to silicium.

3. Basic considerations.

3.1. Separability and integrating factor. Equation (2.23) is a nonlinear first order nonautonomous ODE. If $A = 0$, then (2.23) becomes $\frac{dR}{d\theta} = \frac{1}{R^2}$, which is *separable*, and the solution is

$$R(\theta) = (3\,\theta + R_0^3)^{1/3}. \tag{3.1}$$

However, if $A > 0$, then (2.23) is *not separable*. Moreover, an integrating factor that depends only on R or only on θ does not exist.

3.2. Sign of the derivative and direction field. Equation (2.23) can be rewritten in the equivalent form

$$\frac{dR}{d\theta} = -\frac{A}{R^2\sqrt{\theta}}\left(R + \frac{\theta^{1/4}}{\sqrt{A}}\right)\left(R - \frac{\theta^{1/4}}{\sqrt{A}}\right), \tag{3.2}$$

which shows that the function

$$\rho_0(\theta) \equiv \frac{\theta^{1/4}}{\sqrt{A}} \tag{3.3}$$

plays a special role. On one hand, $\rho_0(\theta)$ satisfies $f(\rho_0(\theta), \theta) = 0$. On the other hand, it follows from (3.2) that

$$\begin{cases} dR/d\theta < 0 & \text{if} \quad R > \rho_0(\theta), \\ dR/d\theta > 0 & \text{if} \quad R < \rho_0(\theta). \end{cases} \tag{3.4}$$

The inequalities (3.4) suggest that orbits have a tendency to remain close to $\rho_0(\theta)$, i.e., that $\rho_0(\theta)$ *is an asymptotic solution* for $\theta \to \infty$. This is indeed the case, in the sense that $\dot{\rho}_0(\theta) - f(\rho_0(\theta), \theta) = \frac{\theta^{-3/4}}{4\sqrt{A}} \to 0$ as $\theta \to \infty$. The *direction field* (DF) of (2.23) was plotted in Figure 3.1 for $A = 0.75$ and $0 \le \theta \le 1$. The DF is *horizontal* on the solid curve $R = \rho_0(\theta)$. The DF is pointing downward along the R-axis, which indicates that orbits dive down for $\theta \approx 0$ and $R(0) > 0$. However, the large θ behavior of the field is consistent with an increasing $R(\theta)$ as θ increases.

3.3. Singularities and numerical solutions. Equation (2.23) is *singular* at $R = 0$ and $\theta = 0$, which is problematic for initial values of the form $R(0) = R_0 > 0$. The singularity at $\theta = 0$ can be circumvented with the change of variable $s = \sqrt{\theta}$, which transforms (2.23) into

$$\frac{dR}{ds} = 2\left(-A + \frac{s}{R^2}\right). \tag{3.5}$$

Equation (3.5) is no longer singular at $s = 0$ but remains singular at $R = 0$. We obtained our numerical solutions[1] by solving (3.5) and then by replacing s by $\sqrt{\theta}$. With this method, we obtained several numerical solutions corresponding to different values of $R_0 > 0$ (Figure 3.2).

[1] In this paper, numerical solutions were obtained with the *Mathematica* function NDSolve, which switches between a nonstiff Adams method and a stiff Gear method.

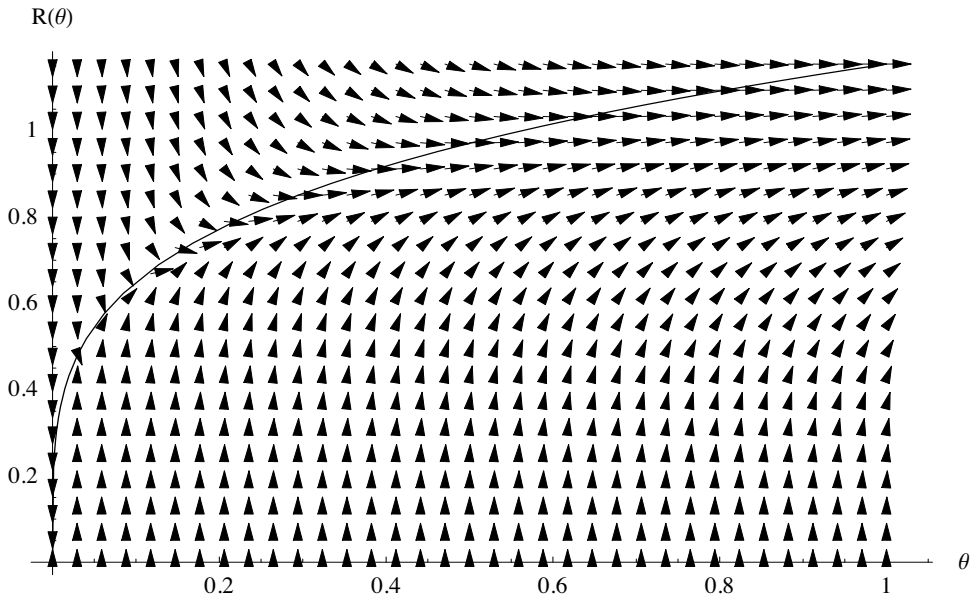


FIG. 3.1. *DF of* (2.23) *for* $A = 0.75$. *The solid curve represents the set of points for which the DF is horizontal, i.e.,* $R = \rho_0(\theta) \equiv \theta^{1/4}/\sqrt{A}$.

Considering Figure 3.2, the first important observation is that $R(\theta, R_0)$ is virtually independent of R_0 for large enough θ. This is fortunate because we did not know a priori which value to choose for R_0. Figure 3.2 informs us that any $R_0 < 0.1$ leads essentially to the same solution for $\theta >> 0.002$. The second observation is that the validity condition $R >> 0.25$ corresponds approximately to the *validity range* $\theta >> \theta_1 \equiv 0.002$.

3.4. Anomalous behavior of $R(\theta)$ around $\theta = 0$. The physics of this problem implies that the melt region expands with time. However, as clearly seen on the DF, all solutions with $R_0 > 0$ dive down in the neighborhood of $\theta = 0$ before eventually going up again. This peculiar behavior occurs in the region where the ODE is not valid. To be cautious, it is important to see if this anomaly can overlap the validity range $\theta > \theta_1$ of the ODE.

We analyzed the behavior of $R(\theta)$ as $\theta \to 0$ in Appendix A and found that

$$\begin{cases} \text{If } R_0 > 0, \ R(\theta) \sim R_0 - 2\ A\ \sqrt{\theta} + \frac{1}{R_0^2}\ \theta, \\ \text{If } R_0 \approx 0, \ \ R(\theta) \sim (3\ \theta)^{1/3} \end{cases} \tag{3.6}$$

as $\theta \to 0$. Solutions with $R_0 > 0$ and $R_0 \approx 0$ are qualitatively different. Indeed, if $R_0 \approx 0$, then $R(\theta)$ *increases* for all $\theta \geq 0$, as it should. However, if $R_0 > 0$, then $R(\theta)$ *decreases*, reaches a minimum around $\theta_{\min} = A^2\ R_0^4$, and then increases ($R(\theta)$ is U-shaped). If we consider, for instance, a solution obtained with $R_0 = 0.001$, then we get $\theta_{\min} \approx 5.6 \times 10^{-13} << \theta_1 = 0.002$. Hence the anomalous behavior (3.6) occurs at very short times and does not overlap the validity range $\theta > \theta_1$ of the ODE.

3.5. Perturbation solution for A small. The solution for $A = 0$ is known exactly, i.e., $R(\theta) = (R_0^3 + 3\ \theta)^{1/3}$, and we assume that $0 \leq A \leq 1$. In this context, it is appropriate to look for a perturbation solution that would be valid for A small. If

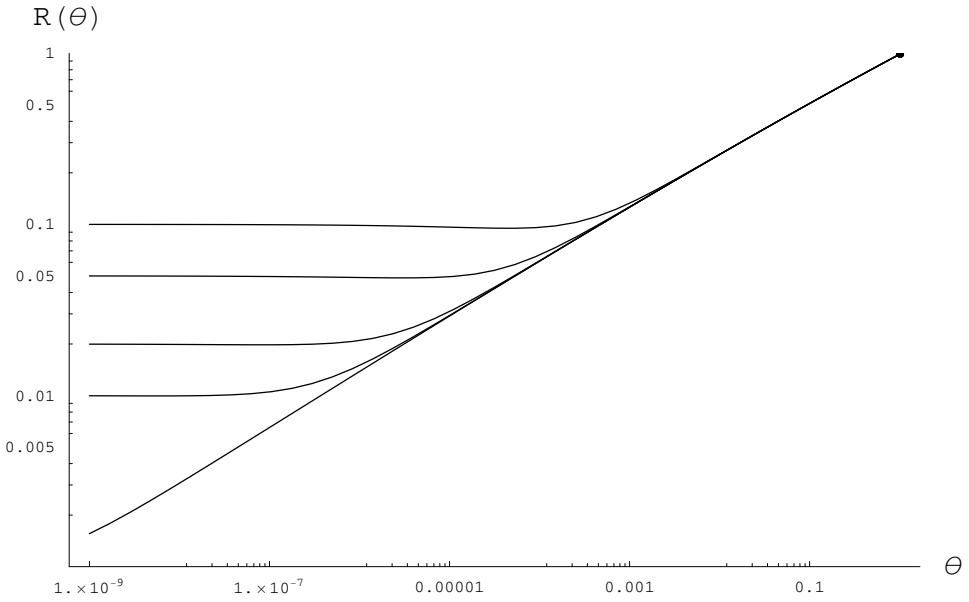


FIG. 3.2. *Numerical solutions for the initial values* $R(0) = (0.001, 0.01, 0.02, 0.05, 0.1)$.

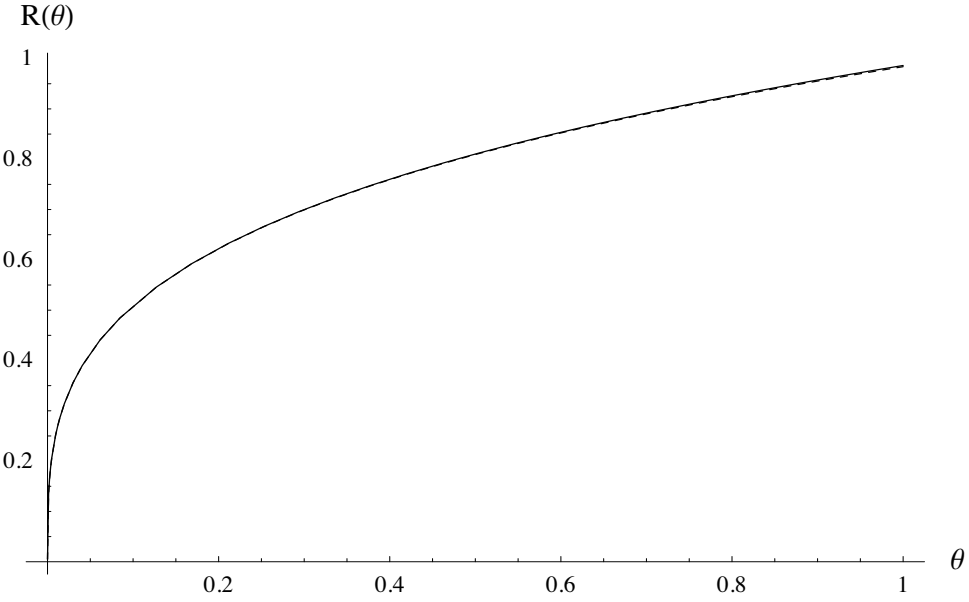


FIG. 3.3. *The solid curve is a numerical solution with* $R(0) = 0.001$ *and* $A = 0.75$*, while the dotted curve is the perturbation solution* (3.8)*. The curves are superposed; i.e., the agreement is excellent.*

we search for a solution of the form

$$R(\theta) = \sum_{n=0}^{\infty} A^n \; R_n(\theta), \tag{3.7}$$

where the R_ns are unknown functions and $R(0) = 0$, then it is straightforward to show (see Appendix B) that the perturbation method leads to the series

$$R(\theta) = (3\ \theta)^{1/3} - \frac{6}{7}\ \sqrt{\theta}\ A + \frac{3}{49}\ (3\ \theta)^{2/3}\ A^2 - \frac{8\ 3^{1/3}}{343}\ \theta^{5/6}\ A^3 - \frac{711}{12005}\ \theta\ A^4 + \frac{4824\ 3^{2/3}}{924385}\ \theta^{7/6}\ A^5 + O(A)^6. \tag{3.8}$$

As shown in Figure 3.3, an excellent agreement of (3.8) with numerical solutions is obtained for $0 \leq \theta \leq 1$.

3.6. Dependence of the solution on the parameter A. We notice that (3.8) can be written in the form $R(\theta) = \sum_{n=0}^{5} c_n\ (\theta^{1/6})^{n+2}\ A^n$, where the c_ns are real coefficients. Factorizing $\theta^{1/3}$ yields $R(\theta) = \theta^{1/3}\ \sum_{n=0}^{5} c_n\ (A\ \theta^{1/6})^n$, which suggests that the solution with $R_0 = 0$ has the general form

$$R(\theta) = \theta^{1/3}\ F(A\ \theta^{1/6}), \tag{3.9}$$

where F is an unknown function. Substituting (3.9) into (2.23) and using the change of variable $u = A\ \theta^{1/6}$ lead to

$$\frac{dF}{du} = \frac{6\ (1 - u\ F^2) - 2\ F^3}{u\ F^2}. \tag{3.10}$$

Equation (3.10) does not depend explicitly on A, and therefore (3.9) is indeed correct in general. The initial condition for (3.10) is $F(0) = 3^{1/3}$, which follows from $R(\theta) \sim (3\ \theta)^{1/3}$ as $\theta \to 0$. Equation (3.10) is a *key equation* that allows us to recover the MacLaurin series of $F(u)$ directly, i.e., without using the perturbation method. Consider, for instance, $F'(0)$. According to (3.10), $F'(0)$ has an indeterminate form $0/0$. However, using l'Hôspital's rule yields $F'(0) = \lim_{u\to 0} \frac{-6\ (F^2 + 2\ u\ F\ F') - 6\ F^2\ F'}{F^2 + 2\ u\ F\ F'} = -6 - 6\ F'(0) \Rightarrow F'(0) = -6/7$, which is correct according to (3.8). Higher order derivatives can also be obtained to recover the whole expansion, i.e., $F(u) = 3^{1/3} - \frac{6}{7}\ u + \frac{9\ 3^{2/3}}{49}\ u^2 - \frac{8\ 3^{1/3}}{343}\ u^3 - \frac{711}{12005}\ u^4 + \frac{4824\ 3^{2/3}}{924385}\ u^5 + O(u)^6$.

4. Asymptotic behavior for $\theta \to \infty$. In the spirit of the geometrical methods of Poincaré, we will try to locate the orbit $R(\theta)$ by examining its distance with respect to a reference curve. We have seen previously that the curve $R = \rho_0(\theta)$, defined by (3.3), is an asymptotic solution as $\theta \to \infty$. We will therefore choose $R = \rho_0(\theta)$ as our reference curve. We consider the time evolution of the distance function $U(\theta)$ defined by

$$U(\theta) \equiv \frac{1}{2}\left(R(\theta) - \rho_0(\theta)\right)^2. \tag{4.1}$$

A time derivative gives $\frac{dU}{d\theta} = (R - \rho_0)(\dot{R} - \dot{\rho}_0) = (R - \rho_0)(f(R,\theta) - \dot{\rho}_0)$. The factor $f(R,\theta) - \dot{\rho}_0$ has two roots $R = \pm\ \rho_1(\theta)$, which yields the factorization

$$\frac{dU}{d\theta} = -\ \frac{(1 + 4\ A^{3/2}\ \theta^{1/4})}{4\ \sqrt{A}\ R^2\ \theta^{3/4}}\ (R + \rho_1(\theta))(R - \rho_0(\theta))(R - \rho_1(\theta)), \tag{4.2}$$

where

$$\rho_1(\theta) = \frac{2\ A^{1/4}\ \theta^{3/8}}{\sqrt{1 + 4\ A^{3/2}\ \theta^{1/4}}}. \tag{4.3}$$

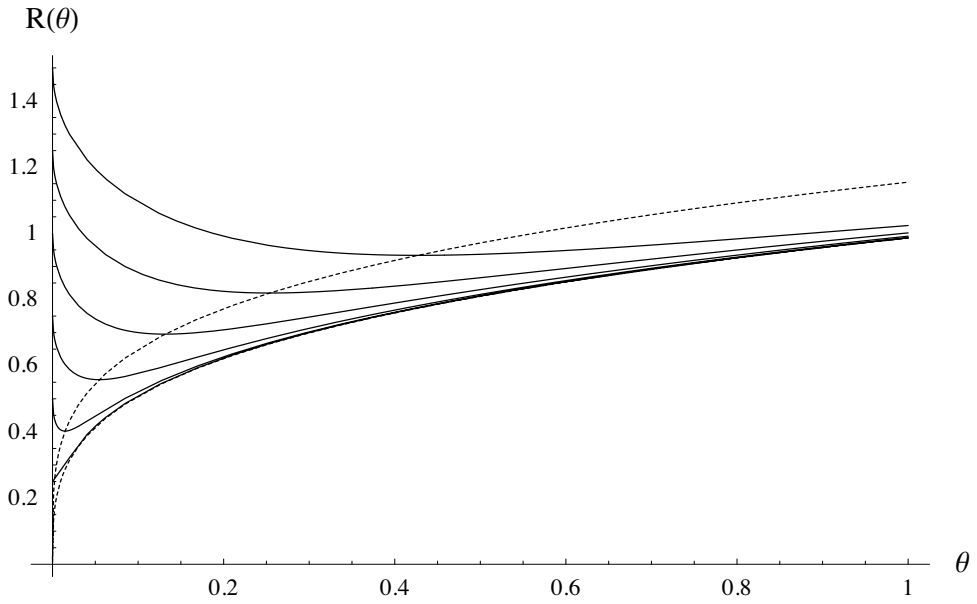


FIG. 4.1. *Dotted lines: Functions $\rho_0(\theta)$ (top) and $\rho_1(\theta)$ (bottom). Solid lines are orbits with different values of $R(0) > 0$, i.e., $(0.01, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5)$. The solutions that correspond to $R(0) = 0.01$ and 0.01 are superposed. Here $A = 0.75$.*

The functions $\rho_0(\theta)$ and $\rho_1(\theta)$, which are plotted in Figure 4.1 for $A = 0.75$, satisfy

$$\rho_0(\theta) > \rho_1(\theta) > 0 \tag{4.4}$$

for all $\theta > 0$ as long as $A > 0$. It follows from (4.2) and (4.4) that

$$\begin{cases} \frac{dU}{d\theta} < 0 & \text{if} \quad R > \rho_0(\theta) \ \text{ or } \ R < \rho_1(\theta), \\ \frac{dU}{d\theta} > 0 & \text{if} \quad \rho_0(\theta) < R < \rho_1(\theta), \\ \frac{dU}{d\theta} = 0 & \text{if} \quad R = \rho_0(\theta) \text{ or } R = \rho_1(\theta). \end{cases} \tag{4.5}$$

In other words, the *crescent-shaped zone* bounded by the curves $R = \rho_0(\theta)$ and $R = \rho_1(\theta)$ is *attractive* for all orbits that are outside the crescent. If $R(0) > 0$, then the orbit is in the region $R > \rho_0(\theta)$ initially, i.e., outside the crescent. If $R(0) = 0$, then according to (3.6) we have $R(\theta) \sim (3\ \theta)^{1/3}$ as $\theta \to 0$, and consequently the orbit lies between the two curves $\rho_0(\theta)$ and $\rho_1(\theta)$ initially. Indeed, $\rho_1(\theta) \sim 2\ A^{1/4}\ \theta^{3/8}$ as $\theta \to 0$, and one shows easily that $2\ A^{1/4}\ \theta^{3/8} < (3\ \theta)^{1/3} < \theta^{1/4}/\sqrt{A}$ as $\theta \to 0$. In summary, the orbit is initially located either *above* the reference curve $R = \rho_0(\theta)$ (for $R_0 > 0$), or *between* $\rho_0(\theta)$ and $\rho_1(\theta)$ (for $R_0 = 0$).

If the orbit starts *above* $R = \rho_0(\theta)$, then it follows from the first line of (4.5) that $R(\theta)$ gets closer to $\rho_0(\theta)$ as θ increases. This behavior is illustrated by the numerical solutions displayed in Figure 4.1. All orbits with $R(0) > 0$ *cross* the curve $R = \rho_0(\theta)$ because the DF is horizontal on this curve. Once inside the *crescent-shaped* region $\rho_0(\theta) < R < \rho_1(\theta)$, then the second line of (4.5) implies that the orbit $R(\theta)$ *goes away* from the upper boundary of the crescent $R = \rho_0(\theta)$. We are going to prove that this orbit *cannot cross the bottom curve* $R = \rho_1(\theta)$ because this would imply $\dot{U}(\theta) > 0$, which is not possible *on* and *below* this curve according to (4.5).

First,

$$\dot{U} = (\rho_0 - R)(\dot{\rho}_0 - \dot{R}) = (\rho_0 - R)\ \dot{D}, \tag{4.6}$$

where $D(\theta) \equiv \rho_0(\theta) - R(\theta)$. The distance $D(\theta)$ satisfies

$$\dot{D} = \dot{\rho}_0(\theta) - \dot{R}(\theta) = \dot{\rho}_0(\theta) - \dot{\rho}_1(\theta) + \dot{\rho}_1(\theta) - \dot{R}(\theta) = \dot{D}_0 + \dot{\rho}_1(\theta) - \dot{R}(\theta), \tag{4.7}$$

where $D_0(\theta) \equiv \rho_0(\theta) - \rho_1(\theta)$ is the distance between the two curves. It can be shown that $\dot{D}_0 > 0$ for all $\theta > 0$; i.e., the distance separating the two curves increases. If we assume that an orbit crosses the curve $R = \rho_1(\theta)$ *from above*, then this orbit has to satisfy $\dot{R} < \dot{\rho}_1 \Rightarrow \dot{\rho}_1(\theta) - \dot{R}(\theta) > 0$ *at the crossing point*. Since $\dot{D}_0 > 0$ and $\dot{\rho}_1(\theta) - \dot{R}(\theta) > 0$ at the crossing point, (4.7) implies that $\dot{D} > 0$. Since $R < \rho_0$ and $\dot{D} > 0$ at the crossing point, it follows from (4.6) that $\dot{U} > 0$, which is not possible *on* or *below* the curve $R = \rho_1(\theta)$ according to (4.5). Hence there is no crossing point.

The orbit *must therefore remain within the crescent*, while going away from the upper curve $R = \rho_0(\theta)$. Orbits that start *within the crescent*, and, in particular, the orbit with $R_0 = 0$, also remain within the crescent for all $\theta > 0$ (for the same reasons). It can be shown that $\lim_{\theta\to\infty} \rho_0(\theta) - \rho_1(\theta) = 1/(8\ A^2)$. Hence the orbit is *sandwiched between two curves that are separated by an asymptotically finite constant distance* and goes away from the upper curve. Numerical solutions (Figure 4.1) indicate that the limit orbit is much closer to $\rho_1(\theta)$ (bottom curve) than to $\rho_0(\theta)$ for $A = 0.75$. It follows that a possible asymptotic behavior for the orbit is

$$R(\theta) \sim \frac{2\ A^{1/4}\ \theta^{3/8}}{\sqrt{1 + 4\ A^{3/2}\ \theta^{1/4}}}\ +\ \mathrm{C}(A, R_0)\ \ \text{as}\ \ \theta \to \infty, \tag{4.8}$$

where $\mathrm{C}(A, R_0) << 1/(8\ A^2)$. Numerical solutions suggest that $\mathrm{C}(A, R_0)$ could be independent of R_0. $\rho_1(\theta)$ satisfies (2.23) asymptotically as $\theta \to \infty$. Indeed, as $\theta \to \infty$, we have $\dot{\rho}_1(\theta) - f(\rho_1(\theta), \theta) \sim \frac{1}{4\sqrt{A}\ \theta^{3/4}} \to 0$. Finally, let us stress that (4.8) has exactly the functional form (3.9), with $F(u) = (2\ u^{1/4})/(\sqrt{1 + 4\ u^{3/2}})$.

5. More accurate asymptotic solutions for $\theta \to \infty$. The asymptotic solution $\rho_1(\theta)$ was shown to be an improvement on the first guess $\rho_0(\theta)$. One may hope that a similar procedure could allow us to further improve the solution. Let us therefore consider the distance function

$$V(\theta) \equiv \frac{1}{2}\left(R(\theta) - \rho_1(\theta)\right)^2. \tag{5.1}$$

Taking the time derivative and factorizing as previously yield

$$\dot{V}(\theta) = -a(\theta)(R + \rho_2(\theta))(R - \rho_1(\theta))(R - \rho_2(\theta)), \tag{5.2}$$

where $\rho_2(\theta) > 0$ and $a(\theta) > 0$ are given by

$$\begin{aligned} \rho_2(\theta) &= \frac{2\ \left(1 + 4\ A^{3/2}\ \theta^{1/4}\right)^{3/4}\ \theta^{5/16}}{\sqrt{3\ A^{1/4} + 4\ A\ \left(1 + 4\ A^{3/2}\ \theta^{1/4}\right)^{3/2}\ \theta^{1/8} + 8\ A^{7/4}\ \theta^{1/4}}}, \\ a(\theta) &= \frac{3\ A^{1/4} + 4\ A\ \left(1 + 4\ A^{3/2}\ \theta^{1/4}\right)^{3/2}\ \theta^{1/8} + 8\ A^{7/4}\ \theta^{1/4}}{4\ R^2\ \left(1 + 4\ A^{3/2}\ \theta^{1/4}\right)^{3/2}\ \theta^{5/8}}. \end{aligned} \tag{5.3}$$

It can be shown that $\rho_2(\theta) > \rho_1(\theta)$ for all $\theta > 0$, which implies with (5.2) that

$$\begin{aligned} &\dot{V}(\theta) < 0 \text{ if } R > \rho_2(\theta) \text{ or } R < \rho_1(\theta), \\ &\dot{V}(\theta) > 0 \text{ if } \rho_1(\theta) < R < \rho_2(\theta); \end{aligned} \tag{5.4}$$

i.e., the crescent region bounded by the curves $R = \rho_1(\theta)$ and $R = \rho_2(\theta)$ is *attractive*. The situation is therefore similar to the previous case in the sense that we have identified another crescent-shaped region which is *attractive* for orbits that are outside this region. However, the arguments that we used in section 4 to prove that orbits entering this crescent are trapped no longer holds in this case. Indeed, the problem is that $\dot{\rho}_2 - \dot{\rho}_1 < 0$; i.e., the two curves get closer to each other as θ increases (which can be shown numerically for $A = 0.75$).

We cannot establish with the same argument that the orbit stays inside this new crescent, but we can at least claim that $\rho_2(\theta)$ is an asymptotic solution because $\rho_2(\theta) \sim \theta^{1/4}/\sqrt{A}$ as $\theta \to \infty$. We may ask if $\rho_2(\theta)$ is a better approximation of $R(\theta)$ than $\rho_1(\theta)$ for $0 \le \theta \le 1$. Numerical solutions (Figure 5.1) indicate that the orbits with $R_0 \approx 0$ are closer to $\rho_2(\theta)$ than to $\rho_1(\theta)$. In Figure 5.1, a low value of A was used to get a clearly visible spacing between the curves bounding the crescent. Indeed, for $A = 0.75$ the curves $\rho_1(\theta)$ and $\rho_2(\theta)$ are almost superposed. We can therefore propose the asymptotic solution

$$R(\theta) \sim \frac{2\ (1 + 4\ A^{3/2}\ \theta^{1/4})^{3/4}\ \theta^{5/16}}{\sqrt{3\ A^{1/4} + 4\ A\ (1 + 4\ A^{3/2}\ \theta^{1/4})^{3/2}\ \theta^{1/8} + 8\ A^{7/4}\ \theta^{1/4}}} \tag{5.5}$$

as $\theta \to \infty$, which is more accurate than $\rho_1(\theta)$ for $0 \le \theta \le 1$. As shown in Figure 5.2, $\rho_2(\theta)$ appears to be a good asymptotic solution. Moreover, the range of validity of $\rho_2(\theta)$ is broader if $R(0) \approx 0$, which is precisely the limit we are interested in.

Remark. The iterative process that allowed us to find $\rho_1(\theta)$ and $\rho_2(\theta)$, starting with $R = \rho_0(\theta)$, can be summarized as follows. In both cases, the new curve $R = \rho_{n+1}(\theta)$ bounding the crescent zone is the positive root of the equation

$$f(\rho_{n+1}, \theta) = \dot{\rho}_n(\theta), \tag{5.6}$$

where $\rho_n(\theta)$ is the previous curve. Using (2.23), (5.6) leads to the iteration formula

$$\rho_{n+1}(\theta) = \frac{\theta^{1/4}}{\sqrt{\sqrt{\theta}\ \dot{\rho}_n(\theta) + A}}. \tag{5.7}$$

Roughly speaking, (5.7) is a kind of *backward Picard iteration*, because we iterate a derivative instead of an integration. We notice that $\rho_n(0) = 0$ for all n and consequently the initial value $R(0) = 0$ is conserved exactly during iteration. Using the initial value $\dot{\rho}_{-1}(\theta) = 0$, we can iterate (5.7) to get successively $\rho_0(\theta)$, $\rho_1(\theta)$, and $\rho_2(\theta)$. It might be possible to generalize this iterative process to find asymptotic solutions for other ODEs.

6. Conclusions. The only *exact* result that we derived about the solution of the ODE (2.23) with $R(0) \approx 0$ is the functional form (3.9). Using (2.22), (3.9) leads to

$$x(\tau, p) = (p\ \tau)^{1/3}\ F(A\ p^{-1/3}\ \tau^{1/6}), \tag{6.1}$$

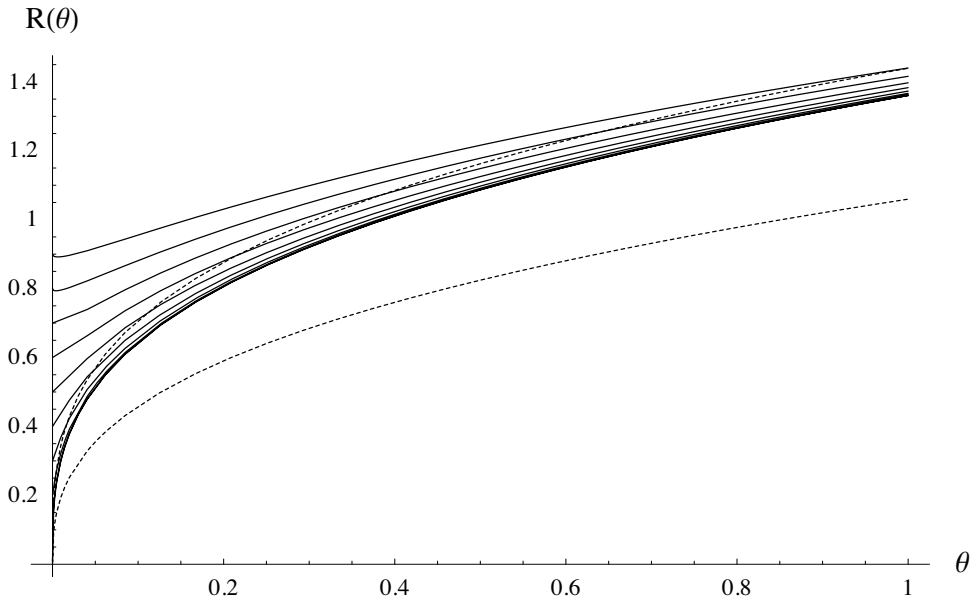


FIG. 5.1. *Results obtained with* $A = 0.1$. *The solid curves are numerical solutions obtained with* $R_0 = (0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$. *Solutions with* $R_0 = 0.01,\ 0.1$ *are superposed. The dashed curves are* $R = \rho_1(\theta)$ *(bottom) and* $R = \rho_2(\theta)$ *(top). In the range* $0 \leq \theta \leq 1$, *orbits that enter the crescent remain in the crescent, move away from* $R = \rho_2(\theta)$ *as expected, but remain nevertheless closer to* $R = \rho_2(\theta)$ *than to* $R = \rho_1(\theta)$.

where F satisfies the ODE (3.10). Equation (6.1) could be useful to represent experimental data obtained with varying beam power p.

To compare the three approximate analytical solutions obtained in this paper, we plotted in Figure 6.1 (top) their relative errors (in percent) with respect to the numerical solution, using $A = 0.75$ and $R_0 = 0.001$. The perturbation solution (3.8) is the most accurate over most of the range, except for $\theta > 0.4$, where the asymptotic solution $R = \rho_2(\theta)$ is more accurate (using more terms in the perturbation solution would increase its accuracy). The asymptotic solution $R = \rho_2(\theta)$ is the second best approximation and its error has a minimum around $\theta = 0.004$. The third best approximation is $R = \rho_1(\theta)$.

The solution $R = \rho_1(\theta)$ is attractive because of its greatest simplicity. In spite of its slightly lower accuracy, we are going to see that this solution is the most useful in practice. First, we should first remember that the ODE studied in this paper is derived from an approximate model. In particular, the melt region is not exactly hemispherical and the model ODE is valid only for $\theta >> \theta_1 = 0.002$, for, say, $\theta \geq 0.02$. Knowing that experimental errors lie in the range of 5–10% [3], we can therefore conclude that the error in the approximate solutions of the model equations is smaller than the error in experimental measurement. From this standpoint, the most interesting solution is the one that offers a good compromise between simplicity and accuracy. From this perspective, the asymptotic solution $R(\theta) = \rho_1(\theta)$ is the simplest and has an error smaller than 3% in the validity range $\theta \geq 0.02$. More importantly, *this error decreases as* θ *increases*, which is not the case for the perturbation solution. To introduce explicitly the dependence on the beam power p, we can combine (2.24) and

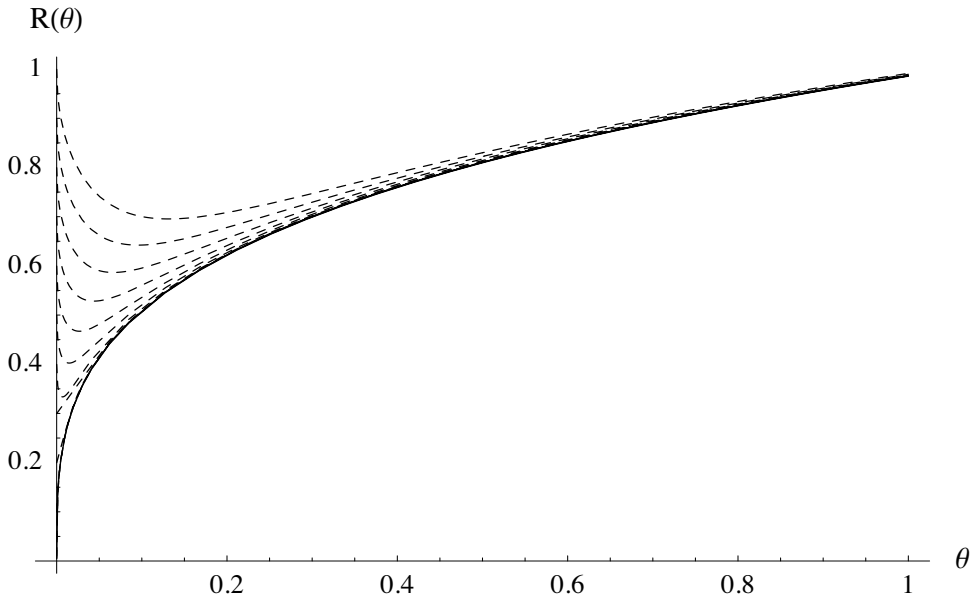


FIG. 5.2. *Results obtained with $A = 0.75$. The dashed curves are numerical solutions obtained with different positive values of $R(0) > 0$. The solid curve is the asymptotic solution $R = \rho_2(\theta)$.*

(4.3) to obtain

$$x(\tau, p) = \frac{2\ A^{1/4}\ p^{1/4}\ \tau^{3/8}}{\sqrt{1 + 4\ A^{3/2}\ p^{-1/2}\ \tau^{1/4}}}, \tag{6.2}$$

which gives an error smaller than 3% for $\tau \geq 12.5$, which corresponds to $t \geq 0.03\ \mu s$. A satisfactory comparison of the model (6.2) with experimental data is presented in [3].

We will conclude on a note about the geometrical method that we used to derive the asymptotic solution $R = \rho_1(\theta)$. This possibly original method leads us to three increasingly accurate asymptotic solutions $\rho_0(\theta)$, $\rho_1(\theta)$, and $\rho_2(\theta)$ that could be obtained by iterating the formula

$$f(\rho_{n+1}, \theta) = \frac{d}{d\theta}\ \rho_n(\theta) \tag{6.3}$$

starting with $\rho_{-1}(\theta) = 0$. In (6.3), f is the function that defines the ODE, i.e., $\frac{dR}{d\theta} = f(R, \theta)$. Roughly speaking, (6.3) is a kind of *backward Picard iteration*, because we iterate a derivative instead of an integration. It would be interesting to see if the iterative process (6.3) could be generalized to find approximate solutions for other ODEs.

Appendix A. Asymptotic behavior of $R(\theta)$ as $\theta \to 0$ for $A > 0$.

To study $R(\theta)$ around $\theta = 0$, we will make the hypothesis

$$R(\theta) \sim R_0 + c\ \theta^{\alpha} \tag{A.1}$$

as $\theta \to 0$, where $\alpha > 0$ and c is a constant. Substituting (A.1) into (2.23) yields

$$\frac{c\ \alpha}{\theta^{1-\alpha}} = -\frac{A}{\theta^{1/2}} + \frac{1}{(R_0 + c\ \theta^{\alpha})^2}. \tag{A.2}$$

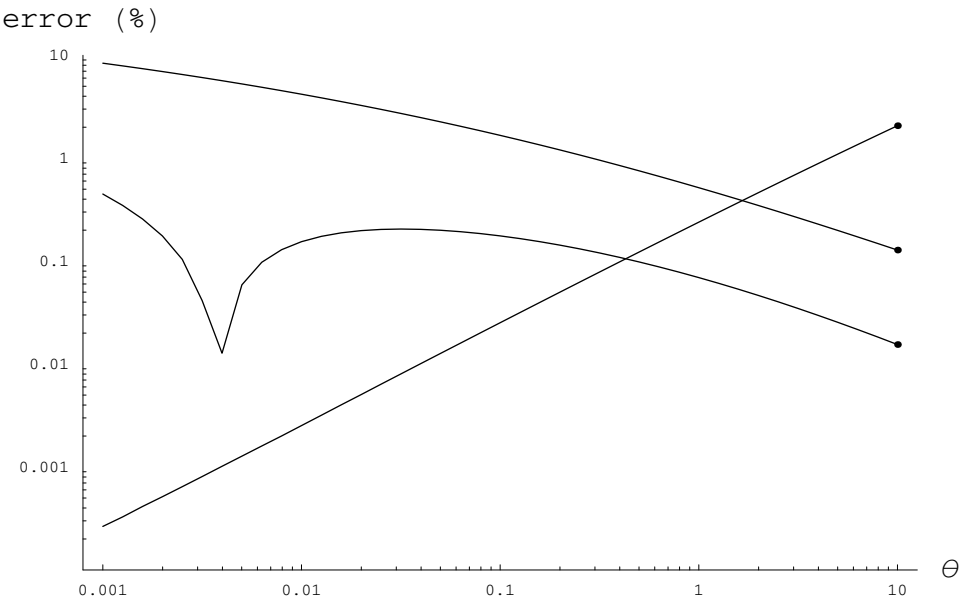


FIG. 6.1. *Comparison of the approximate solutions for* $A = 0.75$ *and* $R_0 = 0.001$. *We plotted the relative error in % with respect to the numerical solution, versus* θ. *Top-left: Error on* $R = \rho_1(\theta)$. *Center-left: Error on* $R = \rho_2(\theta)$. *Bottom-left: Error on the perturbation solution* (3.8).

If $R_0 > 0$, then the term $\frac{A}{\theta^{1/2}}$ on the right-hand side of (A.2) dominates as $\theta \to 0$. It follows that $c\ \alpha = -A$ and $1 - \alpha = 1/2 \Rightarrow \alpha = 1/2$, which in turn implies $c = -2A$. Hence, we have the asymptotic behavior

$$\text{If } R_0 > 0\ ,\ R(\theta) \sim R_0 - 2\ A\ \sqrt{\theta} \tag{A.3}$$

as $\theta \to 0$; i.e., *the orbit dives downward* before it eventually returns to an increasing regime. If $R_0 = 0$, (A.2) becomes

$$\frac{c\ \alpha}{\theta^{1-\alpha}} = -\frac{A}{\theta^{1/2}} + \frac{1}{c^2\ \theta^{2\alpha}}. \tag{A.4}$$

One must examine two cases. First, if we assume that $2\alpha > 1/2 \Rightarrow \alpha > 1/4$, then the term $1/(c^2\ \theta^{2\alpha})$ (right-hand side of (A.4)) dominates as $\theta \to 0$, and therefore $c\ \alpha = 1/c^2$ and $1 - \alpha = 2\ \alpha \Rightarrow \alpha = 1/3$, which also implies $c = 3^{1/3}$. Second, if we assume instead that $2\alpha < 1/2 \Rightarrow \alpha < 1/4$, then we must have $1 - \alpha = 1/2 \Rightarrow \alpha = 1/2$, which contradicts our assumption $\alpha < 1/4$. Hence, for $R_0 = 0$, we have the following asymptotic behavior:

$$\text{If } R_0 = 0,\ R(\theta) \sim (3\ \theta)^{1/3} \tag{A.5}$$

as $\theta \to 0$, which is consistent with (3.1).

Comparing (A.3) with (A.5), we see that solutions are qualitatively different for $R_0 > 0$ and $R_0 = 0$. Indeed, $R(\theta)$ *increases* if $R_0 = 0$, whereas it *decreases* if $R_0 > 0$. According to the DF, $R(\theta)$ should return rapidly to an increasing regime even if $R_0 > 0$. To see how this return occurs, we may try to find a better approximation of $R(\theta)$ around $\theta = 0$ with a MacLaurin expansion. Equation (3.5) implies that the derivatives of $R(s)$ are well defined at $s = 0$ as long as $R_0 > 0$, and therefore

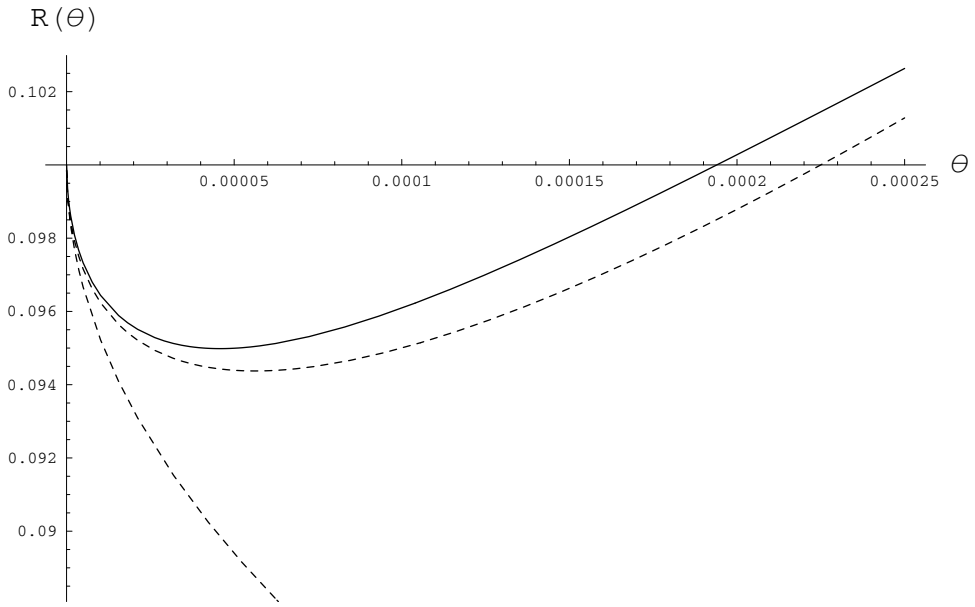


FIG. A.1. *Behavior of $R(\theta)$ around $\theta = 0$ for $R_0 = 0.1$. Here $A = 0.75$, and $\theta_{\min} \approx 0.000056$. The solid line is a numerical solution, and the dashed curves are $R_0 - 2\ A\ \sqrt{\theta}$ (going down) and $R_0 - 2\ A\ \sqrt{\theta} + \frac{1}{R_0^2}\ \theta$ (going back up).*

$R(s) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{d^n R}{ds^n}(0)\ s^n$. Using (3.5) to compute the derivatives $\frac{d^n R}{ds^n}(0)$ leads to $R(s) = R_0 - 2\ A\ s + \frac{s^2}{R_0^2} + \frac{8\ A}{3\ R_0^3}\ s^3 + (\frac{6\ A^2}{R_0^4} - \frac{1}{2\ R_0^5})\ s^4 + O(s)^5$, and replacing s by $\theta^{1/2}$ gives

$$R(\theta) = R_0 - 2\ A\ \theta^{1/2} + \frac{\theta}{R_0^2} + \frac{8\ A}{3\ R_0^3}\ \theta^{3/2} + O(\theta)^2. \tag{A.6}$$

The first three terms of (A.6) give the approximation $\tilde{R}(\theta) \equiv R_0 - 2\ A\ \theta^{1/2} + \frac{\theta}{R_0^2}$. The minimum of $\tilde{R}(\theta)$ occurs at $\theta_{\min} = A^2\ R_0^4$. As shown in Figure A.1, $\tilde{R}(\theta)$ describes fairly well the behavior of $R(\theta)$ around $\theta = 0$ for $R_0 > 0$.

Appendix B. Perturbation solution.

Substituting (3.7) into (2.23) yields

$$\sum_{n=0}^{\infty} A^n\ \dot{R}_n(\theta) = -\frac{A}{\sqrt{\theta}} + \frac{1}{(\sum_{n=0}^{\infty} A^n\ R_n(\theta))^2}. \tag{B.1}$$

Expanding the rightmost term of (B.1) in Taylor series around $A = 0$ to order 2 yields

$$\dot{R}_0 + \dot{R}_1\ A + \dot{R}_2\ A^2 = -\frac{A}{\sqrt{\theta}} + \frac{1}{R_0^2} - \frac{2}{R_0^3}\ R_1\ A + \left(\frac{3R_1^2}{R_0^4} - \frac{2}{R_0^3}\right)\ A^2, \tag{B.2}$$

where $\dot{f} \equiv \frac{df}{d\theta}$ for any function f. Identifying the terms of order 0, 1, and 2 yields

$$\begin{aligned} \dot{R}_0 &= \frac{1}{R_0^2}, \\ \dot{R}_1 &= -\frac{1}{\sqrt{\theta}} - \frac{2}{R_0^3}\, R_1, \\ \dot{R}_2 &= \frac{3R_1^2}{R_0^4} - \frac{2}{R_0^3}\, R_2. \end{aligned} \tag{B.3}$$

We will solve (B.3) with the initial value $R(0) = 0$. The solution of the first equation of (B.3) is $R_0(\theta) = (3\ \theta)^{1/3}$. Substituting $R_0(\theta)$ into the second equation of (B.3), we get

$$\dot{R}_1 + \frac{2}{3\ \theta}\, R_1 = -\frac{1}{\sqrt{\theta}}. \tag{B.4}$$

Fortunately, (B.4) is a nonhomogeneous *linear equation*. Its general solution is $R_1(\theta) = -\frac{6}{7}\ \sqrt{\theta} + \frac{\mathrm{C}}{\theta^{2/3}}$, where C is an arbitrary constant. The initial condition $R_1(0) = 0$ then imposes $\mathrm{C} = 0$ so that

$$R_1(\theta) = -\frac{6}{7}\ \sqrt{\theta}. \tag{B.5}$$

Using previous results for $R_0(\theta)$ and $R_1(\theta)$, the third equation of (B.3) becomes

$$\dot{R}_2 + \frac{2}{3\ \theta}\, R_2 = \frac{k_1}{\theta^{1/3}}, \tag{B.6}$$

where $k_1 = 4\ 3^{2/3}/49$, which is again a linear nonhomogeneous ODE. Its general solution is $R_2(\theta) = \frac{3}{49}\ (3\ \theta)^{2/3} + \frac{\mathrm{C}}{\theta^{2/3}}$, where C is an arbitrary constant. The initial condition $R_2(0) = 0$ again implies $\mathrm{C} = 0$, and therefore

$$R_2(\theta) = \frac{3}{49}\ (3\ \theta)^{2/3}. \tag{B.7}$$

This process can be continued, and the equations remain linear and easy to solve. The first six terms lead to the expansion (3.8).

Acknowledgments. We thank Christiane Rousseau, Pavel Winternitz, and Malidi Ahamadi for their comments on the early stages of this work. We also thank the anonymous referees for their subtle comments and constructive criticism.

REFERENCES

[1] D. Bäuerle, *Laser Processing and Chemistry*, 3rd ed., Springer-Verlag, Berlin, 2000.
[2] H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids*, Clarendon, Oxford, UK, 1988.
[3] J.-Y. Degorce, A. Saucier, and M. Meunier, *A simple analytical method for the characterization of the melt region of a semiconductor under a focused laser irradiation*, Appl. Surface Sci., 208–209 (2003), pp. 267–271.
[4] H. Kisdarjono, A. T. Voutsas, and R. Solanki, *Three-dimensional simulation of rapid melting and resolidification of thin Si films by excimer laser annealing*, J. Appl. Phys., 94 (2003), pp. 4374–4381.
[5] J. F. Ready and D. F. Farson, eds., *LIA Handbook of Laser Materials Processing*, Springer-Verlag, Berlin, 2001.

[6] R. K. Singh and J. Narayan, *Novel method for simulating laser-solid interactions in semiconductors*, Mat. Sci. Engrg., B3 (1989), pp. 217–230.
[7] V. N. Tokarev and A. F. H. Kaplan, *Modeling of time dependent pulsed laser melting*, J. Appl. Phys., 86 (1999), pp. 2836–2846.
[8] R. F. Wood and G. A. Geist, *Modeling of nonequilibrium melting and solidification in laser-irradiated materials*, Phys. Rev. B, 34 (1986), pp. 2606–2620.
[9] R. F. Wood and G. E. Giles, *Macroscopic theory of pulsed-laser annealing. I. Thermal transport and melting*, Phys. Rev. B, 23 (1981), pp. 2923–2942.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 2096–2120

NONLINEAR COUNTERPROPAGATING WAVES, MULTISYMPLECTIC GEOMETRY, AND THE INSTABILITY OF STANDING WAVES*

THOMAS J. BRIDGES† AND FIONA E. LAINE-PEARSON†

Abstract. Standing waves are a fundamental class of solutions of nonlinear wave equations with a spatial reflection symmetry, and they routinely arise in optical and oceanographic applications. At the linear level they are composed of two synchronized counterpropagating periodic traveling waves. At the nonlinear level, they can be defined abstractly by their symmetry properties. In this paper, general aspects of the modulational instability of standing waves are considered. This problem has difficulties that do not arise in the modulational instability of traveling waves. Here we propose a new geometric formulation for the linear stability problem, based on embedding the standing wave in a four-parameter family of nonlinear counterpropagating waves. Multisymplectic geometry is shown to encode the stability properties in an essential way. At the weakly nonlinear level we obtain the surprising result that standing waves are modulationally unstable only if the component traveling waves are modulation unstable. Systems of nonlinear wave equations will be used for illustration, but general aspects will be presented, applicable to a wide range of Hamiltonian PDEs, including water waves.

Key words. modulation instability, variational principles, periodic waves, hyperbolic PDEs, water waves

AMS subject classifications. 70H33, 70S05, 76B07

DOI. 10.1137/S0036139903423753

1. Introduction.

When considering spatially periodic solutions of nonlinear wave equations on the real line, there are two "canonical" classes of temporally periodic solutions: traveling waves and standing waves. Standing waves arise naturally when the system has a reflection symmetry. In this paper the linear stability problem for standing waves is considered.

To illustrate the basic issues, consider the prototype nonlinear wave equation

$$\mathbf{u}_{tt} - \mathbf{C}\mathbf{u}_{xx} + \nabla V(\mathbf{u}) = 0, \quad \mathbf{u} \in \mathbb{R}^m, \quad x \in \mathbb{R}, \tag{1.1}$$

where $\mathbf{C}$ is a symmetric, positive definite, $m \times m$ matrix; $V : \mathbb{R}^m \to \mathbb{R}$ is a given smooth function; and ∇ is the standard gradient on $\mathbb{R}^m$. This class of wave equations appears in a wide range of applications. An example is DNA modeling [30], where a typical case would be $m = 2$, $\mathbf{C} = \mathrm{diag}(1, c^2)$, and $V(\mathbf{u}) = \cos(u_1 + u_2) - 2\cos u_1 - 2\cos u_2$.

For the system (1.1) a standing wave is a spatially periodic and temporally periodic solution which is invariant under reflection $x \mapsto -x$. (A precise definition of standing wave will be given in section 3.)

Suppose a standing wave solution of (1.1) exists and denote it by $\hat{\mathbf{u}}(x,t)$. This existence problem is itself highly nontrivial due to the potential for small divisors. (The relevance of this issue is discussed in section 7.) The linearized stability equation for $\hat{\mathbf{u}}$ is then $\mathbf{u}_{tt} - \mathbf{u}_{xx} + D^2V(\hat{\mathbf{u}})\mathbf{u} = 0$. A modulational instability is a solution of the type $\mathbf{u}(x,t) = \mathrm{Re}(\mathrm{e}^{i\alpha x}\mathbf{v}(x,t))$, where $\mathbf{v}(x,t)$ is periodic in x of the same period as $\hat{\mathbf{u}}(x,t)$ and α is real with $0 < |\alpha| << 1$, and $\|\mathbf{v}\|$ is exponentially growing in

*Received by the editors February 26, 2003; accepted for publication (in revised form) February 12, 2004; published electronically September 2, 2004.
http://www.siam.org/journals/siap/64-6/42375.html
†Department of Mathematics and Statistics, University of Surrey, Guildford, Surrey, GU2 7XH, UK (t.bridges@eim.surrey.ac.uk).

time. Basic technical issues are associated with this instability problem, such as an appropriate function space in which to define the spectral problem, but these issues will not be considered here. There is a more fundamental issue associated with modulational instability that arises even when we suppose that the basic state $\hat{\mathbf{u}}(x,t)$ is a classical solution of (1.1) and a smooth function of the wavenumber and frequency. It is this fundamental issue, which can be attributed to the fact that standing waves are related to a pair of synchronized counterpropagating traveling waves, that we will address here.

Before considering the counterpropagation property of standing waves, it is worth recalling the analogous linear stability problem for traveling waves. Let $\mathbf{u}(x,t) = \hat{\boldsymbol{\phi}}(\theta)$ be a periodic traveling wave solution of (1.1), where $\theta = \omega t + kx + \theta_0$. The solution $\hat{\boldsymbol{\phi}}$ is a 2π-periodic function of θ, ω is the frequency, and k is the wavenumber. The distinction between stability of traveling waves and standing waves can already be seen at small amplitude. Therefore consider two well-known methods for determining whether the traveling wave is modulationally unstable: the Whitham theory [29] and the use of modulation equations such as the nonlinear Schrödinger (NLS) equation.

According to the Whitham modulation theory, a weakly nonlinear wave of amplitude A, of a nonlinear wave equation that can be derived from a Lagrangian formulation, is modulationally unstable if

$$\omega_0''(k)\,\omega_2(k) < 0, \tag{1.2}$$

where $\omega_0(k)$ is the frequency of the linearized wave and $\omega_2(k)$ is the weakly nonlinear correction to the frequency, that is, $\omega(k) = \omega_0(k) + \omega_2(k)|A|^2 + \cdots$ [29].

Using formal asymptotic methods, an NLS equation can be derived for the weakly nonlinear amplitude $A(X,T)$, by letting $\hat{\phi}(\theta) = A(X,T)\mathrm{e}^{\theta} + \text{c.c.} + \cdots$,

$$\mathrm{i}A_T + \tfrac{1}{2}\omega_0''(k)\,A_{XX} = \sigma\,|A|^2 A$$

(cf. [28]; see also [16] for a rigorous justification of this approach for scalar nonlinear wave equations). The basic weakly nonlinear traveling wave is represented in this equation as a solution of the form $A(X,T) = A_0\mathrm{e}^{\mathrm{i}\omega T}$, $A_0 \in \mathbb{C}$, and this state is linearly unstable precisely when (1.2) is satisfied.

Now, the modulational instability of traveling waves, particularly the weakly nonlinear limit, is well understood, from physical, numerical, and rigorous points of view.

The case of standing waves is more difficult. Surprisingly, there is no generalization of the Whitham theory to treat the modulational instability of standing waves. The only theory in the literature that has been proposed for the modulation instability of standing waves is the use of modulation equations (Knobloch and Pierce [18]; see also [17]).

At the linear level, standing waves reduce to a pair of synchronized counterpropagating waves. Therefore one might suspect that a pair of nonlinearly coupled NLS equations of the form

$$\begin{aligned} \mathrm{i}A_T + \mathrm{i}c_g\,A_X &= \tfrac{1}{2}\omega_0''(k)A_{XX} - \sigma\,|A|^2A + 2\sigma(k)\,|B|^2A, \\ \mathrm{i}B_T - \mathrm{i}c_g\,B_X &= \tfrac{1}{2}\omega_0''(k)B_{XX} - \sigma\,|B|^2B + 2\sigma(k)\,|A|^2B \end{aligned} \tag{1.3}$$

would be a suitable model for modulation instability of standing waves. Indeed, in equation (3.3) of Okamura [23], a coupled NLS system of this form is proposed to model the instability of standing waves. The above system was derived specifically to model standing water waves, but the argument is similar for standing waves of any

nonlinear wave equation, although the coefficients in (1.3) would differ. A standing wave is represented in this system by a solution of the form $A = B = B_0 e^{i\omega T}$.

However, Knobloch and Pierce [18] argue that this coupled set of equations is not valid, and this observation is confirmed by the rigorous analysis of Pierce and Wayne [25] and Bambusi, Carati, and Ponno [2]. They argue that the coupling term needs to be replaced by mean-field coupling terms,

$$\begin{aligned} iA^+_T &= \tfrac{1}{2}\omega_0''(k)A^+_{X_+X_+} - \sigma(k)|A^+|^2A^+ + \beta(k)\Lambda^+(A^+)\,A^+, \\ -iA^-_T &= \tfrac{1}{2}\omega_0''(k)A^-_{X_-X_-} - \sigma(k)|A^-|^2A^- + \beta(k)\Lambda^-(A^-)\,A^-, \end{aligned} \tag{1.4}$$

where $X_\pm = X \mp c_g T$ and $\Lambda^\pm(A^\pm) = \frac{1}{P_\pm}\int_0^{P_\pm} |A^\pm|^2\,dX_\pm$.

The distinction between (1.3) and (1.4) is significant as they do not give equivalent results on modulational instability of standing waves. The rational asymptotics presented in [18, 17], and the rigorous theory of [25, 2], provide strong support for the validity of (1.4).

Modulation equations have severe limitations, however. For example, the above modulation equations are limited to weakly nonlinear standing waves. In this paper we present a new theoretical framework for studying the modulational instability of standing waves. The theory is global (i.e., not restricted to small amplitude) and is based on a new variational principle.

Restricting the new theory to small amplitude waves, it predicts the same instability as the modulation equation (1.4). Since the theory presented here is significantly different from the theory used by Okamura and Knobloch and Pierce, it provides additional support for the validity of the modulation instability predicted by (1.4). Physically, the weakly nonlinear result is quite surprising, since weakly nonlinear periodic standing waves are modulationally unstable only if the component weakly nonlinear traveling waves are unstable. However, this correspondence between the instability of traveling and standing waves will not in general carry over to finite-amplitude standing waves.

The theory here will be developed for the modulation instability of standing wave solutions of Hamiltonian PDEs. The theoretical framework has two parts: first, standing waves can be characterized by a constrained variational principle that encodes information about the linear stability problem. Second, by formulating and studying the linear stability problem directly, we show how the information from the variational principle appears explicitly in the linear stability problem. The main result is that the stability exponents for all long-wave instabilities of standing waves of any amplitude (for which they exist) are determined by the roots of a quartic polynomial whose coefficients can be determined explicitly from the existing standing wave.

The obvious variational principle for standing waves does not provide enough information about the linear stability problem. Surprisingly, we find that the natural approach is to embed the family of standing waves in a four-parameter family, initially, construct a variational principle for this larger family, and then take the limit to the original two-parameter family. The argument in favor of this approach is provided by the analogy of standing waves as synchronized counterpropagating waves: the larger parameter family provides information about how the component counterpropagating waves might break up due to instability.

Conservative PDEs can be analyzed from a Lagrangian, Hamiltonian, or multisymplectic Hamiltonian viewpoint. However, neither the Lagrangian nor the classical Hamiltonian perspective provides sufficient geometry to give abstract results—that

is, results that rely only on the Hamiltonian structure and are independent of the particular PDE. It is the multisymplectic formulation of Hamiltonian PDEs that provides sufficient geometry for a general theory. The class of Hamiltonian PDEs that we consider in canonical form is

$$\mathbf{M}Z_t + \mathbf{K}Z_x = \nabla S(Z), \quad Z \in \mathbb{R}^n, \tag{1.5}$$

where $\mathbf{M}$ and $\mathbf{K}$ are constant $n \times n$ skew-symmetric matrices and $S : \mathbb{R}^n \to \mathbb{R}$ is a given smooth function. An example of the multisymplectification process is given in section 2. Most Hamiltonian PDEs can be cast into this form, including water waves, and other examples can be found in [6, 7, 9] and references therein.

Abstractly, these systems can still be characterized as Lagrangian PDEs by considering Lagrangians in the canonical form

$$\mathscr{L} = \int\int \mathrm{L}(Z, Z_t, Z_x)\,\mathrm{d}x\,\mathrm{d}t \quad \text{with} \quad \mathrm{L}(Z, Z_t, Z_x) = \tfrac{1}{2}\langle \mathbf{M}Z_t, Z\rangle + \tfrac{1}{2}\langle \mathbf{K}Z_x, Z\rangle - S(Z), \tag{1.6}$$

where $\langle\cdot,\cdot\rangle$ is a standard inner product on $\mathbb{R}^n$. This Lagrangian, however, retains all the geometry—two symplectic structures and the scalar function S—of the multisymplectic formulation.

An outline of the paper is as follows. In section 2, an example is given of multisymplectification, using (1.1) as an example. In section 3 standing waves are defined and it is shown that a consequence of the definition is that the momentum is identically zero. New variational principles for standing waves and standing waves embedded in a four-parameter family of counterpropagating waves are presented in section 4. There is an interesting connection between the geometry of $\mathbf{O}(2)$-equivariant finite-dimensional Hamiltonian systems, such as the spherical pendulum, and nonlinear wave equations on the real line with periodic boundary conditions, and this connection is explored in Appendix A. The details of the stability analysis for weakly nonlinear and finite-amplitude standing waves are presented in sections 6 and 7.

The small divisor issue that appears in the analysis of standing waves is outside the scope of this paper, but the issue is briefly discussed in section 7. One of the main motivations for studying the modulational instability of standing waves is their importance in the water-wave problem. The theory developed here does not apply directly, but we speculate on some of the implications for water waves in section 8.

2. Multisymplectifying systems of nonlinear wave equations. The theory for instability of standing waves will be developed for the general class of PDEs (1.5). In this section, the general class of nonlinear wave equations (1.1) will be used to illustrate the transformation to multisymplectic form. In sequence, a Lagrangian, a classical Hamiltonian, and then a multisymplectic Hamiltonian formulation of this system will be presented.

The canonical form of the Lagrangian for (1.1) is

$$\mathscr{L} = \int_{\mathscr{V}} \mathrm{L}(\mathbf{u}, \mathbf{u}_t, \mathbf{u}_x)\,\mathrm{d}x \wedge \mathrm{d}t, \quad \mathrm{L}(\mathbf{u}, \mathbf{u}_t, \mathbf{u}_x) = \tfrac{1}{2}\mathbf{u}_t \cdot \mathbf{u}_t - \tfrac{1}{2}\mathbf{u}_x \cdot \mathbf{C}\mathbf{u}_x - \mathrm{V}(\mathbf{u}), \tag{2.1}$$

where $\mathscr{V}$ represents the volume in (x, t) space, and $\cdot$ represents the standard inner product on $\mathbb{R}^m$.

The canonical Hamiltonian formulation for the nonlinear wave equation is obtained by taking the Legendre transform with respect to time only, $\mathbf{v} = \frac{\partial \mathrm{L}}{\partial \mathbf{u}_t} = \mathbf{u}_t$,

and then the governing equations take the form

(2.2)
$$\frac{\partial}{\partial t}\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}=\begin{bmatrix}\mathbf{0} & \mathbf{I}\\ -\mathbf{I} & \mathbf{0}\end{bmatrix}\begin{pmatrix}\delta\mathscr{H}/\delta\mathbf{u}\\ \delta\mathscr{H}/\delta\mathbf{v}\end{pmatrix},\qquad \mathscr{H}(\mathbf{u},\mathbf{v})=\int_{\mathbb{R}}\left[\frac{1}{2}\mathbf{v}\cdot\mathbf{v}+\frac{1}{2}\mathbf{u}_x\cdot\mathbf{C}\mathbf{u}_x+V(\mathbf{u})\right]\,\mathrm{d}x.$$

This Hamiltonian formulation of the nonlinear wave equation has been widely used in analysis (see [19] and references therein). However, a disadvantage of this formulation, when studying pattern formation, is that the Hamiltonian function and symplectic structure associated with (2.2) require specification of a space of functions over the x direction a priori. In the case of modulation instabilities, the basic state is periodic in space but the perturbation class will be in general quasi-periodic. In other words, we may want to determine the spatial variation of the solution set a posteriori.

Multisymplecticity puts space and time on an equal footing. The governing equations are obtained by taking a Legendre transform with respect to all directions,

$$\mathbf{v}=\frac{\partial \mathrm{L}}{\partial \mathbf{u}_t}=\mathbf{u}_t\quad\text{and}\quad \mathbf{w}=\frac{\partial \mathrm{L}}{\partial \mathbf{u}_x}=-\mathbf{C}\mathbf{u}_x.$$

The Legendre transform generates a new Hamiltonian functional,

$$S(\mathbf{u},\mathbf{v},\mathbf{w})=\mathbf{v}\cdot\mathbf{u}_t+\mathbf{w}\cdot\mathbf{u}_x-\mathrm{L}=\tfrac{1}{2}\mathbf{v}\cdot\mathbf{v}-\tfrac{1}{2}\mathbf{w}\cdot\mathbf{C}^{-1}\mathbf{w}+V(\mathbf{u}). \tag{2.3}$$

This function can be thought of as generated by a total Legendre transform as above, or it can be viewed as a secondary Legendre transform: $-S$ is the Legendre transform of the Hamiltonian density H in (2.2).

Now, the new Lagrangian for the system is in standard form for a generalization of Hamilton's principle,

$$\mathscr{L}=\iint \mathrm{L}(\mathbf{u},\mathbf{v},\mathbf{w})\,\mathrm{d}x\wedge\mathrm{d}t,\quad \mathrm{L}(\mathbf{u},\mathbf{v},\mathbf{w})=\mathbf{v}\cdot\mathbf{u}_t+\mathbf{w}\cdot\mathbf{u}_x-S(\mathbf{u},\mathbf{v},\mathbf{w}), \tag{2.4}$$

and the governing equations are given by

$$\begin{aligned}0&=\mathrm{L}_{\mathbf{u}}=-\mathbf{v}_t-\mathbf{w}_x-S_{\mathbf{u}}=-\mathbf{v}_t-\mathbf{w}_x-\nabla V(\mathbf{u}),\\ 0&=\mathrm{L}_{\mathbf{v}}=\mathbf{u}_t-S_{\mathbf{v}}=\mathbf{u}_t-\mathbf{v},\\ 0&=\mathrm{L}_{\mathbf{w}}=\mathbf{u}_x-S_{\mathbf{w}}=\mathbf{u}_x+\mathbf{C}^{-1}\mathbf{w},\end{aligned}$$

using standard fixed endpoint conditions for the variations. While the PDE is now expressed as a first-order system, it has a multisymplectic structure which is awkward for analysis. It can be written in the form $\mathbf{M}Z_x+\mathbf{K}Z_x=\nabla S(Z)$ with $Z\in\mathbb{R}^{3m}$, but the pair of symplectic operators, $\mathbf{M}$ and $\mathbf{K}$, act on $\mathbb{R}^{3m}$ and are always degenerate. This structure can be improved by observing that $\mathbf{v}$ and $\mathbf{w}$ satisfy the constraint $\mathbf{C}^{-1}\mathbf{w}_t+\mathbf{v}_x=0$. Therefore add this constraint to the Lagrangian with vector-valued Lagrange multiplier $\mathbf{p}$, that is,

$$\begin{aligned}\mathscr{L}&=\int_{\mathscr{V}}\mathrm{L}(\mathbf{u},\mathbf{v},\mathbf{w},\mathbf{p})\,\mathrm{d}x\wedge\mathrm{d}t,\\ \mathrm{L}(\mathbf{u},\mathbf{v},\mathbf{w},\mathbf{p})&=\mathbf{v}\cdot\mathbf{u}_t+\mathbf{w}\cdot\mathbf{u}_x-S(\mathbf{u},\mathbf{v},\mathbf{w})+\mathbf{p}\cdot(\mathbf{C}^{-1}\mathbf{w}_t+\mathbf{v}_x).\end{aligned}$$

The governing equations are now

$$
\begin{aligned}
0 = \mathrm{L}_{\mathbf{u}} &= -\mathbf{v}_t - \mathbf{w}_x - S_{\mathbf{u}} = -\mathbf{v}_t - \mathbf{w}_x - \nabla V(\mathbf{u}),\\
0 = \mathrm{L}_{\mathbf{v}} &= \mathbf{u}_t - S_{\mathbf{v}} - \mathbf{p}_x = \mathbf{u}_t - \mathbf{p}_x - \mathbf{v},\\
0 = \mathrm{L}_{\mathbf{w}} &= \mathbf{u}_x - S_{\mathbf{w}} - \mathbf{C}^{-1}\mathbf{p}_t = -\mathbf{C}^{-1}\mathbf{p}_t + \mathbf{u}_x + \mathbf{C}^{-1}\mathbf{w},\\
0 = \mathrm{L}_{\mathbf{p}} &= \mathbf{C}^{-1}\mathbf{w}_t + \mathbf{v}_x = \mathbf{C}^{-1}\mathbf{w}_t + \mathbf{v}_x
\end{aligned}
$$

or

$$
\begin{bmatrix} 0 & -\mathbf{I} & 0 & 0 \\ \mathbf{I} & 0 & 0 & 0 \\ 0 & 0 & 0 & -\mathbf{C}^{-1} \\ 0 & 0 & \mathbf{C}^{-1} & 0 \end{bmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{w} \\ \mathbf{p} \end{pmatrix}_t + \begin{bmatrix} 0 & 0 & -\mathbf{I} & 0 \\ 0 & 0 & 0 & -\mathbf{I} \\ \mathbf{I} & 0 & 0 & 0 \\ 0 & \mathbf{I} & 0 & 0 \end{bmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{w} \\ \mathbf{p} \end{pmatrix}_x = \begin{pmatrix} \nabla \mathrm{V}(\mathbf{u}) \\ \mathbf{v} \\ -\mathbf{C}^{-1}\mathbf{w} \\ 0 \end{pmatrix}. \tag{2.5}
$$

This system can be expressed in canonical multisymplectic form (1.5) with $n = 4m$, and indeed, in this case, $\mathbf{M}$ and $\mathbf{K}$ define symplectic structures on $\mathbb{R}^{4m}$. The two symplectic structures do not commute in general, unless $\mathbf{C} = \mathbf{I}$, since

$$
[\mathbf{M}, \mathbf{K}] = \mathbf{MK} - \mathbf{KM} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \otimes (\mathbf{I} - \mathbf{C}^{-1}).
$$

In the scalar case $m = 1$, scaling can be introduced so that $\mathbf{M}$ and $\mathbf{K}$ always commute.

In summary, the main point of this section is that the system of nonlinear wave equations (1.1) can be characterized in terms of geometric properties: two symplectic structures, and a scalar-valued function $S(Z)$, on a finite-dimensional phase space: $Z \in \mathbb{R}^n$.

A property of the nonlinear wave equation (1.1) that is important for the existence of standing waves is reversibility in x. If $\mathbf{u}(x,t)$ is a solution of (1.1), then $\mathbf{u}(-x,t)$ is also a solution. In the multisymplectification of (1.1), this reversibility is represented by the action

$$
\mathbf{r} \cdot Z(x,t) = \mathbf{R}Z(-x,t) \quad \text{with} \quad \mathbf{R} = \mathrm{diag}(\mathbf{I}, \mathbf{I}, -\mathbf{I}, -\mathbf{I}) \in \mathbb{R}^{4m \times 4m}. \tag{2.6}
$$

The involution $\mathbf{R}$ and its associated action satisfy

$$
\mathbf{RM} = \mathbf{MR}, \quad \mathbf{RK} = -\mathbf{KR}, \quad \text{and} \quad S(\mathbf{r} \cdot Z) = S(Z). \tag{2.7}
$$

In turn, the properties (2.7) imply that $\mathbf{r} \cdot Z$ is a solution of the wave equation in the form (1.5) whenever Z is.

The system of nonlinear wave equations (1.1) is reversible in t as well, and a multisymplectic t-reversor can also be defined, but t-reversibility will not be needed in the general theory for standing waves.

3. Standing wave solutions of Hamiltonian PDEs. The theory for standing waves can be developed based only on the geometric properties of the multisymplectic formulation. Therefore, as in the previous section, we will assume that the PDE has been transformed to a multisymplectic Hamiltonian PDE, and we take the following general class of PDEs as the starting point for the analysis:

$$\mathbf{M}Z_t + \mathbf{K}Z_x = \nabla S(Z), \quad Z \in \mathbb{R}^n. \tag{3.1}$$

The only hypotheses on (3.1) are that $\mathbf{M}$ and $\mathbf{K}$ are constant $n \times n$ skew-symmetric matrices and $S : \mathbb{R}^n \to \mathbb{R}$ is a given smooth function (at least twice continuously differentiable), which does not depend explicitly on x or t. On $\mathbb{R}^n$, the standard inner product will be denoted by $\langle \cdot, \cdot \rangle$.

For the existence of standing waves, we will require that the system (3.1) is x-reversible with a multisymplectic action of the reversor,

$$\mathbf{r} \cdot Z(x,t) = \mathbf{R}Z(-x,t) \tag{3.2}$$

for some isometric involution $\mathbf{R} : \mathbb{R}^n \to \mathbb{R}^n$ satisfying the identities (2.7) (with $\mathbf{M}$, $\mathbf{K}$ and S associated with (3.1)). In this setting, an abstract definition of a standing wave can be given.

DEFINITION. *A solution $\widehat{Z}(x,t)$ of* (3.1) *is called a* standing wave *if it is periodic in both x and t and satisfies $\mathbf{r} \cdot \widehat{Z}(x,t) = \widehat{Z}(x,t)$.*

Curiously, we cannot find anywhere in the literature where a general definition of standing waves for nonlinear PDEs has heretofore been given.

It is sometimes remarked that standing waves are spatially periodic waves with zero momentum. However, we can show that zero momentum is a *consequence* of the above definition.

What is momentum? The momentum here is defined to be the conserved quantity given by Noether's theorem associated with the translation invariance in x of the PDE. If (3.1) represents a physical system, this conserved quantity may indeed be the physical momentum. An application of Noether's theorem to the Lagrangian (1.6) (see Appendix B for this argument) shows that the appropriate form for the momentum on a space of functions that are 2π periodic in x is

$$\mathscr{I}(Z) = \oint \frac{1}{2}\langle \mathbf{M}Z_x(x,t), Z(x,t)\rangle \, \mathrm{d}x \quad \text{where} \quad \oint (\ ) \, \mathrm{d}x := \frac{1}{2\pi}\int_0^{2\pi} (\ ) \, \mathrm{d}x. \tag{3.3}$$

Given this expression for momentum, we can show that $\mathscr{I}(\widehat{Z}) = 0$ if $\widehat{Z}(x,t)$ is a standing wave solution of (3.1):

$$\begin{aligned}
\mathscr{I}(\mathbf{r}\cdot Z) &= \oint \tfrac{1}{2}\langle \mathbf{M}\mathbf{R}(Z(-x,t))_x, \mathbf{R}Z(-x,t)\rangle \, \mathrm{d}x \quad \text{(by definition)} \\
&= -\oint \tfrac{1}{2}\langle \mathbf{M}\mathbf{R}Z_x(-x,t), \mathbf{R}Z(-x,t)\rangle \, \mathrm{d}x \\
&= -\oint \tfrac{1}{2}\langle \mathbf{R}\mathbf{M}Z_x(-x,t), \mathbf{R}Z(-x,t)\rangle \, \mathrm{d}x \quad (\text{using } \mathbf{R}\mathbf{M} = \mathbf{M}\mathbf{R}) \\
&= -\oint \tfrac{1}{2}\langle \mathbf{M}Z_x(-x,t), Z(-x,t)\rangle \, \mathrm{d}x \quad (\text{since } \mathbf{R} \text{ is an isometry}) \\
&= -\oint \tfrac{1}{2}\langle \mathbf{M}Z_x(x,t), Z(x,t)\rangle \, \mathrm{d}x \\
&\quad (\text{using the change of variable } x \mapsto -x \text{ and periodicity}) \\
&= -\mathscr{I}(Z).
\end{aligned}$$

Therefore, if $\widehat{Z}(x,t)$ is a standing wave and so $\mathbf{r}\cdot\widehat{Z} = \widehat{Z}$, it is immediate that $\mathscr{I}(\widehat{Z}) = 0$.

4. Variational principles for standing waves and counterpropagating waves. At the linear level, a standing wave consists of a pair of synchronized counterpropagating waves,

$$Z(x,t) = A\xi\, \mathrm{e}^{\mathrm{i}(\omega t + kx)} + A\xi\, \mathrm{e}^{\mathrm{i}(\omega t - kx)} + \text{c.c.},$$

where $A \in \mathbb{C}$ is a complex amplitude and $\xi \in \mathbb{C}^n$ is an eigenvector associated with the linearization of (3.1). A natural generalization of this form to finite amplitude is to look for *nonlinear* solutions of the form

$$Z(x,t) = \widehat{Z}(\theta_1,\theta_2), \quad \theta_1 = \omega t + kx + \theta_1^o, \quad \theta_2 = \omega t - kx + \theta_2^o, \tag{4.1}$$

where θ_j^o are arbitrary constant phases and $\widehat{Z}$ is a 2π-periodic function of θ_1 and θ_2.

Substituting the form (4.1) into (3.1) results in

$$\omega\,\mathbf{M}\left(\frac{\partial\widehat{Z}}{\partial\theta_1} + \frac{\partial\widehat{Z}}{\partial\theta_2}\right) + k\,\mathbf{K}\left(\frac{\partial\widehat{Z}}{\partial\theta_1} - \frac{\partial\widehat{Z}}{\partial\theta_2}\right) = \nabla S(\widehat{Z}). \tag{4.2}$$

The operators $\mathbf{M}\partial_{\theta_j}$ and $\mathbf{K}\partial_{\theta_j}$ are formally self-adjoint operators on a space of doubly periodic functions. Hence, treating ω and k as Lagrange multipliers, (4.2) can be interpreted as the necessary condition for a constrained variational principle. Let

$$\mathscr{A}(Z) = \int_{\mathbb{T}^2} \tfrac{1}{2}\langle \mathbf{M}(\partial_{\theta_1} + \partial_{\theta_2})Z, Z\rangle\, \mathrm{d}\theta \quad \text{and} \quad \mathscr{B}(Z) = \int_{\mathbb{T}^2} \tfrac{1}{2}\langle \mathbf{K}(\partial_{\theta_1} - \partial_{\theta_2})Z, Z\rangle\, \mathrm{d}\theta,$$

where $\int_{\mathbb{T}^2}(\ )\,\mathrm{d}\theta := (2\pi)^{-2}\int_0^{2\pi}\int_0^{2\pi}(\ )\,\mathrm{d}\theta_1\mathrm{d}\theta_2$. The constrained variational principle is then to find critical points of $\mathscr{S}$, $S(Z)$ averaged over $\mathbb{T}^2$, subject to fixed values of the constraints $\mathscr{A}$ and $\mathscr{B}$. It follows from standard Lagrange multiplier theory that this constrained variational principle is nondegenerate when

$$\det\begin{bmatrix} \mathscr{A}_\omega & \mathscr{A}_k \\ \mathscr{B}_\omega & \mathscr{B}_k \end{bmatrix} \neq 0. \tag{4.3}$$

This variational principle gives a global characterization of any state of (3.1) which is periodic in both x and t. It includes a characterization of standing waves and traveling waves. The special case of strictly traveling waves was considered in [7], and it is shown there that the sign of the determinant (4.3) carries information about linear stability.

Another way to view this variational principle is as a generalization to the spatiotemporal setting of the classical variational principle for periodic solutions of *finite-dimensional* Hamiltonian systems: find critical points of the energy (Hamiltonian) on level sets of the action on a space of periodic functions, with the frequency ω as a Lagrange multiplier. For finite-dimensional Hamiltonian systems this variational principle has been widely used to prove the existence of periodic solutions (cf. [20] and references therein). However, the variational principle associated with (4.1) is more difficult to work with for the case of standing waves. Although standing waves are periodic solutions, the fact that there is an infinite number of modes can cause problems with small divisors (see section 7).

The form of the solution (4.1) is not the most general form for a pair of counterpropagating waves. When considering the linear stability problem for standing waves, it will turn out that a somewhat more general variational principle will be crucial for getting a geometric characterization of linear instability of standing waves. The idea is to embed the family of standing waves in a *four*-parameter family of counterpropagating waves, with the standing wave obtained as a limiting two-parameter case.

Consider the more general class of solutions of (3.1); let

$$Z(x,t) = \widehat{Z}(\theta_1,\theta_2) \quad \text{with} \quad \theta_j = \omega_j\, t + k_j\, x + \theta_j^o, \quad j = 1,2, \tag{4.4}$$

where $\widehat{Z}$ is again a 2π-periodic function of both θ_1 and θ_2. The significant difference here is that the state $\widehat{Z}$ now depends on four parameters, and the interpretation as two counterpropagating waves that are not necessarily synchronized is now evident. Indeed, in general, they may even be propagating in the same direction. However, it is the case of counterpropagating waves, near synchronized standing waves, that is of greatest interest here, that is, $k_1 + k_2 \approx 0$ and $\omega_1 - \omega_2 \approx 0$.

The function $\widehat{Z}$ now satisfies

$$\omega_1 \mathbf{M}\frac{\partial \widehat{Z}}{\partial \theta_1} + \omega_2 \mathbf{M}\frac{\partial \widehat{Z}}{\partial \theta_2} + k_1 \mathbf{K}\frac{\partial \widehat{Z}}{\partial \theta_1} + k_2 \mathbf{K}\frac{\partial \widehat{Z}}{\partial \theta_2} = \nabla \mathscr{S}(\widehat{Z}), \tag{4.5}$$

where $\mathscr{S}$ is S averaged over θ_1 and θ_2. Equation (4.5) can be interpreted as the Lagrange necessary condition for the constrained variational principle: find critical points of S averaged over $\mathbb{T}^2$ restricted to level sets of the four functionals

$$\mathscr{A}_j(Z) = \int_{\mathbb{T}^2} \tfrac{1}{2}\langle \mathbf{M}\partial_{\theta_j} Z, Z\rangle \,\mathrm{d}\theta \quad \text{and} \quad \mathscr{B}_j(Z) = \int_{\mathbb{T}^2} \tfrac{1}{2}\langle \mathbf{K}\partial_{\theta_j} Z, Z\rangle \,\mathrm{d}\theta, \quad j = 1,2. \tag{4.6}$$

The Lagrange necessary condition can be written

$$\nabla \mathscr{S}(\widehat{Z}) = \omega_1\, \nabla \mathscr{A}_1(\widehat{Z}) + \omega_2\, \nabla \mathscr{A}_2(\widehat{Z}) + k_1\, \nabla \mathscr{B}_1(\widehat{Z}) + k_2\, \nabla \mathscr{B}_2(\widehat{Z}). \tag{4.7}$$

The frequencies ω_1, ω_2 and the wavenumbers k_1, k_2 appear as Lagrange multipliers. Using standard Lagrange multiplier theory, this constrained variational principle is nondegenerate if

$$\det \begin{bmatrix} \frac{\delta \mathscr{A}}{\delta \omega} & \frac{\delta \mathscr{A}}{\delta k} \\ \frac{\delta \mathscr{B}}{\delta \omega} & \frac{\delta \mathscr{B}}{\delta k} \end{bmatrix} \neq 0, \quad \text{where} \quad \frac{\delta \mathscr{A}}{\delta \omega} = \begin{pmatrix} \frac{\partial \mathscr{A}_1}{\partial \omega_1} & \frac{\partial \mathscr{A}_1}{\partial \omega_2} \\ \frac{\partial \mathscr{A}_2}{\partial \omega_1} & \frac{\partial \mathscr{A}_2}{\partial \omega_2} \end{pmatrix}, \tag{4.8}$$

with similar expressions for the 2×2 matrices $\frac{\delta \mathscr{A}}{\delta k}, \frac{\delta \mathscr{B}}{\delta \omega}, \frac{\delta \mathscr{B}}{\delta k}$. It is the two-parameter subfamily of two-wave interactions that correspond to standing waves that is of interest. Given a function $\widehat{Z}(\theta_1,\theta_2)$ satisfying this variational principle, a standing wave is recovered formally by taking the limit to synchronized counterpropagating waves

$$\omega_1 \to \omega, \quad k_1 \to k, \quad \omega_2 \to \omega, \quad \text{and} \quad k_2 \to -k$$

if the limits exist. This limit is taken *after* the Jacobian matrices in (4.8) are computed.

At first sight, this limit might seem a bit questionable: taking the limit on a torus from irrational values to a resonance? However, there is additional structure

here. The translation invariance in x restricted to periodic functions along with the x-reversibility generates the group $\mathbf{O}(2)$. Translation invariance in time restricted to periodic functions generates an action of $\mathbb{S}^1$. Combining these groups gives $\mathbf{O}(2)\times\mathbb{S}^1$: a *toral symmetry*. The toral symmetry is almost enough structure to allow for smooth variation of parameters on the torus. Indeed, if the system was finite-dimensional, this would be true, and this case is discussed briefly in Appendix A. The obstacle to smoothness for the above limit leading to standing waves is again the potential for small divisors due to a countable number of purely imaginary eigenvalues (see section 7).

5. Stability analysis of nonlinear standing waves. It is in the study of the stability of standing waves that the importance of the embedding of standing waves in the four-parameter family becomes apparent. In this section the linear stability problem for standing waves is formulated and it is shown that the entries in the determinant (4.8) appear in the linear stability analysis in a central way. The strategy is to linearize (3.1) about the full four-parameter two-wave interaction. Then, after the stability condition is deduced, the limit to standing waves is taken.

Substitute $Z(x,t)=\widehat{Z}(\theta_1,\theta_2)+\widehat{U}(\theta_1,\theta_2,x,t)$, where $\widehat{Z}$ is the wave (4.4) and $\widehat{U}$ is a perturbation, into (3.1) and linearize about $\widehat{Z}$,

$$\mathbf{M}\widehat{U}_t+\mathbf{K}\widehat{U}_x=\mathbf{L}(\theta_1,\theta_2)\widehat{U}, \tag{5.1}$$

where

$$\begin{aligned}\mathbf{L}(\theta_1,\theta_2)&=D^2\mathscr{S}(\widehat{Z})-\mathbf{M}\left[\omega_1\frac{\partial}{\partial\theta_1}+\omega_2\frac{\partial}{\partial\theta_2}\right]-\mathbf{K}\left[k_1\frac{\partial}{\partial\theta_1}+k_2\frac{\partial}{\partial\theta_2}\right]\\&=D^2\mathscr{S}(\widehat{Z})-\omega_1D^2\mathscr{A}_1(\widehat{Z})-\omega_2D^2\mathscr{A}_2(\widehat{Z})-k_1D^2\mathscr{B}_1(\widehat{Z})-k_2D^2\mathscr{B}_2(\widehat{Z}).\end{aligned} \tag{5.2}$$

The operator $\mathbf{L}$ is a linear partial differential operator with nonconstant (periodic) coefficients depending on θ_1 and θ_2. Introduce a class of perturbations of modulation type

$$\widehat{U}(\theta_1,\theta_2,x,t)=\mathrm{Re}\left(U(\theta_1,\theta_2)\mathrm{e}^{\lambda t+\mathrm{i}\alpha x}\right),$$

where $\lambda\in\mathbb{C}$ is the stability exponent and $\alpha\in\mathbb{R}$ is the modulation parameter associated with the x-direction. The eigenvalue problem for $(\lambda,U(\theta_1,\theta_2))$ is then

$$\mathbf{L}(\theta_1,\theta_2)U=\lambda\mathbf{M}U+\mathrm{i}\alpha\mathbf{K}U. \tag{5.3}$$

Definition. *If there exists a solution $U(\theta_1,\theta_2)$ of (5.3) which is 2π-periodic in θ_1 and θ_2, for some $\lambda\in\mathbb{C}$ and $\alpha\in\mathbb{R}$, with $\mathrm{Re}(\lambda)>0$, then we say that the basic state $\widehat{Z}(\theta_1,\theta_2)$ is linearly unstable or* spectrally unstable.

The application of this definition and the development of the geometric stability condition are not rigorous. For example, identification of the precise space of functions in which $U(\theta_1,\theta_2)$ might exist is beyond the scope of this paper. The obstacle to rigor is the potential small divisor problem, which would result in the range of the operator $\mathbf{L}$ not being closed. In some special cases, for example, if there is enough symmetry [9], the theory can be made rigorous using a Lyapunov–Schmidt reduction.

The eigenvalue problem (5.3) is still a PDE in θ_1 and θ_2. It has considerable structure (combination of symmetric and antisymmetric operators), but we do not

expect to be able to analyze this spectral problem completely. However, we can get complete results on *long-wave* instability, that is, when $|\alpha| \ll 1$. In this case, the geometry of (4.4) can be used to give a geometric characterization of the spectrum for α small. Eigenvalue problems of this type have been studied geometrically before in a different but related context [9, section 5], and therefore we can appeal to those results.

The kernel of $\mathbf{L}(\theta_1, \theta_2)$ has (at least) two elements,

$$\mathrm{Ker}(\mathbf{L}) \supseteq \mathrm{span}\left\{ \frac{\partial \widehat{Z}}{\partial \theta_1}, \frac{\partial \widehat{Z}}{\partial \theta_2} \right\}, \tag{5.4}$$

and this can be verified by differentiating (4.7) with respect to θ_1 and θ_2. Assume that these two functions are the only elements in the kernel, a property that is generically satisfied. (For certain parameter values, the kernel might be larger.) Then look for long-wave instabilities $\alpha \ll 1$ by expanding U in a Taylor series. Consider the ansatz

$$U = c_1 \left(\widehat{Z}_{\theta_1} + \lambda \widehat{Z}_{\omega_1} + \mathrm{i}\alpha \widehat{Z}_{k_1} \right) + c_2 \left(\widehat{Z}_{\theta_2} + \lambda \widehat{Z}_{\omega_1} + \mathrm{i}\alpha \widehat{Z}_{k_2} \right) + \mathcal{O}(|\lambda|^2 + |\alpha|^2), \tag{5.5}$$

where $\mathbf{c} = (c_1, c_2)$ are arbitrary complex constants. By differentiating (4.7) with respect to ω_1, ω_2, k_1, and k_2 it can be verified that this expression is indeed the solution to (5.3) to leading order.

It is worth remarking here that it is precisely in the leading-order expression for U that the deformation from standing waves to the general two-wave interaction is necessary. Four derivatives of $\widehat{Z}$ with respect to parameters are needed in (5.5), whereas if there was just one frequency and one wavenumber, only $\widehat{Z}_\omega$ and $\widehat{Z}_k$ would be available for (5.5).

Since $\mathbf{L}$ is formally self-adjoint, the solvability condition for (5.3) is

$$\begin{aligned} [\![\widehat{Z}_{\theta_1}, \lambda \mathbf{M} U + \mathrm{i}\alpha \mathbf{K} U]\!] &= 0, \\ [\![\widehat{Z}_{\theta_2}, \lambda \mathbf{M} U + \mathrm{i}\alpha \mathbf{K} U]\!] &= 0, \end{aligned}$$

where

$$[\![f, g]\!] = \int_{\mathbb{T}^2} \langle f(\theta_1, \theta_2), g(\theta_1, \theta_2) \rangle \, \mathrm{d}\theta = \frac{1}{(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} \langle f(\theta_1, \theta_2), g(\theta_1, \theta_2) \rangle \, \mathrm{d}\theta_1 \mathrm{d}\theta_2.$$

This solvability condition still contains the unknown function U, but we have a leading-order expression for U. Substituting the leading-order expression for U into the solvability condition leads to the pair of algebraic equations

$$\left[\mathbf{N}_0 \lambda^2 + \mathrm{i}\alpha\lambda \mathbf{N}_1 + (\mathrm{i}\alpha)^2 \mathbf{N}_2 + \cdots \right] \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{5.6}$$

where $\mathbf{N}_j$, $j = 0, 1, 2$, are 2×2 matrices depending only on the properties of the basic wave $\widehat{Z}$. The derivation of the expression for $\mathbf{N}_0$ will be given, and then the result for the other two will be stated. From the solvability condition we have that

$$\mathbf{N}_0 = \begin{bmatrix} \int_{\mathbb{T}^2} \langle \widehat{Z}_{\theta_1}, \mathbf{M} \widehat{Z}_{\omega_1} \rangle \, \mathrm{d}\theta & \int_{\mathbb{T}^2} \langle \widehat{Z}_{\theta_1}, \mathbf{M} \widehat{Z}_{\omega_2} \rangle \, \mathrm{d}\theta \\ \int_{\mathbb{T}^2} \langle \widehat{Z}_{\theta_2}, \mathbf{M} \widehat{Z}_{\omega_1} \rangle \, \mathrm{d}\theta & \int_{\mathbb{T}^2} \langle \widehat{Z}_{\theta_2}, \mathbf{M} \widehat{Z}_{\omega_2} \rangle \, \mathrm{d}\theta \end{bmatrix}.$$

However, by differentiating the functionals (4.6) with respect to ω_1 and ω_2 we find that the matrix simplifies to

$$\mathbf{N}_0 = -\begin{bmatrix} \frac{\partial \mathscr{A}_1}{\partial \omega_1} & \frac{\partial \mathscr{A}_1}{\partial \omega_2} \\ \frac{\partial \mathscr{A}_2}{\partial \omega_1} & \frac{\partial \mathscr{A}_2}{\partial \omega_2} \end{bmatrix} = -\frac{\partial \mathscr{A}}{\partial \omega}.$$

Similarly, $\mathbf{N}_1 = -\frac{\partial \mathscr{A}}{\partial k} - \frac{\partial \mathscr{B}}{\partial \omega}$ and $\mathbf{N}_2 = -\frac{\partial \mathscr{B}}{\partial k}$, and so (5.6) reduces to

$$\left(\lambda^2 \frac{\delta \mathscr{A}}{\delta \omega} + \mathrm{i}\alpha\lambda \left(\frac{\delta \mathscr{A}}{\delta k} + \frac{\delta \mathscr{B}}{\delta \omega}\right) + (\mathrm{i}\alpha)^2 \frac{\delta \mathscr{B}}{\delta k} + \cdots\right) \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Therefore, if $|\lambda| + |\alpha|$ is sufficiently small and the matrix (4.8) is nondegenerate, the *long-wave stability* of the basic two-wave interaction is determined by the quartic

$$\begin{aligned} \Delta(\lambda, \alpha) &= \det\left[\lambda^2 \frac{\delta \mathscr{A}}{\delta \omega} + \mathrm{i}\alpha\lambda \left(\frac{\delta \mathscr{A}}{\delta k} + \frac{\delta \mathscr{B}}{\delta \omega}\right) + (\mathrm{i}\alpha)^2 \frac{\delta \mathscr{B}}{\delta k}\right] \\ &= \det\left[\sigma^{\mathrm{T}} \otimes \mathbf{I}_2 \begin{pmatrix} \delta \mathscr{A}/\delta \omega & \delta \mathscr{A}/\delta k \\ \delta \mathscr{B}/\delta \omega & \delta \mathscr{B}/\delta k \end{pmatrix} \sigma \otimes \mathbf{I}_2\right], \quad \sigma = \begin{pmatrix} \lambda \\ \mathrm{i}\alpha \end{pmatrix}. \end{aligned} \tag{5.7}$$

The second form shows that central role played by the nondegeneracy condition from the constrained variational principle of section 4.

Expanding out the determinant in (5.7) leads to a quartic polynomial for λ,

$$\Delta(\lambda, \alpha) = a_4\lambda^4 + a_3\lambda^3 + a_2\lambda^2 + a_1\lambda + a_0 \tag{5.8}$$

with

$$\begin{aligned} a_4 &= \det\left(\frac{\delta \mathscr{A}}{\delta \omega}\right), \\ a_3 &= \mathrm{i}\alpha \mathrm{Tr}\left(\frac{\delta \mathscr{A}}{\delta \omega}^{\#} \left(\frac{\delta \mathscr{A}}{\delta k} + \frac{\delta \mathscr{B}}{\delta \omega}\right)\right), \\ a_2 &= -\alpha^2 \det\left(\frac{\delta \mathscr{A}}{\delta k} + \frac{\delta \mathscr{B}}{\delta \omega}\right) - \alpha^2 \mathrm{Tr}\left(\frac{\delta \mathscr{A}}{\delta \omega}^{\#} \frac{\delta \mathscr{B}}{\delta k}\right), \\ a_1 &= -\mathrm{i}\alpha^3 \mathrm{Tr}\left(\frac{\delta \mathscr{B}}{\delta k}^{\#} \frac{\delta \mathscr{A}}{\delta k} + \frac{\delta \mathscr{B}}{\delta \omega}\right), \\ a_0 &= \alpha^4 \det\left(\frac{\delta \mathscr{B}}{\delta k}\right), \end{aligned} \tag{5.9}$$

where the superscript # indicates adjugate, i.e.,

$$\mathbf{C}^{\#} = \begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix}^{\#} = \mathbf{J}^{-1}\mathbf{C}\mathbf{J} = \begin{pmatrix} c_3 & -c_2 \\ -c_2 & c_1 \end{pmatrix}, \quad \text{where } \mathbf{J} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

This stability quartic applies to both standing waves and the deformed two-wave interaction, which may have independent interest. Given a basic four-parameter wave, $(\widehat{Z}; \omega_1, \omega_2, k_1, k_2)$, the coefficients of the quartic can in principle be computed, and then the quartic solved for the four roots, thereby determining whether there is a long-wave instability.

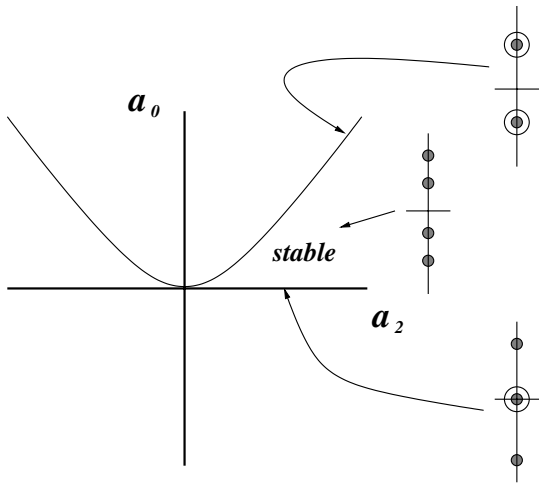


FIG. 1. *Position of the roots of the quartic* $\Delta(\lambda,\alpha)=0$ *when* $a_1=a_3=0$ *and* $a_4=1$*, showing the stable region and its boundary in the* (a_2,a_0) *plane. In all other regions there is at least one unstable root.*

5.1. The stability quartic when $a_1=a_3=0$. A special case of the quartic that can be analyzed in detail is when $a_1=a_3=0$. This case will not arise in general for standing waves at finite amplitude, but it does arise in the limit as the amplitude of the wave tends to zero—the weakly nonlinear limit.

In the analysis of the stability quartic $\Delta(\lambda,\alpha)=0$, the term "instability" will mean that there is at least one root of $\Delta(\lambda,\alpha)=0$ with positive real part, and "stability" will mean that all four roots are purely imaginary and simple (spectral stability). We have the following complete classification of the roots of (5.8) when $a_1=a_3=0$:

$$\begin{aligned} a_4a_0<0 &\Rightarrow \text{instability},\\ a_4a_0\geq 0 \text{ but } a_4a_2<0 &\Rightarrow \text{instability},\\ a_4a_0>0 \text{ but } a_4a_2=0 &\Rightarrow \text{instability},\\ a_4a_0>0 \text{ and } a_4a_2>0 \text{ but } a_2^2-4a_4a_0<0 &\Rightarrow \text{instability},\\ a_4a_0>0,\ a_4a_2>0 \text{ and } a_2^2-4a_4a_0>0 &\Rightarrow \text{stability}. \end{aligned}$$

There are also two special cases where the spectrum is purely imaginary but there are multiple eigenvalues. When $a_0a_4=0$, $a_2a_4>0$, and $a_2^2-4a_4a_0>0$, there is a pair of distinct purely imaginary eigenvalues and a double zero eigenvalue. When $a_4a_0>0$, $a_4a_2>0$ but $a_2^2-4a_4a_0=0$ there is a pair of purely imaginary eigenvalues each of multiplicity two. These special cases lie on the boundary of the region of stability, as illustrated in Figure 1.

6. Instability of weakly nonlinear standing waves. The purpose of this section is threefold. It illustrates in the simplest possible setting how the variational principle and stability theory accumulate information on the spectral problem. It shows explicitly the importance of the limit from the four-parameter two-wave interaction to the two-parameter standing wave. Third, it shows that the theory of this paper recovers the modulation instability predicted by coupled NLS equations with mean-field coupling.

To construct weakly nonlinear counterpropagating waves, take a Fourier ansatz,

(6.1)
$$\begin{aligned}\widehat{Z}(\theta_1,\theta_2) = {} & A_1\boldsymbol{\xi}_1 e^{i\theta_1} + \overline{A_1\boldsymbol{\xi}_1}e^{-i\theta_1} + A_2\boldsymbol{\xi}_2 e^{i\theta_2} + \overline{A_2\boldsymbol{\xi}_2}e^{-i\theta_2} \\ & + \Upsilon_{20} + \Upsilon_{21}e^{2i\theta_1} + \overline{\Upsilon_{21}}e^{-2i\theta_1} + \Upsilon_{22}e^{2i\theta_2} + \overline{\Upsilon_{22}}e^{-2i\theta_2} \\ & + \Upsilon_{23}e^{i(\theta_1+\theta_2)} + \overline{\Upsilon_{23}}e^{-i(\theta_1+\theta_2)} + \Upsilon_{24}e^{i(\theta_1-\theta_2)} + \overline{\Upsilon_{24}}e^{-i(\theta_1-\theta_2)} + \cdots,\end{aligned}$$

where $\theta_j = k_j x + \omega_j t$ $(j = 1, 2)$, A_1 and A_2 are complex amplitudes, and $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ have unit length. This ansatz is substituted into the Lagrangian (1.6),

(6.2)
$$\begin{aligned}\overline{\mathscr{L}}(A_1, A_2, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \mu_1, \mu_2, \Upsilon, \dots) = {} & \mathscr{S}(\widehat{Z}) - \omega_1\mathscr{A}_1 - \omega_2\mathscr{A}_2 - k_1\mathscr{B}_1 - k_2\mathscr{B}_2 \\ & - \mu_1(\|\boldsymbol{\xi}_1\|^2 - 1) - \mu_2(\|\boldsymbol{\xi}_2\|^2 - 1).\end{aligned}$$

Here, μ_1 and μ_2 are Lagrange multipliers associated with the constraints on $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$. The vectors $\boldsymbol{\xi}_j$ are eigenvectors of a linear Hermitian operator, and the Lagrange multipliers μ_j give a way of extending the dispersion relation to the nonlinear case in a coordinate-free way.

Formally solving this finite-dimensional Lagrangian system leads to the reduced Lagrangian

$$\overline{\mathscr{L}} = \mu_1|A_1|^2 + \mu_2|A_2|^2 + \tfrac{1}{2}\sigma_{11}|A_1|^4 + \sigma_{12}|A_1|^2|A_2|^2 + \tfrac{1}{2}\sigma_{22}|A_2|^4 + \cdots$$

and to amplitude equations for A_1 and A_2 of the general form

(6.3)
$$\begin{aligned}A_1\left(\mu(\omega_1, k_1) + \sigma_{11}|A_1|^2 + \sigma_{12}|A_2|^2 + \cdots\right) = 0, \\ A_2\left(\mu(\omega_2, k_2) + \sigma_{12}|A_1|^2 + \sigma_{22}|A_2|^2 + \cdots\right) = 0.\end{aligned}$$

To leading order, the Lagrange multipliers μ_1 and μ_2 are the dispersion relation for the linearized problem evaluated at (ω_1, k_1) and (ω_2, k_2), respectively. To compute the elements needed for the stability analysis, we need the functionals $\mathscr{A}_j$ and $\mathscr{B}_j$. To leading order they are

$$\mathscr{A}_j(\omega, k) = -\frac{\partial}{\partial\omega_j}\mu_j(\omega, k)|A_j|^2 + \cdots \quad \text{and} \quad \mathscr{B}_j(\omega, k) = -\frac{\partial}{\partial k_j}\mu_j(\omega, k)|A_j|^2 + \cdots,$$

where $(\omega, k) := (\omega_1, \omega_2, k_1, k_2)$. These expressions are verified by substituting (6.1) into the functionals (4.6). Using these expressions we compute

$$\frac{\delta\mathscr{A}}{\delta\omega} = \begin{bmatrix} \frac{\partial\mathscr{A}_1}{\delta\omega_1} & \frac{\partial\mathscr{A}_1}{\delta\omega_2} \\ \frac{\partial\mathscr{A}_2}{\delta\omega_1} & \frac{\partial\mathscr{A}_2}{\delta\omega_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial\omega_1}a_1|A_1|^2 + a_1\frac{\partial}{\partial\omega_1}|A_1|^2, & a_1\frac{\partial}{\partial\omega_2}|A_1|^2 \\ a_2\frac{\partial}{\partial\omega_1}|A_2|^2, & \frac{\partial}{\partial\omega_2}a_2|A_2|^2 + a_2\frac{\partial}{\partial\omega_2}|A_2|^2 \end{bmatrix} + \cdots,$$

where $a_j = -\frac{\partial}{\partial\omega_j}\mu_j$, $j = 1, 2$. Now apply the standing wave limit to this matrix,

(6.4)
$$\omega_2 \to \omega_1 := \omega, \quad k_2 \to -k_1 := -k, \quad |A_2| \to |A_1| := |A|,$$

to find

$$\frac{\delta\mathscr{A}}{\delta\omega} = -D_\omega^2\Lambda^{-1} + D_{\omega\omega}|A|^2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \cdots,$$

where $D(\omega,k)=\lim_{\to \mathrm{SWs}}\mu_1(\omega_1,k_1)=\lim_{\to \mathrm{SWs}}\mu_2(\omega_2,k_2)$, and

$$\Lambda:=\begin{bmatrix}a & b\\ b & a\end{bmatrix}=\lim_{\to \mathrm{SWs}}\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{bmatrix}.$$

Similarly we find

$$\frac{\delta\mathscr{A}}{\delta k}=-D_\omega D_k\,\Lambda^{-1}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}+D_{\omega k}|A|^2\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}+\cdots,$$

$$\frac{\delta\mathscr{B}}{\delta\omega}=\left(\frac{\delta\mathscr{A}}{\delta k}\right)^T=-D_\omega D_k\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\Lambda^{-1}+D_{\omega k}|A|^2\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}+\cdots,$$

and so

$$\frac{\delta\mathscr{A}}{\delta k}+\frac{\delta\mathscr{B}}{\delta\omega}=-\frac{2a}{|\Lambda|}D_\omega D_k\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}+2D_{\omega k}|A|^2\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}+\cdots.$$

For the third term in the matrix (5.7),

$$\frac{\delta\mathscr{B}}{\delta k}=-D_k^2\frac{\Lambda}{|\Lambda|}+D_{kk}|A|^2\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}+\cdots.$$

Now, the stability quartic (5.8) in the standing wave limit takes the form

$$\Delta(\lambda,\alpha)=a_4\lambda^4+a_3\lambda^3+a_2\lambda^2+a_1\lambda+a_0$$

with

$$\begin{aligned}
a_4&=\det\left(\frac{\delta A}{\delta\omega}\right)=\frac{D_\omega^4}{|\Lambda|}-\frac{2a}{|\Lambda|}D_\omega^2D_{\omega\omega}|A|^2+\cdots,\\
a_3&=\mathrm{i}\alpha\mathrm{Tr}\left(\frac{\delta A}{\delta\omega}^{\#}\left(\frac{\delta A}{\delta k}+\frac{\delta B}{\delta\omega}\right)\right)=0,\\
a_2&=-\alpha^2\det\left(\frac{\delta A}{\delta k}+\frac{\delta B}{\delta\omega}\right)-\alpha^2\mathrm{Tr}\left(\frac{\delta A}{\delta\omega}^{\#}\frac{\delta B}{\delta k}\right)\\
&=-\alpha^2\left(-\frac{2}{|\Lambda|}D_\omega^2D_k^2+\frac{2a}{|\Lambda|}(-D_\omega^2D_{kk}-D_k^2D_{\omega\omega}+4D_\omega D_kD_{\omega k})|A|^2+\cdots\right)\\
a_1&=(\mathrm{i}\alpha)^3\mathrm{Tr}\left(\frac{\delta B}{\delta k}^{\#}\left(\frac{\delta A}{\delta k}+\frac{\delta B}{\delta\omega}\right)\right)=0,\\
a_0&=\alpha^4\det\left(\frac{\delta B}{\delta k}\right)=\alpha^4\left(\frac{D_k^4}{|\Lambda|}-\frac{2a}{|\Lambda|}D_k^2D_{kk}|A|^2+\cdots\right).
\end{aligned}$$

Let

$$\delta = -D_\omega^2 D_{kk} - D_k^2 D_{\omega\omega} + 2D_\omega D_k D_{\omega k} = \det \begin{bmatrix} D_{\omega\omega} & D_{\omega k} & D_\omega \\ D_{k\omega} & D_{kk} & D_k \\ D_\omega & D_k & 0 \end{bmatrix}. \tag{6.5}$$

Then, to summarize, the stability quartic in the standing wave limit is

$$\Delta(\lambda,\alpha) = \det\left[\lambda^2 \frac{\delta\mathscr{A}}{\delta\omega} + \mathrm{i}\alpha\lambda\left(\frac{\delta\mathscr{A}}{\delta k} + \frac{\delta\mathscr{B}}{\delta\omega}\right) + (\mathrm{i}\alpha)^2 \frac{\delta\mathscr{B}}{\delta k}\right] = a_4\lambda^4 + a_2(\mathrm{i}\alpha)^2\lambda^2 + a_0(\mathrm{i}\alpha)^4$$

with

$$\begin{aligned} a_4 &= +\frac{D_\omega^4}{|\Lambda|} - \frac{2a}{|\Lambda|} D_\omega^2 D_{\omega\omega}|A|^2 + \cdots, \\ a_2 &= -\frac{2}{|\Lambda|} D_\omega^2 D_k^2 + \frac{2a}{|\Lambda|}(\delta + 2D_\omega D_k D_{\omega k})|A|^2 + \cdots, \\ a_0 &= +\frac{D_k^4}{|\Lambda|} - \frac{2a}{|\Lambda|} D_k^2 D_{kk}|A|^2 + \cdots. \end{aligned}$$

Apply the stability-instability classification in section 5.1, which requires the expressions

$$\begin{aligned} a_0 a_4 &= \frac{D_\omega^4 D_k^4}{|\Lambda|^2} + \cdots > 0, \\ -a_2 a_4 &= \frac{2}{|\Lambda|^2} D_k^2 D_\omega^6 + \cdots > 0, \\ a_2^2 - 4a_0 a_4 &= -16\, a\delta\, \frac{D_\omega^2 D_k^2}{|\Lambda|^2}|A|^2 + \cdots . \end{aligned}$$

Hence, from the stability-instability classification in section 5.1, if we assume the conditions

$$D_\omega \neq 0, \quad D_k \neq 0, \quad \det(\Lambda) \neq 0, \quad a \neq 0, \quad \text{and} \quad \delta \neq 0$$

are satisfied, we can conclude, for $|A|$ sufficiently small, that the stability quartic (5.7) has an unstable eigenvalue if and only if

$$a\delta > 0.$$

A significant feature of this result is that the instability of standing waves is independent of the standing wave frequency correction. To see this, go back to (6.3) and take the standing wave limit,

$$\begin{aligned} A_1\left(D(\omega,k) + a\,|A_1|^2 + b\,|A_2|^2 + \cdots\right) &= 0, \\ A_2\left(D(\omega,k) + b\,|A_1|^2 + a\,|A_2|^2 + \cdots\right) &= 0. \end{aligned} \tag{6.6}$$

Hence

$$\begin{aligned} \omega^{TW} &= \omega_0 - \frac{a}{D_\omega}|A_1|^2 + \cdots, \quad A_2 = 0, \\ \omega^{SW} &= \omega_0 - \frac{(a+b)}{D_\omega}|A_1|^2 + \cdots, \quad A_2 = A_1. \end{aligned}$$

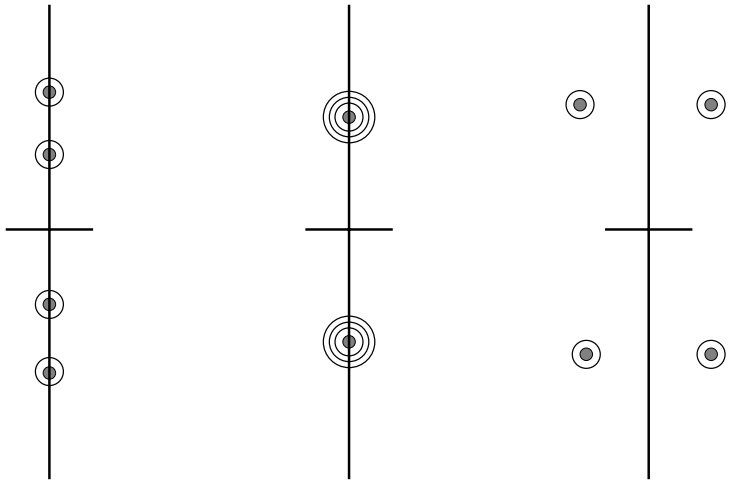


FIG. 2. *Schematic of the position of the eigenvalues at the threshold of instability for weakly nonlinear standing waves.*

Although the weakly nonlinear correction to the frequency for standing waves depends on b, the modulation instability at low amplitude is independent of b. Effectively, the stability problem decouples at low amplitude.

To confirm that a weakly nonlinear traveling wave has the same instability condition as the standing wave, we can compare $a\delta > 0$ with the Whitham condition. If $D(\omega_0(k), k) = 0$ and $D_\omega \neq 0$, then $D_\omega \omega_0'(k) + D_k = 0$ and so

$$\omega_0''(k) = \frac{1}{D_\omega^3} \det \begin{bmatrix} D_{\omega\omega} & D_{\omega k} & D_\omega \\ D_{k\omega} & D_{kk} & D_k \\ D_\omega & D_k & 0 \end{bmatrix} = \frac{1}{D_\omega^3} \delta.$$

From the above expression for the frequency of the traveling waves, ω^{TW}, we have that $\omega_2 = -a/D_\omega$ and so

$$\text{sign}(a\delta) = \text{sign}\left(-\omega_2 D_\omega\, \omega_0''(k) D_\omega^4\right) = -\text{sign}(\omega_0''(k)\omega_2),$$

showing that $a\delta > 0$ is equivalent to $\omega_0''(k)\omega_2 < 0$.

This instability condition for weakly nonlinear standing waves agrees with the instability condition derived by Knobloch and Pierce [18], obtained from the coupled NLS equations with mean-field terms for counterpropagating waves.

For finite-amplitude standing waves, the instability of standing waves will in general differ from the instability of traveling waves. It is an interesting open problem to determine precisely how this instability will change for finite amplitude standing waves. It can be studied either by carrying the amplitude expansion to the next order or numerically.

There is another subtle difference between the traveling wave instability and standing wave instability, which shows up even for weakly nonlinear standing waves. For the standing wave, the unstable subspace is twice as large as the case of traveling waves. A schematic is shown in Figure 2. This figure shows the temporal eigenvalues of the linear stability problem for weakly nonlinear standing waves, with

instability arising through a collision of eigenvalues of opposite signature. Note that each eigenvalue is double, and so the unstable subspace is four-dimensional (whereas for traveling waves it is two-dimensional). The multiple eigenvalues will not persist for finite-amplitude standing waves, suggesting that the behavior of the instability for finite-amplitude standing waves will be more dramatic than traveling waves in general and weakly nonlinear standing waves in particular.

6.1. Example: Calculations for a scalar nonlinear wave equation. An elementary example of the theory is obtained by considering the scalar nonlinear wave equation: (1.1) with $m = 1$. Let $V(u)$ be any smooth function with leading Taylor expansion

$$V(u) = \frac{1}{2}a_1u^2 + \frac{1}{3}a_2u^3 + \frac{1}{4}a_3u^4 + \cdots, \quad a_1 > 0.$$

Then a straightforward calculation leads to the reduced Lagrangian

$$\overline{\mathscr{L}} = \mu_1|A_1|^2 + \mu_2|A_2|^2 + \tfrac{1}{2}\sigma_{11}|A|^4 + \sigma_{12}|A_1|^2|A_2|^2 + \tfrac{1}{2}\sigma_{22}|A_2|^4 + \cdots$$

with $\mu_j(\omega, k) = k_j^2 - \omega_1^2 + a_1$, $\sigma_{11} = \sigma_{22} = -\frac{5}{3a_1}a_2^2 + \frac{9}{4}a_3$, and

$$\sigma_{12} = -2a_2^2\left(\frac{1}{a_1} + \frac{1}{a_1 - (\omega_1+\omega_2)^2 + (k_1+k_2)^2} + \frac{1}{a_1 - (\omega_1-\omega_2)^2 + (k_1-k_2)^2}\right) + 3a_3.$$

Computing the parameter Jacobian (4.8) and taking the limit $\omega_2 \to \omega_1 := \omega$ and $k_2 \to -k_1 := -k$ we find

$$D(\omega, k) = k^2 - \omega^2 + a_1, \quad \delta = -8a_1, \quad \text{and} \quad a = \lim_{\to \mathrm{SWs}} \sigma_{11} = \sigma_{11},$$

and so $a\delta = -8a_1(-\frac{5}{3a_1}a_2^2 + \frac{9}{4}a_3)$. Hence both traveling waves and standing wave solutions of (1.1) are unstable in the weakly nonlinear limit whenever

$$20a_2^2 - 27a_1a_3 > 0 \quad \text{and} \quad a_1 > 0.$$

The instability of finite-amplitude standing wave solutions of even this scalar nonlinear wave equation is an open problem, but the theory of this paper can be applied, given (either numerical or analytic) expressions for the finite-amplitude standing waves.

7. Small divisors and the equivariant Lyapunov center theorem. The obstacle to a rigorous proof of the existence of smooth families of standing waves and the linear stability theory is a potential small divisor problem. This issue can be illustrated by considering the scalar version of the nonlinear wave equation (1.1), which can be written

$$u_{tt} - u_{xx} + a_1u = V'(u) - a_1u,$$

where $a_1 = V''(0)$ is some positive real number. In application of the implicit function theorem to the existence of standing waves, linear systems of the type of the left-hand side have to be inverted on the complement of its kernel, on a space of space-time periodic functions. Such systems can be written in the general form

$$V_t = \mathbf{L}V + \mathbf{f}(x,t), \quad \mathbf{L} = \begin{bmatrix} 0 & I \\ \partial_{xx} - a_1 & 0 \end{bmatrix},$$

where $\mathbf{f}(x,t)$ is a vector-valued periodic function of x and t. Now, the spectrum of $\mathbf{L}$ on a space of 2π-periodic functions is

$$\lambda_n = \mathrm{i}\sqrt{n^2 + a_1} := \mathrm{i}\omega_n, \quad n \in \mathbb{Z}.$$

The spectrum consists of a countable number of purely imaginary eigenvalues. Now, a_1 can be chosen so that

$$\lambda_1 j - \lambda_n \neq 0, \quad j \geq 1,\ n \neq 1,$$

which is the usual nonresonance condition of the Lyapunov center theorem. However, the distance $|\omega_1 j - \omega_n|$ may tend to zero as j, n tend to infinity, creating a small divisor problem. In other words, a frequency ω_n when n is large enough may get arbitrarily close to a resonant multiple of ω_1.

Effectively, what is needed is a version of the Lyapunov center theorem in infinite dimensions. The first result of this type is due to Craig and Wayne [12] and uses Nash–Moser theory to overcome the small divisor problem. However, the resulting branches of periodic solutions are not smooth but lie on a Cantor-like subset of parameter space.

By imposing the stronger diophantine condition on the frequencies

$$|\omega j - \omega_n| \geq \frac{\gamma}{j}, \quad j \geq 1, \quad n \geq 2, \quad \text{for some } \gamma > 0,$$

Bambusi [1] and Bambusi and Paleari [4] prove that the ordinary implicit function can be used, and this leads to partial smoothness of branches of periodic solutions.

These results are encouraging, but they still do not provide sufficient smoothness for the limits required in section 5. Moreover, the present analysis uses symmetry in a central way, and so an equivariant version of the Lyapunov center theorem [22] generalized to infinite dimensions would be required. Some intriguing results in this direction are given by Bambusi and Gaeta [3].

8. Instability of standing water waves. One of the most interesting examples of standing waves is standing water waves. These waves are most commonly observed and studied in the context of sloshing of fluid in a vessel. However, they are also a central part of pattern formation in the open ocean. The first nonlinear theory for standing waves was proposed by Rayleigh [27]. Indeed, he showed that they arise naturally along with traveling waves in any analysis of weakly nonlinear space and time periodic water waves. Since Rayleigh's work there has been a wide range of analytical and numerical theories for standing waves; see [11] for a list of references.

Recently, progress has been made in developing a rigorous theory for existence of standing waves. In finite depth small divisors arise, and a rigorous proof in this case for weakly nonlinear standing waves has been given by Plotnikov and Toland [26]. The proof uses a Nash–Moser framework, and therefore the branches of periodic solutions are not smooth, certainly not smooth enough to embed them in a higher parameter family. Surprisingly, the problem in infinite depth is more difficult. The kernel of the linearized problem is infinite-dimensional and the dispersion relation is algebraic, but recent significant progress has been made [14, 15].

For weakly nonlinear standing waves, stability results have been reported by Okamura [23] and Knobloch and Pierce [18] using modulation equations. The paper [18] gives the first correct analytic result for instability of weakly nonlinear standing waves. For finite-amplitude standing waves, the only results in the literature on the linear

stability is the work of Mercer and Roberts [21]. There are several interesting results in [21]. They compute long-wave instabilities at finite amplitude (called subharmonic instabilities there). They compute the action ($\mathscr{A}_1$ here) as a function of frequency (ω_1 here) and show that there is a point where $\partial_{\omega_1}\mathscr{A}_1 = 0$. This latter point will have an effect on the modulation instability at that point.

The theory of this paper suggests a new approach to the numerical computation of standing waves. The standing waves should be embedded in a four-parameter family and then the elements of (4.8) computed to study a wider range of stability properties. This embedding would not increase computation time (for standing waves or four-parameter two-wave interaction, the solution is expanded in a double Fourier series) but would increase the range of parameter space. However, it is only the parameter space *near the standing waves* that is of interest, and the computation of the functionals and their parameter dependence is a secondary calculation.

On a formal level one can draw a number of conclusions about the instability of standing water waves from the theory reported in this paper. First, the water-wave problem can be formulated as a multisymplectic system [6, 7] and the framework of this paper applied. For weakly nonlinear standing waves the conclusion for deep water is immediate: weakly nonlinear standing waves are unstable to a Benjamin–Feir instability in the same way that traveling waves are unstable. This is in agreement with the results of [18]. Further numerical calculations would be needed to test the theory of this paper at large amplitude to compare with and extend the results of [21].

The case of standing waves in finite depth may also have mean flow generation. For traveling waves, it is well known that reducing the depth creates a mean flow that stabilizes the Benjamin–Feir instability. Therefore an interesting open problem would be to determine the effect of mean flow on the stability of standing water waves in finite depth. Results obtained using modulation equations by Knobloch and Pierce [18] suggest that the weakly nonlinear finite-depth standing waves are affected by mean flow in exactly the same way as traveling waves. However, the role of mean flow in the stability of *finite-amplitude* standing waves is an open question.

9. Concluding remarks. The basic strategy here—embed a multiperiodic, say, N-periodic, pattern in an N-wave interaction with $2N$ parameters, compute parameter Jacobians, then take a limit to the original N-parameter wave to obtain stability information—has wider applicability. For example, in [10] this idea is generalized to determine stability conditions for short-crested Stokes waves in three space dimensions.

Short-crested Stokes waves are solutions of the form

$$Z(x,y,t) = \widehat{Z}(\theta_1,\theta_2), \quad \theta_1 = kx + \ell y + \omega t, \quad \theta_2 = kx - \ell y + \omega t.$$

They are three-parameter doubly periodic solutions and have been widely studied by oceanographers and engineers because they arise as a secondary bifurcation from classical Stokes waves and are known to influence sand transport. Short-crested waves are a generalization of standing waves in the sense that they can be characterized as synchronized oblique traveling waves, and in the limit as the angle between the two waves becomes zero they reduce to standing waves.

There are a number of open questions in the fluid mechanics literature about their stability. The theory of this paper generalizes to this problem in a straightforward way. The short-crested wave is embedded in the six-parameter two-phase wavetrain $Z(x,y,t) = \widehat{Z}(\theta_1,\theta_2)$ but with

$$\theta_j = \omega_j t + k_j x + \ell_j y + \theta_j^o, \quad j = 1,2.$$

These waves, when characterized by a constrained variational principle, generate the 6×6 matrix

$$\begin{pmatrix} \frac{\delta \mathscr{A}}{\delta \omega} & \frac{\delta \mathscr{A}}{\delta k} & \frac{\delta \mathscr{A}}{\delta \ell} \\ \frac{\delta \mathscr{B}}{\delta \omega} & \frac{\delta \mathscr{B}}{\delta k} & \frac{\delta \mathscr{B}}{\delta \ell} \\ \frac{\delta \mathscr{C}}{\delta \omega} & \frac{\delta \mathscr{C}}{\delta k} & \frac{\delta \mathscr{C}}{\delta \ell} \end{pmatrix}.$$

Proceeding as in section 5, a stability theory can be developed that predicts all long-wave instabilities using the entries of the above matrix. The complete details are given elsewhere [10].

Another generalization of interest is the study of the stability of *hexagonal* ocean patterns by embedding them in a three-phase wave train—with nine parameters—and then following the strategy in this paper to develop a theory for long-wave instability.

Appendix A. $O(2)$-invariant Hamiltonian systems and the spherical pendulum. Some insight into the geometry of nonlinear wave equations on the real line with periodic boundary conditions and x-reflection symmetry can be gained by examining the finite-dimensional analogue. This analogy is useful for illustrating the toral structure, but the analogy can be taken only so far, since the most interesting consequence for nonlinear wave equations—the geometry of modulational instability—is absent in finite dimensions.

Consider a Hamiltonian system on $\mathbb{R}^4$ with standard symplectic operator $\mathbf{J}$,

$$\mathbf{J}U_t = \nabla H(U), \quad U \in \mathbb{R}^4, \tag{A.1}$$

and suppose H is smooth and the system is $\mathbf{O}(2)$-equivariant. That is, there is a representation Γ of $\mathbf{O}(2)$ acting on $\mathbb{R}^4$ such that H is Γ-invariant and the action of Γ is symplectic [22]. Near a Γ-invariant equilibrium there are generically two classes of periodic solutions: traveling waves and standing waves [22].

The canonical example of an $\mathbf{O}(2)$-equivariant Hamiltonian system is the spherical pendulum (see [22, 5]), and it is sufficient to restrict attention to this example. For the spherical pendulum, the geometry and nature of solutions can be seen explicitly. The traveling waves are the conical pendulum solutions, and standing waves are the planar pendulum solutions. There are two traveling waves (one rotating clockwise and one rotating counterclockwise), and there is an $\mathbf{SO}(2)$ orbit of standing waves (the plane of motion of the planar pendulum can be rotated around). The standing waves of the spherical pendulum have zero angular momentum.

There is another well-known class of solutions of the spherical pendulum: the toral solutions which have a smoothly varying rotation number and nonzero angular momentum. These are sometimes called *relative periodic orbits*. Physically, they correspond to a precessing planar pendulum.

The solutions of the spherical pendulum can be usefully viewed in the energy-momentum space, as shown in Figure 3, where E represents the value of the energy and I represents the value of the momentum. The standing waves are along the $I = 0$ axis. The traveling waves correspond to minima of the energy restricted to level sets of the momentum and lie along the two curves shown. There are no solutions associated with (I, E) values below the traveling wave curves, and the region between the two curves excluding $I = 0$ is filled with toral solutions with smoothly varying rotation number.

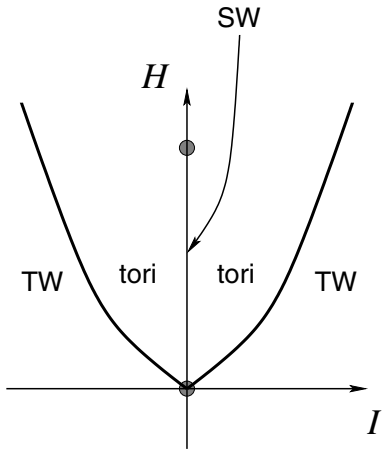


FIG. 3. *Schematic of the energy-momentum space for the spherical pendulum. The two highlighted points on the line of zero momentum are the two equilibria (vertical up and vertical down) of the spherical pendulum.*

A periodic solution, of period $T = \frac{2\pi}{\omega}$, of a finite-dimensional Hamiltonian system such as (A.1) can be characterized by a variational principle: a critical point of H restricted to level sets of the action,

$$\mathscr{A} = \frac{1}{2\pi}\int_0^{2\pi} A(U)\,\mathrm{d}\theta, \quad A(U) = \frac{1}{2}\langle \mathbf{J}U_\theta, U\rangle, \tag{A.2}$$

with the frequency, ω, a Lagrange multiplier. The Lagrange necessary condition for this variational principle is

$$\nabla H - \omega \nabla A = \nabla H - \omega \mathbf{J} U_\theta = 0, \quad \theta = \omega\, t + \theta^o.$$

A standing wave state would have the additional requirement that it is invariant under reflection (the reflection subgroup of $\mathbf{O}(2)$).

Now we come to the main point of this section. Can the limit $I \to 0$ be taken in the class of toral solutions leading to a standing wave?

Consider embedding the standing wave in a toral solution. Let $U(t) = \widehat{U}(\theta_1, \theta_2)$ with $\theta_j = \omega_j t + \theta_j^o$ for $j = 1, 2$. Then a variational characterization is again natural, and the Lagrange necessary condition is

$$\nabla H - \omega_1 \nabla A_1 - \omega_2 \nabla A_2 = \nabla H(\widehat{Z}) - \omega_1 \mathbf{J}\partial_{\theta_1}\widehat{U} - \omega_2 \mathbf{J}\partial_{\theta_2}\widehat{U} = 0, \quad \widehat{U} : \mathbb{T}^2 \to \mathbb{R}^4. \tag{A.3}$$

It follows from standard Lagrange multiplier theory that a state satisfying (A.3) is nondegenerate when

$$\det \begin{bmatrix} \frac{\partial \omega_1}{\partial I_1} & \frac{\partial \omega_1}{\partial I_2} \\ \frac{\partial \omega_2}{\partial I_1} & \frac{\partial \omega_2}{\partial I_2} \end{bmatrix} \neq 0, \tag{A.4}$$

where I_1 and I_2 represent values of the two actions. Now, solutions of this variational principle are smooth functions of the frequencies, away from the singularities (the

two equilibrium points and the two branches of traveling waves). This can be seen explicitly using the integrability of the spherical pendulum [13], but it follows more generally from symmetry. The more surprising smoothness result is that the frequency map (A.4) exists—in the limit from a toral state to the line $I = 0$—and is well defined (away from the two equilibrium points). Suppose I_2 represents the angular momentum in (A.4); then Horosov [13] proves that

$$\lim_{I_2\to 0} \det \begin{bmatrix} \frac{\partial\omega_1}{\partial I_1} & \frac{\partial\omega_1}{\partial I_2} \\ \frac{\partial\omega_2}{\partial I_1} & \frac{\partial\omega_2}{\partial I_2} \end{bmatrix} = \det \begin{bmatrix} \left(\frac{\partial\omega_1}{\partial I_1}\right)^0 & 0 \\ 0 & \left(\frac{\partial\omega_2}{\partial I_2}\right)^0 \end{bmatrix} = \left(\frac{\partial\omega_1}{\partial I_1}\right)^0 \left(\frac{\partial\omega_2}{\partial I_2}\right)^0 \neq 0. \tag{A.5}$$

Although the limit $I_2 \to 0$ results in a degeneration from a toral solution to a periodic solution, the toral frequency map does not degenerate! This nondegeneracy arises because the planar pendulum solutions lie on a torus and so the tangent space of the *manifold* of standing waves is two-dimensional. The first term in the product on the right-hand side of (A.5) is the change in frequency with amplitude (or energy) of the planar pendulum, and the second term in the product just says that the second frequency changes smoothly in going from negative to positive angular momentum (or vice versa).

The above geometry is also central to the standing wave problem associated with nonlinear wave equations on the real line. In the second variational principle in section 4, the standing wave is being embedded in a generalized multisymplectic relative-periodic orbit. In other words, geometrically there should be a smooth variation of the parameters. This would be true of any finite-dimensional approximation of the standing wave problem. In the limit as the number of modes goes to infinity, the small divisor issue again appears.

Appendix B. Multisymplectic Noether theory and momentum conservation. In this appendix, classical Noether theory is applied to the Lagrangian in the canonical multisymplectic form (1.6). A Lagrangian $\mathscr{L} = \int\int \mathrm{L}(Z, Z_t, Z_x)\,\mathrm{d}x\wedge\mathrm{d}t$, with $Z(x,t)$ vector valued, which does not depend explicitly on x, has a symmetry with generator $\mathbf{v} = \frac{\partial}{\partial x}$. Using Noether's theorem [24, section 4.4], this symmetry generates a conservation law,

$$\mathrm{I}(Z)_t + \mathrm{P}(Z)_x = 0$$

with

$$\mathrm{I}(Z) = \left\langle Z_x, \frac{\partial \mathrm{L}}{\partial Z_t} \right\rangle \quad \text{and} \quad \mathrm{P}(Z) = \left\langle Z_x, \frac{\partial \mathrm{L}}{\partial Z_x} \right\rangle - \mathrm{L}(Z, Z_t, Z_x),$$

where $\langle\cdot,\cdot\rangle$ is the standard inner product on $\mathbb{R}^n$. Applying these formulas to L in canonical form,

$$\mathrm{L}(Z, Z_t, Z_x) = \tfrac{1}{2}\langle \mathbf{M}Z_t, Z\rangle + \tfrac{1}{2}\langle \mathbf{K}Z_x, Z\rangle - S(Z),$$

leads to

$$\mathrm{I}(Z) = \tfrac{1}{2}\langle \mathbf{M}Z_x, Z\rangle, \quad \mathrm{P}(Z) = S(Z) - \tfrac{1}{2}\langle \mathbf{M}Z_t, Z\rangle.$$

This conservation law can be confirmed by direct calculation,

$$\begin{aligned}
\mathrm{I}(Z)_t &= \frac{1}{2}\langle \mathbf{M}Z_{xt}, Z\rangle + \frac{1}{2}\langle \mathbf{M}Z_x, Z_t\rangle \\
&= \frac{\partial}{\partial x}\left(\frac{1}{2}\langle \mathbf{M}Z_t, Z\rangle\right) - \langle \mathbf{M}Z_t, Z_x\rangle \quad \text{(using skew-symmetry of } \mathbf{M}\text{)} \\
&= \frac{\partial}{\partial x}\left(\frac{1}{2}\langle \mathbf{M}Z_t, Z\rangle\right) - \langle \nabla S(Z) - \mathbf{K}Z_x, Z_x\rangle \quad \text{(substituting for } \mathbf{M}Z_t \text{ using (3.1))} \\
&= \frac{\partial}{\partial x}\left(\frac{1}{2}\langle \mathbf{M}Z_t, Z\rangle - S(Z)\right) \quad \text{(using skew-symmetry of } \mathbf{K}\text{)},
\end{aligned}$$

and hence $\mathrm{I}(Z)_t + (S(Z) - \frac{1}{2}\langle \mathbf{M}Z_t, Z\rangle)_x = 0$.

REFERENCES

[1] D. Bambusi, *Lyapunov center theorem for some nonlinear PDEs: A simple proof*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 29 (2000), pp. 823–837.

[2] D. Bambusi, A. Carati, and A. Ponno, *The nonlinear Schrödinger equation as a resonant normal form*, Discrete Contin. Dyn. Syst. Ser. B, 2 (2002), pp. 109–128.

[3] D. Bambusi and C. Gaeta, *On the persistence of invariant tori and a theorem of Nekhoroshev*, Math. Phys. Electron. J., 8 (2002), Paper 1.

[4] D. Bambusi and S. Paleari, *Families of periodic solutions of resonant PDEs*, J. Nonlinear Sci., 11 (2001), pp. 69–87.

[5] L. M. Bates and R. H. Cushman, *Global Aspects of Classical Integrable Systems*, Birkhäuser-Verlag, Basel, 1997.

[6] T. J. Bridges, *Periodic patterns, linear instability, symplectic structure and mean-flow dynamics for 3D surface waves*, Philos. Trans. Roy. Soc. London Ser. A, 354 (1996), pp. 533–574.

[7] T. J. Bridges, *Multi-symplectic structures and wave propagation*, Math. Proc. Cambridge Philos. Soc., 121 (1997), pp. 147–190.

[8] T. J. Bridges, F. Dias, and D. Menasce, *Steady three-dimensional water-wave patterns on a finite-depth fluid*, J. Fluid Mech., 436 (2001), pp. 145–175.

[9] T. J. Bridges and F. E. Laine-Pearson, *Multisymplectic relative equilibria, multiphase wavetrains, and coupled NLS equations*, Stud. Appl. Math., 107 (2001), pp. 137–155.

[10] T. J. Bridges and F. E. Laine-Pearson, *The Long-Wave Instability of Short-Crested Waves, via Embedding in the Oblique Two-Wave Interaction*, submitted.

[11] P. J. Bryant and M. Stiassne, *Different forms for nonlinear standing waves in deep water*, J. Fluid Mech., 272 (1994), pp. 135–156.

[12] W. Craig and C. E. Wayne, *Newton's method and periodic solutions of nonlinear wave equations*, Comm. Pure Appl. Math., 46 (1993), pp. 1409–1501.

[13] E. Horosov, *Perturbations of the spherical pendulum and Abelian integrals*, J. Reine Angew. Math., 408 (1990), pp. 114–135.

[14] G. Iooss, *Semi-analytic theory of standing waves in deep water for several dominant modes*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 455 (1999), pp. 3513–3526.

[15] G. Iooss, *On the standing wave problem in deep water*, J. Math. Fluid Mech., 4 (2002), pp. 155–185.

[16] P. Kirrmann, G. Schneider, and A. Mielke, *The validity of modulation equations for extended systems with cubic nonlinearities*, Proc. Roy. Soc. Edinburgh Sect. A, 122 (1992), pp. 85–91.

[17] E. Knobloch and J. D. Gibbon, *Coupled NLS equations for counter propagating waves in systems with reflection symmetry*, Phys. Lett. A, 145 (1991), pp. 353–356.

[18] E. Knobloch and R. D. Pierce, *On the modulational instability of travelling and standing water waves*, Phys. Fluids, 6 (1994), pp. 1177–1190.

[19] S. B. Kuksin, *Analysis of Hamiltonian PDEs*, Oxford University Press, Oxford, UK, 2000.

[20] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems*, Springer-Verlag, Berlin, 1989.

[21] G. N. Mercer and A. J. Roberts, *Standing waves in deep water. Their stability and extreme form*, Phys. Fluids A, 4 (1992), pp. 259–269.

[22] J. A. Montaldi, R. M. Roberts, and I. N. Stewart, *Periodic solutions near equilibria of symmetric Hamiltonian systems*, Philos. Trans. Roy. Soc. London Ser. A, 325 (1988), pp. 237–293.
[23] M. Okamura, *Instabilities of weakly nonlinear standing gravity waves*, J. Phys. Soc. Japan, 53 (1984), pp. 3788–3796.
[24] P. J. Olver, *Application of Lie Groups to Differential Equations*, Springer-Verlag, New York, 1986.
[25] R. D. Pierce and C. E. Wayne, *On the validity of mean-field amplitude equations for counterpropagating wavetrains*, Nonlinearity, 8 (1995), pp. 769–799.
[26] P. I. Plotnikov and J. F. Toland, *Nash-Moser theory for standing water waves*, Arch. Ration. Mech. Anal., 159 (2001), pp. 1–83.
[27] Lord Rayleigh, *Deep water waves, progressive or stationary, to the third order approximation*, Proc. Roy. Soc. London A, 91 (1915), pp. 345–353.
[28] C. Sulem and P.-L. Sulem, *The Nonlinear Schrödinger Equation. Self-Focusing and Wave Collapse*, Appl. Math. Sci. 139, Springer-Verlag, New York, 1999.
[29] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley-Interscience, New York, 1974.
[30] L. V. Yakushevich, *Nonlinear Physics of DNA*, 2nd rev. ed., Wiley-VCH, Weinheim, Germany, 2004.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 2121–2143

NONCRYSTALLOGRAPHIC MOTION OF A DISLOCATION AS A FINE MIXTURE OF RECTILINEAR PATHS*

TIZIANA ARMANO† AND PAOLO CERMELLI†

Abstract. In this work we discuss the convergence of an approximation scheme for the solution, near an attractor, of a discontinuous dynamical system arising in the theory of dislocations in crystalline solids. It is well known that dislocations can move only along a finite number of crystallographic directions: in two dimensions, the resulting trajectories are piecewise rectilinear paths. However, in special situations such as near an attractor, dislocations are forced to move along curved paths: we characterize this class of motions as fine mixtures of crystallographic motions, using the notion of generalized curves due to Young, and we explicitly compute the parametrized measure associated to a sequence of polygonals. The result is then used to motivate a simple numerical scheme and show that it is physically consistent. Numerical simulations based on this scheme are also presented and discussed.

Key words. dislocation motion, dislocations and cracks, force on a dislocation, Young measures

AMS subject classifications. 74B99, 74H15, 65P99, 74E15

DOI. 10.1137/S003613990343063X

1. Introduction.

The goal of this work is to study a special problem arising in the theory of defects in crystalline materials.

Specifically, we are interested in studying the motion of a screw dislocation in a cylindrical crystalline elastic body. Dislocations are the most common line defects in crystals, and their mobility is responsible for the plastic behavior and the ductility of most metals [6], [7], [8].

We use here the model developed by Cermelli and Gurtin [2] to describe the motion of a dislocation. Under some simplified hypotheses, the motion of a rectilinear dislocation can be described in terms of the motion of the intersection point of the defect line with an orthogonal plane. Peculiar to crystalline materials is the fact that such points are restricted to move along special directions, the so-called glide or slip directions (glide and slip planes in a general three-dimensional framework).

In elastic materials, a state of stress induces a force on a dislocation, the so-called Peach–Köhler force (see [2], [3], [4], and [9]) and the defect moves parallel to the direction on which the projection of this force is maximal (maximum dissipation criterion). Hence, the motion of a dislocation can be viewed as the solution of a dynamical system in a plane domain, obtained by projecting the Peach–Köhler force on the crystallographic directions. Since the number of such directions in a crystal is finite, it follows that the trajectories are piecewise rectilinear paths.

The general properties of this dynamical system have been studied in [2]; we focus here on a special situation, namely, the motion near a curve S which is an attractor. The dislocation is attracted by S: when it reaches it, it cannot escape (since it would violate the maximum dissipation criterion), but it cannot move along S either, since it would, in general, violate the crystallographic restriction on the direction of motion.

*Received by the editors June 25, 2003; accepted for publication (in revised form) February 9, 2004; published electronically September 14, 2004. This work was supported by the Italian M.U.R.S.T. research project "Modelli matematici per la scienza dei materiali" (Cofin 2002).
http://www.siam.org/journals/siap/64-6/43063.html
†Dipartimento di Matematica, Università di Torino, Via Carlo Alberto 10, I–10123 Torino, Italy (tiziana.armano@unito.it, paolo.cermelli@unito.it).

Hence, it seems natural to approximate the motion of the defect on S by a sequence of polygonals, which are piecewise parallel to the crystallographic directions but do not necessarily satisfy the maximum dissipation criterion at all times.

The main result of this paper is the proof that if such a sequence is a maximizing sequence for the dissipation, it converges to a unique smooth motion on S, which we refer to as *fine cross slip*.[1]

To study the limits of maximizing sequences we make use of the notion of generalized curves due to Young, in their recent formulation known as parametrized (or Young) measures in the literature on the calculus of variations. Young measures provide a richer characterization of finely oscillating sequences than their weak limits. We compute the Young measure associated to sequences of polygonals maximizing the dissipation, and we characterize fine cross slip as a fine mixture of crystallographic rectilinear motions, with weights depending on the direction of the attractor S.

From the mathematical point of view, the problem is equivalent to finding the solution of a dynamical system, in a neighborhood of an attracting curve at which the velocity field is multivalued. If the solution is computed numerically by any time-discretized scheme, the trajectory oscillates finely near the attractor, on a polygonal which is only approximately a solution of the original dynamical system. We show that any approximation scheme based on choosing a time step h determines, as $h \to 0$, a maximizing sequence for the dissipation and therefore converges to a unique smooth motion on S, which can be described as a fine mixture of crystallographic motions. Further, the limit motion coincides with the fine cross slip introduced above.

Finally, we present the results of numerical simulations of the motion of a screw dislocation in a domain with a crack or a rigid inclusion, based on the approximation scheme described above.

We also discuss an explicit example for which the velocity field, the attracting curves, and the motion by fine cross slip can be computed analytically. The numerical results support our theoretical considerations: the solution near the attracting curve, as computed by classical ODE solvers (such as Euler and Runge–Kutta), shows a good agreement with the limit motion on the attracting curve.

2. Statement of the problem. We summarize in this section the model discussed in [2]. Consider an elastic cylinder $B = \Omega \times \mathbb{R}$, where Ω is a domain in $\mathbb{R}^2$. A *screw Volterra dislocation* is a singular displacement field on B which can be constructed by the following ideal procedure [12]: first cut the cylinder B along a vertical half-plane Π, then translate one of the faces along the cut by a constant vertical vector $\boldsymbol{b}$, glue back the faces along Π, and let the cylinder relax to an elastic equilibrium state (Figure 2.1). The resulting displacement field, measured with respect to the initial configuration, is smooth in $B \setminus \Pi$ but is discontinuous across Π with constant jump $\boldsymbol{b}$. The vertical line $\partial\Pi$ is called the *dislocation line*, and $\boldsymbol{b}$ is the *Burgers vector*. To avoid dealing with discontinuous displacement fields, it can be shown that a screw dislocation can be characterized equivalently in terms of a deformation field on $B \setminus \partial\Pi$, singular at $\partial\Pi$. In simple cases, the deformation field generated by a dislocation is independent of the vertical coordinate, and the problem admits a two-dimensional formulation in terms of planar fields on Ω, which are singular at $\boldsymbol{z} = \partial\Pi \cap \Omega$ [2].

Precisely, let Ω be a domain in $\mathbb{R}^2$, with Cartesian coordinates (x, y) and associated basis $(\boldsymbol{e}_1, \boldsymbol{e}_2)$, and let $\boldsymbol{x}$ denote a generic point in Ω.

[1] Fine cross slip of screw dislocations has indeed been experimentally observed (see, e.g., [5] and [7]).

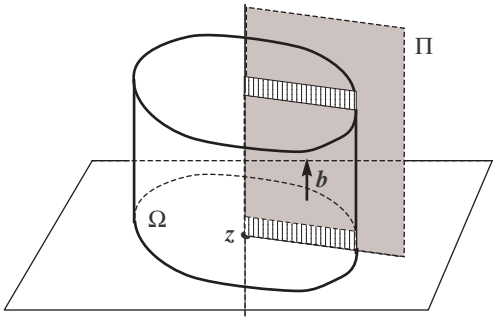


FIG. 2.1. *A screw Volterra dislocation in the cylinder* $\Omega \times \mathbb{R}$.

Fix a defect position $\boldsymbol{z} \in \Omega$ and consider the solution $u : \Omega \to \mathbb{R}$ of the Neumann problem

$$\begin{cases} \Delta u = 0 & \text{in } \Omega, \\ \dfrac{\partial u}{\partial n} = -\boldsymbol{g}_0 \cdot \boldsymbol{n} + \sigma_0 & \text{on } \partial\Omega, \end{cases} \tag{2.1}$$

where Δ is the Laplace operator, $\partial/\partial n$ is the normal time derivative on $\partial\Omega$, $\boldsymbol{n}$ is the outward unit normal to $\partial\Omega$, and

$$\boldsymbol{g}_0 = \boldsymbol{g}_0(\boldsymbol{x}, \boldsymbol{z}) = \frac{b}{2\pi|\boldsymbol{x} - \boldsymbol{z}|^2}\, \boldsymbol{e}_3 \times (\boldsymbol{x} - \boldsymbol{z}), \tag{2.2}$$

where b is a real constant, $\boldsymbol{e}_3 = \boldsymbol{e}_1 \times \boldsymbol{e}_2$ is a unit vector in $\mathbb{R}^3$ orthogonal to the plane containing Ω (so that $\boldsymbol{e}_3 \times (\cdot)$ represents a counterclockwise $\pi/2$-rotation in the Ω-plane), and $\sigma_0 = \sigma_0(\boldsymbol{x})$ is an assigned function on $\partial\Omega$. The field u represents the regular part of the displacement due to the dislocation at $\boldsymbol{z}$, while $\boldsymbol{g}_0$ is related to the singular part of the deformation.

For each fixed $\boldsymbol{z} \in \Omega$, the Neumann problem (2.1) has a unique smooth solution[2] (modulo an additive constant), which we henceforth denote by

$$u = u(\boldsymbol{x}, \boldsymbol{z}), \qquad \boldsymbol{x} \in \Omega. \tag{2.3}$$

Consider now the smooth vector field in Ω,

$$\boldsymbol{\jmath}(\boldsymbol{x}) = b\, \nabla u(\boldsymbol{x}, \boldsymbol{x}) \times \boldsymbol{e}_3, \qquad \boldsymbol{x} \in \Omega, \tag{2.4}$$

where $\nabla u(\boldsymbol{x}, \boldsymbol{x}) = \nabla_{\boldsymbol{x}} u(\boldsymbol{x}, \boldsymbol{z})|_{\boldsymbol{z}=\boldsymbol{x}}$ is the gradient of the solution $u(\boldsymbol{x}, \boldsymbol{z})$ of (2.1), for a dislocation located at $\boldsymbol{z} = \boldsymbol{x}$. The vector field $\boldsymbol{\jmath}(\boldsymbol{x})$ depends only on the domain Ω and the boundary conditions σ_0 and may be identified to the Peach–Köhler force on a dislocation located at $\boldsymbol{x} \in \Omega$.

[2]The solution u of (2.1) can be represented explicitly with the help of Green's formula,

$$u(\boldsymbol{x}, \boldsymbol{z}) = -\int_{\partial\Omega} N(\boldsymbol{x}, \boldsymbol{\xi}(s)) \left(\sigma_0(\boldsymbol{\xi}(s)) - \frac{b}{2\pi|\boldsymbol{\xi}(s) - \boldsymbol{z}|^2}\, \boldsymbol{n}(\boldsymbol{\xi}(s)) \cdot \boldsymbol{e}_3 \times (\boldsymbol{\xi}(s) - \boldsymbol{z}) \right) ds + \mathrm{c},$$

where $N(\boldsymbol{x}, \boldsymbol{\xi})$ is the Neumann function for the domain Ω, $\boldsymbol{\xi}(s)$ is a parametrization of $\partial\Omega$ with arc length s, and c is a constant.

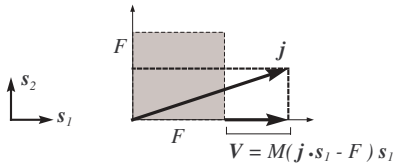


FIG. 2.2. *The definition of the vector field* $\boldsymbol{V}$.

Let now t denote time and $[0,T]$ be the time interval of interest. To study the behavior of a defect under the action of the force (2.4), consider a dislocation motion

$$\boldsymbol{z} : [0,T] \to \Omega.$$

Introducing the (finite) set of *crystallographic directions*

$$\mathcal{C} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n\}$$

with $\boldsymbol{s}_i$ fixed unit vectors in $\mathbb{R}^2$, the basic physical idea is that a dislocation can only move parallel to a crystallographic direction $\boldsymbol{s} \in \mathcal{C}$ on which the projection of the force $\boldsymbol{\jmath} \cdot \boldsymbol{s}$ is maximal, provided this is greater than a given threshold F, the so-called Peierls force (Figure 2.2). Therefore, we write the basic equation governing the motion of a dislocation as

$$\dot{\boldsymbol{z}} = \boldsymbol{V}(\boldsymbol{z}), \quad \boldsymbol{z} \in \Omega, \tag{2.5}$$

where the superposed dot denotes time derivative and where the vector field $\boldsymbol{V}$ is defined by

$$\boldsymbol{V}(\boldsymbol{x}) := \begin{cases} \boldsymbol{0} & \text{if } \boldsymbol{\jmath}(\boldsymbol{x}) \cdot \boldsymbol{s} \le F \ \ \forall \boldsymbol{s} \in \mathcal{C}, \\ M\,(\boldsymbol{\jmath}(\boldsymbol{x}) \cdot \boldsymbol{e}(\boldsymbol{x}) - F)\,\boldsymbol{e}(\boldsymbol{x}) & \text{otherwise,} \end{cases} \tag{2.6}$$

where $M > 0$ and $F \ge 0$ are given constants and $\boldsymbol{e}(\boldsymbol{x}) \in \mathcal{C}$ is determined by the *maximum dissipation criterion*, i.e., the requirement that the projection of $\boldsymbol{\jmath}(\boldsymbol{x})$ on $\boldsymbol{e}(\boldsymbol{x})$ be maximal, i.e.,

$$\boldsymbol{\jmath}(\boldsymbol{x}) \cdot \boldsymbol{e}(\boldsymbol{x}) = \max_{\boldsymbol{s} \in \mathcal{C},\, \boldsymbol{\jmath} \cdot \boldsymbol{s} > F} \{\boldsymbol{\jmath}(\boldsymbol{x}) \cdot \boldsymbol{s}\}. \tag{2.7}$$

Notice that $\boldsymbol{e}(\boldsymbol{x})$ takes its values in the finite set $\mathcal{C}$ and, if not identically constant, cannot be continuous in the whole Ω, which implies that also $\boldsymbol{V}(\boldsymbol{x})$ cannot be continuous in Ω. More precisely, it may happen that at some point $\boldsymbol{x}$ the maximization problem (2.7) admits two solutions: at such points the field $\boldsymbol{e}(\boldsymbol{x})$, and by consequence also $\boldsymbol{V}(\boldsymbol{x})$, is multivalued. Indeed, $\boldsymbol{\jmath} \cdot \boldsymbol{s}$ can have at most two maxima in $\mathcal{C}$ for $\boldsymbol{\jmath}$ given. Assume in fact that there exist three distinct unit vectors $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, $\boldsymbol{s}_3$ such that $\boldsymbol{\jmath} \cdot \boldsymbol{s}_1 = \boldsymbol{\jmath} \cdot \boldsymbol{s}_2 = \boldsymbol{\jmath} \cdot \boldsymbol{s}_3$; then the endpoints of $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, and $\boldsymbol{s}_3$ belong to the same straight line perpendicular to $\boldsymbol{\jmath}$, which is impossible since the $\boldsymbol{s}_i$ are unit vectors.

A detailed analysis of the phase portrait of the dynamical system (2.5) was performed in [2], where it is shown that Ω splits into (i) regions where $\boldsymbol{V}(\boldsymbol{x}) = \boldsymbol{0}$, and the dislocation is stationary; (ii) *single slip regions* $R(\boldsymbol{s})$ (open regions in $\mathbb{R}^2$), in which $\boldsymbol{e}(\boldsymbol{x}) = \boldsymbol{s}$ is constant; and (iii) curves S on which $\boldsymbol{e}(\boldsymbol{x})$ is multivalued, which can be further subdivided into *cross-slip curves*, *sources*, and *attracting curves* according to the relative orientation of the curve S and the vectors $\boldsymbol{e}(\boldsymbol{x})$ (Figure 2.3).

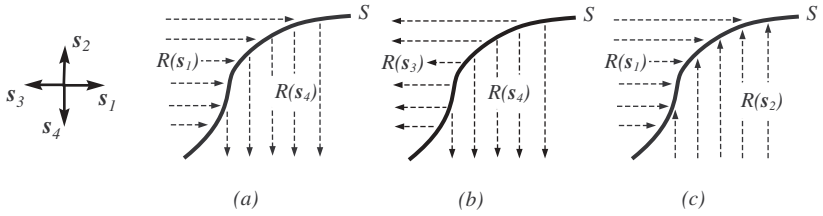


FIG. 2.3. (a) *Cross-slip curve;* (b) *source;* (c) *attracting curve—separating two single-slip regions.*

The motion of a dislocation, solution of (2.5), can be described as follows. Consider, to fix ideas, a dislocation initially at $\boldsymbol{z}_0 \in R(\boldsymbol{s}_1)$. The evolution equation (2.5) reduces to

$$\dot{\boldsymbol{z}} = V_1(\boldsymbol{z})\boldsymbol{s}_1$$

with $V_1(\boldsymbol{z}) = M(\boldsymbol{\jmath}(\boldsymbol{z}) \cdot \boldsymbol{s}_1 - F)$. Hence, the dislocation moves along a straight line parallel to $\boldsymbol{s}_1$, until it reaches some point at the boundary of $R(\boldsymbol{s}_1)$. If this point belongs to a curve on which $\boldsymbol{e}(\boldsymbol{x})$ is multivalued, two situations may occur. If the curve is a cross-slip curve, separating two single slip regions $R(\boldsymbol{s}_1)$ and (say) $R(\boldsymbol{s}_4)$ (corresponding to the directions $\boldsymbol{s}_1$ and $\boldsymbol{s}_4$), then the solution can be prolonged across the curve, and the dislocation moves into $R(\boldsymbol{s}_4)$ on a straight line parallel to $\boldsymbol{s}_4$; this phenomenon is known as *cross slip* (Figure 2.3(a)). If the curve is an attractor, then the solution of (2.5) cannot be prolonged into the adjacent region, since it would violate the maximum dissipation criterion (Figure 2.3(c)).

Hence, the problem seems to be ill-posed in the presence of an attractor. To remove the ambiguity, it was suggested in [2] that, when the dislocation reaches an attractor S, it continues to move along it according to an evolution equation of the form

$$\dot{\boldsymbol{z}} = \boldsymbol{w}(\boldsymbol{z}) \quad \text{with } \boldsymbol{w}(\boldsymbol{z}) = V_{12}(\boldsymbol{z})(\alpha_1(\boldsymbol{z})\boldsymbol{s}_1 + \alpha_2(\boldsymbol{z})\boldsymbol{s}_2), \tag{2.8}$$

where $V_{12}(\boldsymbol{z}) := \boldsymbol{\jmath}(\boldsymbol{z}) \cdot \boldsymbol{s}_1 - F = \boldsymbol{\jmath}(\boldsymbol{z}) \cdot \boldsymbol{s}_2 - F$, and α_1, α_2 are determined by solving

$$\begin{cases} \alpha_1 + \alpha_2 = 1, \\ \alpha_1(\boldsymbol{s}_1 - \boldsymbol{s}_2) \cdot (\nabla \boldsymbol{\jmath})\boldsymbol{s}_1 + \alpha_2(\boldsymbol{s}_1 - \boldsymbol{s}_2) \cdot (\nabla \boldsymbol{\jmath})\boldsymbol{s}_2 = 0. \end{cases} \tag{2.9}$$

The resulting smooth motion of the dislocation, referred to as *fine cross slip*, is therefore noncrystallographic, since it does not occur along a crystallographic direction $\boldsymbol{s} \in \mathcal{C}$. The purpose of the next section is to show that motion by fine cross slip (2.8) can be realized as the limit of a sequence of infinitesimal cross slips across the attracting curve S and to provide a basis for a numerical scheme based on the maximum dissipation criterion.

Remark. Letting

$$\hat{V}(\boldsymbol{e}, \boldsymbol{\jmath}) := \begin{cases} 0 & \text{if } \boldsymbol{\jmath} \cdot \boldsymbol{e} \le F, \\ M(\boldsymbol{\jmath} \cdot \boldsymbol{e} - F) & \text{if } \boldsymbol{\jmath} \cdot \boldsymbol{e} > F, \end{cases} \tag{2.10}$$

we may rewrite condition (2.7) as the requirement that motion may occur only in those directions $\boldsymbol{e}$ which maximize the *dissipation* $\hat{V}(\boldsymbol{s}, \boldsymbol{\jmath})\boldsymbol{\jmath} \cdot \boldsymbol{s}$, i.e.,

$$\hat{V}(\boldsymbol{e}, \boldsymbol{\jmath})\, \boldsymbol{\jmath} \cdot \boldsymbol{e} = \max_{\boldsymbol{s} \in \mathcal{C}} [\hat{V}(\boldsymbol{s}, \boldsymbol{\jmath})\, \boldsymbol{\jmath} \cdot \boldsymbol{s}], \tag{2.11}$$

provided that $\hat{V}(\boldsymbol{e},\boldsymbol{\jmath}) > 0$. The equivalence of (2.7) and (2.11) follows from the fact that the function $M(\xi - F)\xi$ is monotonic with respect to ξ for $\xi > F$.

3. Convergence of sequences of admissible polygonals. We study here the motion of a dislocation near an attracting curve, in order to justify (2.8) rigorously. From now on we regard the vector field $\boldsymbol{\jmath}(\boldsymbol{x})$ in (2.4) as assigned and smooth in Ω.

Let $\boldsymbol{z} : [0,T] \to \Omega$ be a given motion (not necessarily a solution of (2.5), (2.7), and (2.6)). Writing

$$\dot{\boldsymbol{z}}(t) = V(t)\boldsymbol{e}(t), \quad t \in [0,T], \tag{3.1}$$

with $V = |\dot{\boldsymbol{z}}|$ and $\boldsymbol{e} = \dot{\boldsymbol{z}}/|\dot{\boldsymbol{z}}|$, we say that $\boldsymbol{z}$ is *admissible* if

(i) $\boldsymbol{z}$ is continuous and piecewise smooth,

(ii) the direction of motion $\boldsymbol{e}(t)$ belongs to the set of crystallographic directions, and the velocity is a function of the projection of the force on that direction,[3] i.e.,

$$\boldsymbol{e}(t) \in \mathcal{C} \quad \text{and} \quad V(t) = \hat{V}(\boldsymbol{e}(t), \boldsymbol{\jmath}(\boldsymbol{z}(t))), \tag{3.2}$$

at each time t, with $\hat{V}$ given by (2.10).

An admissible motion does not necessarily satisfy the maximum dissipation criterion at all times, but its trajectory is a polygonal with edges parallel to the crystallographic directions.

We assume from now on that the set of crystallographic directions is

$$\mathcal{C} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, -\boldsymbol{s}_1, -\boldsymbol{s}_2\} \tag{3.3}$$

with $\boldsymbol{s}_1 = \boldsymbol{e}_1$ and $\boldsymbol{s}_2 = \boldsymbol{e}_2$, and we consider two adjacent single slip regions $R(\boldsymbol{s}_1)$ and $R(\boldsymbol{s}_2)$, connected open sets in Ω such that[4] $\overline{R(\boldsymbol{s}_1)} \cap \overline{R(\boldsymbol{s}_2)} \neq \emptyset$ and $\overline{R(\boldsymbol{s}_1)} \cap \partial\Omega = \emptyset$, $\overline{R(\boldsymbol{s}_2)} \cap \partial\Omega = \emptyset$. By definition, in $R(\boldsymbol{s}_1)$ and $R(\boldsymbol{s}_2)$ the dissipation is maximal in the directions $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, respectively, i.e.,

$$\begin{cases} \boldsymbol{x} \in R(\boldsymbol{s}_1) \Rightarrow \boldsymbol{s}_1 \cdot \boldsymbol{\jmath}(\boldsymbol{x}) > \boldsymbol{s} \cdot \boldsymbol{\jmath}(\boldsymbol{x}) & \forall \boldsymbol{s} \in \mathcal{C},\ \boldsymbol{s} \neq \boldsymbol{s}_1, \\ \boldsymbol{x} \in R(\boldsymbol{s}_2) \Rightarrow \boldsymbol{s}_2 \cdot \boldsymbol{\jmath}(\boldsymbol{x}) > \boldsymbol{s} \cdot \boldsymbol{\jmath}(\boldsymbol{x}) & \forall \boldsymbol{s} \in \mathcal{C},\ \boldsymbol{s} \neq \boldsymbol{s}_2. \end{cases} \tag{3.4}$$

Also, we assume that

$$\boldsymbol{\jmath}(\boldsymbol{x})\cdot\boldsymbol{s}_1 > F \ \text{ and } \ \boldsymbol{\jmath}(\boldsymbol{x})\cdot\boldsymbol{s}_2 > F, \quad \boldsymbol{x} \in \overline{R(\boldsymbol{s}_1)} \cup \overline{R(\boldsymbol{s}_2)}.$$

3.1. The definition of attracting curve. Let

$$G(\boldsymbol{x}) := (\boldsymbol{s}_2 - \boldsymbol{s}_1) \cdot \boldsymbol{\jmath}(\boldsymbol{x}), \tag{3.5}$$

and assume that $\boldsymbol{\jmath}$ is such that $\nabla G \neq 0$ in Ω. By definition,

$$G(\boldsymbol{x}) < 0 \ \text{ for } \ \boldsymbol{x} \in R(\boldsymbol{s}_1) \quad \text{and} \quad G(\boldsymbol{x}) > 0 \ \text{ for } \ \boldsymbol{x} \in R(\boldsymbol{s}_2),$$

so that, by the smoothness of G and the fact that $\nabla G \neq 0$, the set

$$S = \overline{R(\boldsymbol{s}_1)} \cap \overline{R(\boldsymbol{s}_2)}$$

[3] For $\boldsymbol{z}$ continuous and piecewise smooth, $\dot{\boldsymbol{z}}$ is the right time derivative at corner points.

[4] Here $\bar{R}$ denotes the closure of a set $R \subset \Omega$.

is a smooth curve on which G vanishes, i.e.,

$$G(\boldsymbol{x}) = 0 \Leftrightarrow \boldsymbol{s}_1 \cdot \boldsymbol{\jmath}(\boldsymbol{x}) = \boldsymbol{s}_2 \cdot \boldsymbol{\jmath}(\boldsymbol{x}), \quad \boldsymbol{x} \in S. \tag{3.6}$$

We say that S is an *attracting curve* for $R(\boldsymbol{s}_1)$ and $R(\boldsymbol{s}_2)$ if it satisfies the supplementary conditions

$$\boldsymbol{s}_1 \cdot \nabla G(\boldsymbol{x}) > 0, \quad \boldsymbol{s}_2 \cdot \nabla G(\boldsymbol{x}) < 0, \quad \boldsymbol{x} \in S. \tag{3.7}$$

Hence, at an attracting curve, $\boldsymbol{s}_1$ points into $R(\boldsymbol{s}_2)$ and $\boldsymbol{s}_2$ points into $R(\boldsymbol{s}_1)$ (Figure 2.3(c)). We denote by

$$\boldsymbol{\tau} = \boldsymbol{e}_3 \times \frac{\nabla G}{|\nabla G|}$$

the tangent vector to S.

No admissible motion satisfying the maximum dissipation criterion can originate from an attracting curve S. To see this, consider an admissible motion along $\boldsymbol{s}_1$ with initial point on S: by $(3.7)_1$, G is increasing along $\boldsymbol{s}_1$, and the dislocation moves into the single slip region $R(\boldsymbol{s}_2)$. But in this region the dissipation is maximal in the direction $\boldsymbol{s}_2$, and the maximum dissipation criterion is violated.

Moreover, writing

$$\begin{cases} V_1(\boldsymbol{x}) := \hat{V}(\boldsymbol{s}_1, \boldsymbol{\jmath}(\boldsymbol{x})) = M(\boldsymbol{s}_1 \cdot \boldsymbol{\jmath}(\boldsymbol{x}) - F), \\ V_2(\boldsymbol{x}) := \hat{V}(\boldsymbol{s}_2, \boldsymbol{\jmath}(\boldsymbol{x})) = M(\boldsymbol{s}_2 \cdot \boldsymbol{\jmath}(\boldsymbol{x}) - F) \end{cases} \tag{3.8}$$

for the admissible velocities in the directions $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ at $\boldsymbol{x} \in \overline{R(\boldsymbol{s}_1)} \cup \overline{R(\boldsymbol{s}_2)}$, (3.6) implies that $V_1(\boldsymbol{x}) = V_2(\boldsymbol{x})$ at $\boldsymbol{x} \in S$, and we denote by

$$V(\boldsymbol{x}) := V_1(\boldsymbol{x}) = V_2(\boldsymbol{x}), \quad \boldsymbol{x} \in S,$$

their common value. However, since at S the maximum dissipation criterion admits both $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ as solutions, the vector field $\boldsymbol{V}$ in (2.6) is multivalued, with values

$$V(\boldsymbol{x})\boldsymbol{s}_1 \quad \text{and} \quad V(\boldsymbol{x})\boldsymbol{s}_2$$

at $\boldsymbol{x} \in S$.

3.2. Admissible polygonals. We study here admissible motions that do not necessarily satisfy the maximum dissipation criterion. By definition, an admissible motion $\boldsymbol{z}$ is a time-parametrized polygonal with sides parallel to the crystallographic directions $\boldsymbol{s}_i \in \mathcal{C}$ and piecewise continuous speed given by (2.10). Restricting to admissible motions occurring in $\overline{R(\boldsymbol{s}_1)} \cup \overline{R(\boldsymbol{s}_2)}$ along the directions $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ only, we have

$$\text{either} \quad \dot{\boldsymbol{z}}(t) = V_1(\boldsymbol{z}(t))\boldsymbol{s}_1 \quad \text{or} \quad \dot{\boldsymbol{z}}(t) = V_2(\boldsymbol{z}(t))\boldsymbol{s}_2$$

for $t \in [0, T]$, where V_1 and V_2 are given by (3.8) and $\dot{\boldsymbol{z}}(t)$ is the right time derivative at the corner points of the polygonal.

Hence, an admissible polygonal is a piecewise smooth curve

$$\boldsymbol{z}(t) = x(t)\boldsymbol{s}_1 + y(t)\boldsymbol{s}_2$$

such that there exists a partition $\{[\tau_i, \tau_{i+1}]\}$ of the time interval $[0,T]$ with $i = 0,1,2,\dots$, for which

(3.9)
$$x(t) = \begin{cases} \text{given by } t - \tau_{2j} = \displaystyle\int_{x(\tau_{2j})}^{x(t)} \frac{d\xi}{V_1(\boldsymbol{z}(\tau_{2j}) + (\xi - x(\tau_{2j}))\boldsymbol{s}_1)}, & t \in [\tau_{2j}, \tau_{2j+1}], \\ x(\tau_{2j+1}), & t \in [\tau_{2j+1}, \tau_{2j+2}], \end{cases}$$

and

(3.10)
$$y(t) = \begin{cases} y(\tau_{2j}), & t \in [\tau_{2j}, \tau_{2j+1}], \\ \text{given by} \\ t - \tau_{2j+1} = \displaystyle\int_{y(\tau_{2j+1})}^{y(t)} \frac{d\eta}{V_2(\boldsymbol{z}(\tau_{2j+1}) + (\eta - y(\tau_{2j+1}))\boldsymbol{s}_2)}, & t \in [\tau_{2j+1}, \tau_{2j+2}]. \end{cases}$$

Since V_1 and V_2 are continuous in $\overline{R(\boldsymbol{s}_1)} \cup \overline{R(\boldsymbol{s}_2)}$, they are also bounded on compact subsets, and this in turn implies the following property of admissible motions.

PROPOSITION 3.1. *For any $\boldsymbol{z}_0 \in \overline{R(\boldsymbol{s}_1)} \cup \overline{R(\boldsymbol{s}_2)}$ there exists $T > 0$ such that the set of all admissible motions originating from $\boldsymbol{z}_0$ is bounded in $W^{1,\infty}((0,T),\mathbb{R}^2)$.*

3.3. Sequences of admissible motions. The natural notion of convergence for sequences of admissible motions should account for the fact that the velocity oscillates between the directions $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ and therefore may converge only in average. Weak-$*$ convergence in $W^{1,\infty}((0,T),\mathbb{R}^2)$ serves the purpose. We say that a sequence of Lipschitz motions $\{\boldsymbol{z}_k\}$ converges weak-$*$ in $W^{1,\infty}((0,T),\mathbb{R}^2)$ if there exists a motion $\boldsymbol{\xi} \in W^{1,\infty}((0,T),\mathbb{R}^2)$ such that $\boldsymbol{z}_k \to \boldsymbol{\xi}$ strongly in $C([0,T],\mathbb{R}^2)$ and $\dot{\boldsymbol{z}}_k \overset{*}{\rightharpoonup} \dot{\boldsymbol{\xi}}$ in $L^\infty([0,T],\mathbb{R}^2)$, i.e.,

$$\sup_{t\in[0,T]} |\boldsymbol{z}_k(t) - \boldsymbol{\xi}(t)| \to 0$$

and

$$\int_I (\dot{\boldsymbol{z}}_k(t) - \dot{\boldsymbol{\xi}}(t))\, dt \to 0$$

for any interval $I \subset [0,T]$, provided $\{\dot{\boldsymbol{z}}_k(t)\}$ is bounded in $L^\infty([0,T],\mathbb{R}^2)$.

The weak limit of a sequence of admissible motions is characterized by the Young measure associated to the sequence of the velocities (see Young [13] or, for a more recent approach, [10]). Consider in fact a sequence $\{\boldsymbol{w}_k : (0,T) \to \mathbb{R}^2\}$ converging weak-$*$ to $\boldsymbol{w}_0$ in $L^\infty((0,T),\mathbb{R}^2)$. A Young measure associated with the sequence $\{\boldsymbol{w}_k\}$ is a family of probability measures $\{\nu_t\}_{t\in(0,T)}$ in $\mathbb{R}^2$ which depends measurably on t, i.e., for any $\varphi : \mathbb{R}^2 \to \mathbb{R}$ continuous, the function

$$\bar{\varphi}(t) = \int_{\mathbb{R}^2} \varphi(\boldsymbol{w}) d\nu_t(\boldsymbol{w}) \tag{3.11}$$

is measurable. The fundamental property of ν_t is that, for any continuous φ, the sequence $\{\varphi(\boldsymbol{w}_k)\}$ converges (modulo a subsequence) weak-$*$ to $\bar{\varphi}$ in $L^\infty((0,T),\mathbb{R})$, i.e.,

$$\int_I \varphi(\boldsymbol{w}_k(t))\, dt \to \int_I \int_{\mathbb{R}^2} \varphi(\boldsymbol{w}) d\nu_t(\boldsymbol{w})\, dt, \tag{3.12}$$

for any interval $I \subset [0,T]$, provided that $\{\varphi(\boldsymbol{w}_k)\}$ is bounded in $L^\infty([0,T],\mathbb{R})$. In particular, the weak limit $\boldsymbol{w}_0$ is the expected value of $\boldsymbol{w}$ with respect to ν_t, i.e.,

$$\boldsymbol{w}_0(t) = \int_{\mathbb{R}^2} \boldsymbol{w}\, d\nu_t(\boldsymbol{w}). \tag{3.13}$$

An explicit characterization of ν_t is as follows. For every measurable $E \subset \mathbb{R}^2$,

$$\nu_t(E) = \lim_{R\to 0} \lim_{k\to+\infty} \frac{|\{s \in (t-R, t+R) \,:\, \boldsymbol{w}_k(s) \in E\}|}{2R}, \tag{3.14}$$

where $|\cdot|$ denotes the Lebesgue measure on $\mathbb{R}$.

THEOREM 3.2. *Consider a sequence of admissible polygonals $\boldsymbol{z}_k(t)$ in the directions $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, converging weak-$*$ in $W^{1,\infty}((0,T),\mathbb{R}^2)$ as $k \to +\infty$ to a Lipschitz motion $\boldsymbol{\xi} \in W^{1,\infty}((0,T),\mathbb{R}^2)$. Then the Young measure associated to the sequence $\{\dot{\boldsymbol{z}}_k\}$ is*

$$\nu_t = \lambda_1(t)\, \delta_{V_1(\boldsymbol{\xi}(t))\boldsymbol{s}_1} + \lambda_2(t)\, \delta_{V_2(\boldsymbol{\xi}(t))\boldsymbol{s}_2}, \qquad t \in (0,T), \tag{3.15}$$

with $\delta_{V_1(\boldsymbol{\xi}(t))\boldsymbol{s}_1}$ and $\delta_{V_2(\boldsymbol{\xi}(t))\boldsymbol{s}_2}$ Dirac measures localized at $V_1(\boldsymbol{\xi}(t))\boldsymbol{s}_1$ and $V_2(\boldsymbol{\xi}(t))\boldsymbol{s}_2$, respectively, and

$$\lambda_1(t) = \frac{\dot{\boldsymbol{\xi}}(t)\cdot \boldsymbol{s}_1}{V_1(\boldsymbol{\xi}(t))}, \qquad \lambda_2(t) = \frac{\dot{\boldsymbol{\xi}}(t)\cdot \boldsymbol{s}_2}{V_2(\boldsymbol{\xi}(t))}. \tag{3.16}$$

Proof. We use property (3.14) to compute the measure ν_t associated to $\{\dot{\boldsymbol{z}}_k\}$. Fix $\bar{t} \in (0,T)$, and consider the sets

$$E_{1,\epsilon} = \{\boldsymbol{w} \in \mathbb{R}^2 \,:\, \boldsymbol{w} = w\boldsymbol{s}_1,\ w \in (V_1(\boldsymbol{\xi}(\bar{t})) - \epsilon, V_1(\boldsymbol{\xi}(\bar{t})) + \epsilon)\},$$
$$E_{2,\epsilon} = \{\boldsymbol{w} \in \mathbb{R}^2 \,:\, \boldsymbol{w} = w\boldsymbol{s}_2,\ w \in (V_2(\boldsymbol{\xi}(\bar{t})) - \epsilon, V_2(\boldsymbol{\xi}(\bar{t})) + \epsilon)\}$$

for fixed $\epsilon > 0$, and

$$E_1 = \{V_1(\boldsymbol{\xi}(\bar{t}))\boldsymbol{s}_1\}, \quad E_2 = \{V_2(\boldsymbol{\xi}(\bar{t}))\boldsymbol{s}_2\}.$$

We want to compute

$$\nu_{\bar{t}}(E_{1,\epsilon}) = \lim_{R\to 0} \lim_{k\to+\infty} \frac{|I_{1,k,\epsilon}|}{2R} \quad \text{and} \quad \nu_{\bar{t}}(E_{2,\epsilon}) = \lim_{R\to 0} \lim_{k\to+\infty} \frac{|I_{2,k,\epsilon}|}{2R},$$

where

$$I_{1,k,\epsilon} = \{s \in (\bar{t}-R, \bar{t}+R) \,:\, \dot{\boldsymbol{z}}_k(s) \in E_{1,\epsilon}\},$$
$$I_{2,k,\epsilon} = \{s \in (\bar{t}-R, \bar{t}+R) \,:\, \dot{\boldsymbol{z}}_k(s) \in E_{2,\epsilon}\}.$$

Since every polygonal $\boldsymbol{z}_k$ is admissible, it satisfies (3.2), and

$$\text{either} \quad \dot{\boldsymbol{z}}_k(t) = V_1(\boldsymbol{z}_k(t))\boldsymbol{s}_1 \quad \text{or} \quad \dot{\boldsymbol{z}}_k(t) = V_2(\boldsymbol{z}_k(t))\boldsymbol{s}_2 \tag{3.17}$$

for $t \in [0,T]$, with V_1 and V_2 given by (3.8). Hence, since $V_1(\boldsymbol{z}_k(t)) \to V_1(\boldsymbol{\xi}(t))$ and $V_2(\boldsymbol{z}_k(t)) \to V_2(\boldsymbol{\xi}(t))$ uniformly in t it follows that $\dot{\boldsymbol{z}}_k(s) \in E_{1,\epsilon} \cup E_{2,\epsilon}$ for $s \in (\bar{t}-R, \bar{t}+R)$ for k sufficiently large and R sufficiently small. Hence, by (3.14), if E does not contain either $E_{1,\epsilon}$ or $E_{2,\epsilon}$, then $\nu_{\bar{t}}(E) = 0$, so that the measure $\nu_{\bar{t}}$ is concentrated

on $E_{1,\epsilon}$ and $E_{2,\epsilon}$. Since ϵ is arbitrary, $\nu_{\bar t}(E_{1,\epsilon})$ and $\nu_{\bar t}(E_{2,\epsilon})$ are independent of ϵ, and since (see [11])

$$\nu_{\bar t}(E_1) = \inf_{\epsilon}\{\nu_{\bar t}(E_{1,\epsilon})\}, \qquad \nu_{\bar t}(E_2) = \inf_{\epsilon}\{\nu_{\bar t}(E_{2,\epsilon})\},$$

it follows that the measure $\nu_{\bar t}$ is concentrated on E_1 and E_2, so that (3.15) holds with

$$\lambda_1(\bar t) = \nu_{\bar t}(E_1), \qquad \lambda_2(\bar t) = \nu_{\bar t}(E_2).$$

Moreover, admissibility implies that $|I_{1,k,\epsilon}|+|I_{2,k,\epsilon}| = |I_{1,k,\epsilon}\cup I_{2,k,\epsilon}| = |(\bar t-R, \bar t+R)| = 2R$, and passing to the limit this in turn implies that

$$\lambda_1 + \lambda_2 = 1, \tag{3.18}$$

an identity which is also an immediate consequence of the fact that $\nu_{\bar t}$ is a probability measure. Finally, by (3.13),

$$\dot{\boldsymbol{\xi}}(\bar t\,) = \int_{\mathbb{R}^2} \boldsymbol{w}\, d\nu_{\bar t}(\boldsymbol{w}) = \lambda_1(\bar t\,)\, V_1(\boldsymbol{\xi}(\bar t\,))\boldsymbol{s}_1 + \lambda_2(\bar t\,)\, V_2(\boldsymbol{\xi}(\bar t\,))\boldsymbol{s}_2,$$

from which (3.16) follows. □

Notice that since the velocity of the limit motion is

$$\dot{\boldsymbol{\xi}}(t) = \lambda_1(t)V_1(\boldsymbol{\xi}(t))\boldsymbol{s}_1 + \lambda_2(t)V_2(\boldsymbol{\xi}(t))\boldsymbol{s}_2, \tag{3.19}$$

it follows that the weak limit of a sequence of admissible motions is not necessarily admissible but can be represented as a fine mixture of crystallographic motions in the admissible directions $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$.

COROLLARY 3.3. *Let S be an attracting curve separating two single slip regions $R(\boldsymbol{s}_1)$ and $R(\boldsymbol{s}_2)$: any sequence of admissible polygonals $\boldsymbol{z}_k(t)$ with directions $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ such that*

$$\operatorname{dist}(\boldsymbol{z}_k(t), S) \to 0, \tag{3.20}$$

uniformly in $t \in [0,T]$ as $k \to +\infty$, converges weak-$*$ in $W^{1,\infty}((0,T),\mathbb{R}^2)$ (and, in particular, uniformly) to a smooth motion $\boldsymbol{\xi}(t)$ on S with velocity

$$\dot{\boldsymbol{\xi}}(t) = \frac{V(\boldsymbol{\xi}(t))}{\boldsymbol{\tau}(\boldsymbol{\xi}(t))\cdot(\boldsymbol{s}_1+\boldsymbol{s}_2)}\ \boldsymbol{\tau}(\boldsymbol{\xi}(t)), \tag{3.21}$$

where $\boldsymbol{\tau}$ is the unit tangent vector to S and $V(\boldsymbol{x}) := V_1(\boldsymbol{x}) = V_2(\boldsymbol{x})$ the speed evaluated at $\boldsymbol{x} \in S$ (see (3.8)). Moreover, the Young measure associated to the sequence $\{\dot{\boldsymbol{z}}_k\}$ is

$$\nu_t = \lambda_1(\boldsymbol{\xi}(t))\,\delta_{V(\boldsymbol{\xi}(t))\boldsymbol{s}_1} + \lambda_2(\boldsymbol{\xi}(t))\,\delta_{V(\boldsymbol{\xi}(t))\boldsymbol{s}_2} \tag{3.22}$$

with

$$\lambda_1(\boldsymbol{x}) = \frac{\boldsymbol{\tau}(\boldsymbol{x})\cdot\boldsymbol{s}_1}{\boldsymbol{\tau}(\boldsymbol{x})\cdot(\boldsymbol{s}_1+\boldsymbol{s}_2)}, \qquad \lambda_2(\boldsymbol{x}) = \frac{\boldsymbol{\tau}(\boldsymbol{x})\cdot\boldsymbol{s}_2}{\boldsymbol{\tau}(\boldsymbol{x})\cdot(\boldsymbol{s}_1+\boldsymbol{s}_2)} \tag{3.23}$$

for a.e. $\boldsymbol{x} \in S$.

Proof. Since every polygonal $\boldsymbol{z}_k$ is admissible, it satisfies (3.2) and (2.10), i.e.,

$$\text{either} \quad \dot{\boldsymbol{z}}_k(t) = V_1(\boldsymbol{z}_k(t))\boldsymbol{s}_1 \quad \text{or} \quad \dot{\boldsymbol{z}}_k(t) = V_2(\boldsymbol{z}_k(t))\boldsymbol{s}_2,$$

for $t \in [0,T]$. Hence, since V_1 and V_2 are bounded in a neighborhood of S, the sequence $\{\boldsymbol{z}_k\}$ is bounded in $W^{1,\infty}((0,T),\mathbb{R}^2)$ (see Proposition 3.1), so that there exists a subsequence (not relabeled) extracted from $\{\boldsymbol{z}_k\}$, converging weak-$*$ to a parametrized curve $\boldsymbol{\xi} \in W^{1,\infty}((0,T),\mathbb{R}^2)$. By (3.20), $\boldsymbol{\xi}(t) \in S$ for all $t \in T$. Writing

$$\dot{\boldsymbol{\xi}} = W\boldsymbol{\ell}, \tag{3.24}$$

with $\boldsymbol{\ell}(t) = \dot{\boldsymbol{\xi}}(t)/|\dot{\boldsymbol{\xi}}(t)|$ the unit vector of $\dot{\boldsymbol{\xi}}(t)$, and applying (3.16) and (3.18) we find

$$\lambda_1 + \lambda_2 = \frac{\dot{\boldsymbol{\xi}} \cdot \boldsymbol{s}_1}{V_1} + \frac{\dot{\boldsymbol{\xi}} \cdot \boldsymbol{s}_2}{V_2} = 1,$$

from which it follows that

$$W = \frac{V_1(\boldsymbol{\xi})V_2(\boldsymbol{\xi})}{\boldsymbol{\ell} \cdot (V_1(\boldsymbol{\xi})\boldsymbol{s}_2 + V_2(\boldsymbol{\xi})\boldsymbol{s}_1)}. \tag{3.25}$$

Also, writing

$$\lambda_1 = \frac{\dot{\boldsymbol{\xi}} \cdot \boldsymbol{s}_1/V_1}{\dot{\boldsymbol{\xi}} \cdot (\boldsymbol{s}_1/V_1 + \boldsymbol{s}_2/V_2)} = \frac{V_2\dot{\boldsymbol{\xi}} \cdot \boldsymbol{s}_1}{\dot{\boldsymbol{\xi}} \cdot (V_2\boldsymbol{s}_1 + V_1\boldsymbol{s}_2)},$$

and recalling (3.24), we also obtain the coefficients of the Young measure associated to $\{\dot{\boldsymbol{z}}_k\}$ in the form

$$\lambda_1 = \frac{V_2(\boldsymbol{\xi})\boldsymbol{\ell} \cdot \boldsymbol{s}_1}{\boldsymbol{\ell} \cdot (V_2(\boldsymbol{\xi})\boldsymbol{s}_1 + V_1(\boldsymbol{\xi})\boldsymbol{s}_2)}, \qquad \lambda_2 = \frac{V_1(\boldsymbol{\xi})\boldsymbol{\ell} \cdot \boldsymbol{s}_2}{\boldsymbol{\ell} \cdot (V_2(\boldsymbol{\xi})\boldsymbol{s}_1 + V_1(\boldsymbol{\xi})\boldsymbol{s}_2)}. \tag{3.26}$$

Now, the function $\boldsymbol{\xi}$ defines a motion on S; hence $\boldsymbol{\ell} = \boldsymbol{\tau}$ (the unit tangent to S) in (3.25) and (3.26), and recalling that $V_1 = V_2 =: V$ on S and S is smooth, we obtain (3.21), (3.22) and (3.23).

Finally, notice that (3.21) uniquely defines a smooth motion on S, and its expression does not depend on the subsequence of $\{\boldsymbol{z}_k\}$ used to construct it. Hence, we can conclude that every converging subsequence has the same limit, so that the original sequence $\{\boldsymbol{z}_k\}$ converges to $\boldsymbol{\xi}$. □

Notice that although each admissible motion $\boldsymbol{z}_k(t)$ does not necessarily satisfy the maximum dissipation criterion for all $t \in [0,T]$, the sequence $\boldsymbol{z}_k$ is a maximizing sequence for the dissipation, since the limit motion $\boldsymbol{\xi}$ satisfies the maximum dissipation criterion (recall, however, that the limit motion is not admissible). To see this, let $J(\boldsymbol{x}) := \boldsymbol{\jmath}(\boldsymbol{x}) \cdot \boldsymbol{s}_1 = \boldsymbol{\jmath}(\boldsymbol{x}) \cdot \boldsymbol{s}_2$ and $V(\boldsymbol{x}) := V_1(\boldsymbol{x}) = V_2(\boldsymbol{x})$ for $\boldsymbol{x} \in S$ (see (3.6)). The maximum dissipation (among all admissible motions) at $\boldsymbol{x} \in S$ is (see (2.11) and (3.4))

$$\max_{\boldsymbol{s} \in \mathcal{C}} \{\hat{V}(\boldsymbol{s}, \boldsymbol{\jmath}(\boldsymbol{x}))\, \boldsymbol{\jmath}(\boldsymbol{x}) \cdot \boldsymbol{s}\} = J(\boldsymbol{x})V(\boldsymbol{x}), \tag{3.27}$$

while the dissipation relative to the limit motion $\boldsymbol{\xi}(t)$ is

$$\boldsymbol{\jmath}(\boldsymbol{\xi}(t)) \cdot \dot{\boldsymbol{\xi}}(t) = \frac{V(\boldsymbol{\xi}(t))}{\boldsymbol{\tau}(\boldsymbol{\xi}(t)) \cdot (\boldsymbol{s}_1 + \boldsymbol{s}_2)}\, \boldsymbol{\jmath}(\boldsymbol{\xi}(t)) \cdot \boldsymbol{\tau}(\boldsymbol{\xi}(t)) = V(\boldsymbol{\xi}(t))J(\boldsymbol{\xi}(t)), \tag{3.28}$$

since $\boldsymbol{\jmath} = J(\boldsymbol{s}_1 + \boldsymbol{s}_2)$, and these expressions coincide at $\boldsymbol{x} = \boldsymbol{\xi}(t)$.

Also, it is not difficult to prove that (3.21) coincides with (2.8). In fact, solving system $(2.8)_2$ and recalling (3.5), we obtain

$$\begin{cases} \alpha_1 = \dfrac{(\boldsymbol{s}_2 - \boldsymbol{s}_1)\cdot(\nabla \boldsymbol{\jmath})\boldsymbol{s}_2}{(\boldsymbol{s}_2 - \boldsymbol{s}_1)\cdot(\nabla \boldsymbol{\jmath})\boldsymbol{s}_2 - (\boldsymbol{s}_2 - \boldsymbol{s}_1)\cdot(\nabla \boldsymbol{\jmath})\boldsymbol{s}_1} = \dfrac{\nabla G\cdot \boldsymbol{s}_2}{\nabla G\cdot \boldsymbol{s}_2 - \nabla G\cdot \boldsymbol{s}_1}, \\ \alpha_2 = \dfrac{-(\boldsymbol{s}_2 - \boldsymbol{s}_1)\cdot(\nabla \boldsymbol{\jmath})\boldsymbol{s}_1}{(\boldsymbol{s}_2 - \boldsymbol{s}_1)\cdot(\nabla \boldsymbol{\jmath})\boldsymbol{s}_2 - (\boldsymbol{s}_2 - \boldsymbol{s}_1)\cdot(\nabla \boldsymbol{\jmath})\boldsymbol{s}_1} = -\dfrac{\nabla G\cdot \boldsymbol{s}_1}{\nabla G\cdot \boldsymbol{s}_2 - \nabla G\cdot \boldsymbol{s}_1}, \end{cases}$$

with G given by (3.5). Now, noting that $\nabla G\cdot \boldsymbol{s}_2 = \nabla G\cdot \boldsymbol{e}_3\times \boldsymbol{s}_1 = -\boldsymbol{e}_3\times \nabla G\cdot \boldsymbol{s}_1 = -|\nabla G|\boldsymbol{\tau}\cdot \boldsymbol{s}_1$, and $\nabla G\cdot \boldsymbol{s}_1 = -\nabla G\cdot \boldsymbol{e}_3\times \boldsymbol{s}_2 = \boldsymbol{e}_3\times \nabla G\cdot \boldsymbol{s}_2 = |\nabla G|\boldsymbol{\tau}\cdot \boldsymbol{s}_2$, we find

$$\alpha_1 = \frac{\boldsymbol{\tau}\cdot \boldsymbol{s}_1}{\boldsymbol{\tau}\cdot \boldsymbol{s}_1 + \boldsymbol{\tau}\cdot \boldsymbol{s}_2}, \qquad \alpha_2 = \frac{\boldsymbol{\tau}\cdot \boldsymbol{s}_2}{\boldsymbol{\tau}\cdot \boldsymbol{s}_1 + \boldsymbol{\tau}\cdot \boldsymbol{s}_2},$$

which yields (3.21), recalling that V_{12} coincides with V in our present notation.

3.4. Sequences of admissible polygonals maximizing the dissipation. In this section we show that every sequence of polygonals maximizing the dissipation converges to the smooth motion $\boldsymbol{\xi}$ on S given by (3.21).

For $\boldsymbol{x}\in\Omega$, let $V_M(\boldsymbol{x})$ and $\boldsymbol{e}_M(\boldsymbol{x})$ denote the speed and direction of motion selected by the maximum dissipation criterion (2.11) among all admissible velocity fields, i.e., such that

$$V_M(\boldsymbol{x})\boldsymbol{e}_M(\boldsymbol{x})\cdot \boldsymbol{\jmath}(\boldsymbol{x}) = \max_{\boldsymbol{s}\in\mathcal{C}}\{\hat{V}(\boldsymbol{s},\boldsymbol{\jmath}(\boldsymbol{x}))\,\boldsymbol{s}\cdot \boldsymbol{\jmath}(\boldsymbol{x})\}, \tag{3.29}$$

where $\hat{V}$ is given by (2.10). Notice that although $\boldsymbol{e}_M(\boldsymbol{x})$ is in general multivalued at S, the maximum dissipation (3.29) is single valued everywhere. Consider the function $D\colon \Omega\times\mathbb{R}^2\to\mathbb{R}$ defined by

$$D(\boldsymbol{x},\boldsymbol{w}) = \boldsymbol{\jmath}(\boldsymbol{x})\cdot(V_M(\boldsymbol{x})\boldsymbol{e}_M(\boldsymbol{x}) - \boldsymbol{w})\,. \tag{3.30}$$

For a given motion $\boldsymbol{z}\in W^{1,\infty}((0,T),\mathbb{R}^2)$ the real function $D(\boldsymbol{z}(t),\dot{\boldsymbol{z}}(t))$ belongs to $L^\infty((0,T),\mathbb{R})$ and measures the difference between the maximum possible dissipation and the actual dissipation at each time.

Fix $\boldsymbol{z}_0\in S$ and consider the set of admissible curves originating from $\boldsymbol{z}_0$:

$$\begin{aligned}\mathcal{A} = \big\{\boldsymbol{z}:[0,T]\to\mathbb{R}^2 :\ & \boldsymbol{z} \text{ piecewise smooth, } \boldsymbol{z}(0) = \boldsymbol{z}_0\in S \quad \text{and} \\ & \text{either} \quad \dot{\boldsymbol{z}}(t) = V_1(\boldsymbol{z}(t))\boldsymbol{s}_1 \quad \text{or} \quad \dot{\boldsymbol{z}}(t) = V_2(\boldsymbol{z}(t))\boldsymbol{s}_2,\ t\in[0,T]\big\}\,,\end{aligned}$$

where $\dot{\boldsymbol{z}}$ denotes the right time derivative at corner points of the polygonals.

By definition

$$D(\boldsymbol{z}(t),\dot{\boldsymbol{z}}(t))\geq 0 \qquad \forall \boldsymbol{z}\in\mathcal{A},\forall t\in[0,T], \tag{3.31}$$

although D can be negative for some nonadmissible motion.

Consider now the functional associated to D,

$$E(\boldsymbol{z}) = \int_0^T D(\boldsymbol{z}(t),\dot{\boldsymbol{z}}(t))\,dt = \int_0^T \boldsymbol{\jmath}(\boldsymbol{z}(t))\cdot(V_M(\boldsymbol{z}(t))\boldsymbol{e}_M(\boldsymbol{z}(t)) - \dot{\boldsymbol{z}}(t))\,dt, \tag{3.32}$$

defined for $\boldsymbol{z} \in W^{1,\infty}((0,T),\mathbb{R}^2)$. By the discussion following (3.7), no admissible motion satisfying the maximum dissipation criterion can originate from S, so that E is strictly positive on $\mathcal{A}$. Indeed, as we shall show in the next section,

$$\inf_{\boldsymbol{z}\in\mathcal{A}} E(\boldsymbol{z}) = 0, \tag{3.33}$$

and the infimum is not attained on $\mathcal{A}$.

THEOREM 3.4. *Any sequence of admissible polygonals $\{\boldsymbol{z}_k\} \subset \mathcal{A}$ minimizing E (or, equivalently, maximizing the dissipation), i.e., such that*

$$\lim_{k\to+\infty} E(\boldsymbol{z}_k) = 0, \tag{3.34}$$

converges weak-$*$ in $W^{1,\infty}((0,T),\mathbb{R}^2)$ to the smooth motion $\boldsymbol{\xi}(t)$ on S, whose velocity is (3.21).

Proof. By Corollary 3.3, it is enough to prove that every sequence of admissible curves $\{\boldsymbol{z}_k\}$ minimizing E converges to a motion on S.

Notice first that, by Proposition 3.1, $\mathcal{A}$ is bounded in $W^{1,\infty}((0,T),\mathbb{R}^2)$, and every sequence $\{\boldsymbol{z}_k\} \subset \mathcal{A}$ admits a subsequence (not relabeled) which converges weak-$*$ in $W^{1,\infty}((0,T),\mathbb{R}^2)$ to a (in general nonadmissible) Lipschitz motion $\boldsymbol{\xi}$. In particular,

$$\lim_{k\to+\infty} \int_I \dot{\boldsymbol{z}}_k(t)\,dt = \int_I \dot{\boldsymbol{\xi}}(t)\,dt$$

for every interval $I \subset [0,T]$, and by Theorem 3.2 (see (3.19)),

$$\dot{\boldsymbol{\xi}}(t) = \lambda_1(t)\,V_1(\boldsymbol{\xi}(t))\boldsymbol{s}_1 + \lambda_2(t)\,V_2(\boldsymbol{\xi}(t))\boldsymbol{s}_2. \tag{3.35}$$

Assume now that $\{\boldsymbol{z}_k\}$ is a minimizing sequence for E. Since all $\boldsymbol{z}_k$ are admissible, the integrand $D(\boldsymbol{z}_k(t),\dot{\boldsymbol{z}}_k(t))$ is nonnegative for all t, so that (3.34) is equivalent to weak-$*$ convergence of $D(\boldsymbol{z}_k(t),\dot{\boldsymbol{z}}_k(t))$ to zero, i.e.,

$$\lim_{k\to+\infty} \int_I D(\boldsymbol{z}_k(t),\dot{\boldsymbol{z}}_k(t))\,dt = 0,$$

for every interval $I \subset [0,T]$. Moreover, since $D(\boldsymbol{x},\boldsymbol{w})$ is continuous with respect to $\boldsymbol{x}$ and affine in $\boldsymbol{w}$, it is continuous under weak-$*$ convergence, and

$$\lim_{k\to+\infty} \int_I D(\boldsymbol{z}_k(t),\dot{\boldsymbol{z}}_k(t))\,dt = \int_I D(\boldsymbol{\xi}(t),\dot{\boldsymbol{\xi}}(t))\,dt.$$

Hence,

$$\int_I D(\boldsymbol{\xi}(t),\dot{\boldsymbol{\xi}}(t))\,dt = 0 \tag{3.36}$$

for any interval $I \subset [0,T]$.

We now show by contradiction that $\boldsymbol{\xi}(t) \in S$ for all $t \in [0,T]$. Suppose, to fix ideas, that $\boldsymbol{\xi}(t) \in R(\boldsymbol{s}_1)$ for $t \in (0,\epsilon)$, and recall that $\boldsymbol{\xi}(0) \in S$. Since by definition $V_M(\boldsymbol{x})\boldsymbol{\jmath}(\boldsymbol{x})\cdot\boldsymbol{e}_M(\boldsymbol{x}) = V_1(\boldsymbol{x})\boldsymbol{s}_1\cdot\boldsymbol{\jmath}(\boldsymbol{x})$ in $R(\boldsymbol{s}_1)$, (3.36) yields

$$\int_0^\epsilon \boldsymbol{\jmath}(\boldsymbol{\xi}(t))\cdot(V_1(\boldsymbol{\xi}(t))\boldsymbol{s}_1 - \dot{\boldsymbol{\xi}}(t))\,dt = 0,$$

which becomes in turn, by (3.35),

$$0 = \int_0^\epsilon \boldsymbol{\jmath}(\boldsymbol{\xi}(t)) \cdot ((1-\lambda_1(t))V_1(\boldsymbol{\xi}(t))\boldsymbol{s}_1 - \lambda_2(t)\, V_2(\boldsymbol{\xi}(t))\boldsymbol{s}_2)\; dt$$
$$= \int_0^\epsilon (1-\lambda_1(t))\,(V_1(\boldsymbol{\xi}(t))\boldsymbol{s}_1 \cdot \boldsymbol{\jmath}(\boldsymbol{\xi}(t)) - V_2(\boldsymbol{\xi}(t))\boldsymbol{s}_2 \cdot \boldsymbol{\jmath}(\boldsymbol{\xi}(t)))\; dt,$$

and since $V_1(\boldsymbol{\xi}(t))\boldsymbol{s}_1 \cdot \boldsymbol{\jmath}(\boldsymbol{\xi}(t)) > V_2(\boldsymbol{\xi}(t))\boldsymbol{s}_2 \cdot \boldsymbol{\jmath}(\boldsymbol{\xi}(t))$ in $R(\boldsymbol{s}_1)$, the integrand is nonnegative, and we obtain

$$\lambda_1(t) = 1, \quad \lambda_2(t) = 0,$$

a.e. in $(0,\epsilon)$. Hence $\boldsymbol{\xi}$ is admissible and

$$\dot{\boldsymbol{\xi}}(t) = V_1(\boldsymbol{\xi}(t))\boldsymbol{s}_1 \quad \text{and} \quad \boldsymbol{\xi}(0) \in S,$$

which is a contradiction since, by $(3.7)_2$, any admissible motion in the direction $\boldsymbol{s}_1$ originating from S necessarily occurs in the region $R(\boldsymbol{s}_2)$ for some time interval containing $t = 0$. Repeating this argument for any interval $[\epsilon, \epsilon']$, it follows that $\boldsymbol{\xi}(t) \in S$ for each $t \in [0,T]$, so that, in particular, $\operatorname{dist}(\boldsymbol{z}_k(t), S) \to 0$ uniformly in t, and we can apply Corollary 3.3 to obtain the thesis. □

4. Explicit construction of a sequence of polygonals maximizing the dissipation. The results of the previous section provide a theoretical justification of the evolution equation (2.8) postulated in [2] for a dislocation moving on an attracting curve, but they can also be used to motivate a simple numerical scheme which involves only a time-discretized ODE solver of the dynamical system (2.5).

The idea is as follows. Assume that the field $\boldsymbol{\jmath}$, and by consequence the vector field $\boldsymbol{V}(\boldsymbol{x})$ in (2.6), is known. Since, away from the attracting curve, the motion of the dislocation is a solution of the dynamical system (2.5), it can be computed by any time-discretized numerical method for ODEs, for instance a simple Euler method. To be specific, choose an initial point in $R(\boldsymbol{s}_1)$, and propagate the dislocation along $\boldsymbol{s}_1$ by solving numerically (2.5) with time step h. Since S is an attracting curve, the solution $\boldsymbol{z}(t)$ approaches S as time increases. At some time t_0, it happens that $\boldsymbol{z}(t_0) \in R(\boldsymbol{s}_1)$, but $\boldsymbol{z}(t_0+h)$, which is obtained by propagating $\boldsymbol{z}(t_0)$ using $\boldsymbol{V}(\boldsymbol{z}(t_0)) = V_1(\boldsymbol{z}(t_0))\boldsymbol{s}_1$, belongs to $R(\boldsymbol{s}_2)$. In other terms, since the time steps are discretized, the dislocation crosses S and moves for a time interval strictly smaller than h along $\boldsymbol{s}_1$ into the region $R(\boldsymbol{s}_2)$, thereby violating the maximum dissipation criterion.

At the next iterative step, the point $\boldsymbol{z}(t_0+h) \in R(\boldsymbol{s}_2)$ is propagated in the $\boldsymbol{s}_2$ direction using the "correct" value $\boldsymbol{V}(\boldsymbol{z}(t_0+h)) = V_2(\boldsymbol{z}(t_0+h))\boldsymbol{s}_2$ and again approaches the attractor S. Iterating the procedure, we obtain a zig-zag approximation of the motion near the attracting curve S, which only requires the knowledge of the vector field (2.6) away from S.

This motion is a numerical approximation of an admissible polygonal, which however does not satisfy the maximum dissipation criterion at all times. The purpose of this section is to construct a sequence of admissible polygonals consistent with the above argument and to show that it maximizes the dissipation as the time step $h \to 0$, which implies in turn that, by Theorem 3.4, it converges as $h \to 0$ to the smooth non-admissible motion on S given by (3.21).

We now make this idea precise. Let S be an attracting curve separating $R(\boldsymbol{s}_1)$ and $R(\boldsymbol{s}_2)$ as in section 3.1. Fix $h = T/k$ with k integer, and choose $\boldsymbol{z}_0 = x_0\boldsymbol{s}_1 + y_0\boldsymbol{s}_2 \in S$.

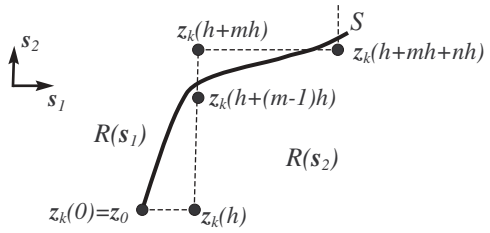


FIG. 4.1. *Approximating polygonal.*

For $t \in [0, h)$, let $\boldsymbol{z}_k(t)$ be the rectilinear motion in the direction $\boldsymbol{s}_1$ solution of

$$\dot{\boldsymbol{z}} = V_1(\boldsymbol{z})\boldsymbol{s}_1$$

with initial condition $\boldsymbol{z}_0$, i.e., writing $\boldsymbol{z}_k(t) = x_k(t)\boldsymbol{s}_1 + y_k(t)\boldsymbol{s}_2$,

$$\begin{cases} x_k(t) \text{ is defined by} \quad t = \displaystyle\int_{x_0}^{x_k(t)} \frac{d\xi}{V_1(\boldsymbol{z}_0 + (\xi - x_0)\boldsymbol{s}_1)}, \qquad t \in [0, h). \\ y_k(t) \equiv y_0, \end{cases}$$

Since, by (3.7), $\boldsymbol{s}_1$ at S points into the single slip region $R(\boldsymbol{s}_2)$, it follows that $\boldsymbol{z}_k(t) \in R(\boldsymbol{s}_2)$ for $t \in [0, h)$. Should the motion continue along $\boldsymbol{s}_1$, it would further violate the maximum dissipation criterion, and therefore at time $t = h$ the motion switches to the preferred direction $\boldsymbol{s}_2$. Consider now the solution of

$$\dot{\boldsymbol{z}} = V_2(\boldsymbol{z})\boldsymbol{s}_2$$

for $t \geq h$, with initial condition $\boldsymbol{z}_k(h)$, and denote this solution by $\bar{\boldsymbol{z}}(t)$. Since S is an attractor for $R(\boldsymbol{s}_2)$, there exists a time $\bar{t}$ such that $\bar{\boldsymbol{z}}(\bar{t}) \in S$, so that there exists an integer m such that $\bar{\boldsymbol{z}}(h + (m-1)h) \in R(\boldsymbol{s}_2)$ but $\bar{\boldsymbol{z}}(h + mh) \in R(\boldsymbol{s}_1)$. We let $\boldsymbol{z}_k(t) = \bar{\boldsymbol{z}}(t)$ for $t \in [h, h + mh]$, i.e.,

$$\begin{cases} x_k(t) \equiv x_k(h), \\ y_k(t) \text{ defined by} \quad t - h = \displaystyle\int_{y_k(h)}^{y_k(t)} \frac{d\eta}{V_2(\boldsymbol{z}_k(h) + (\eta - y_k(h))\boldsymbol{s}_2)}, \qquad t \in [h, h + mh). \end{cases}$$

Again, since $\boldsymbol{z}_k(h + mh) \in R(\boldsymbol{s}_1)$, to satisfy the maximum dissipation criterion the motion switches to the preferred direction $\boldsymbol{s}_1$ at time $t = h + mh$ and moves toward the attractor S (Figure 4.1).

The above procedure can now be iterated, and we obtain an admissible polygonal

$$\boldsymbol{z}_k(t) = x_k(t)\boldsymbol{s}_1 + y_k(t)\boldsymbol{s}_2, \tag{4.1}$$

defined on all $[0, T]$, where $x_k(t)$ and $y_k(t)$ are defined by relations analogous to (3.9) and (3.10), with the τ_i now integer multiples of h.

PROPOSITION 4.1. *The sequence of polygonals* $\{\boldsymbol{z}_k\}$ *defined by* (4.1) *converges to* S*, i.e.,*

$$\operatorname{dist}(\boldsymbol{z}_k(t), S) \to 0, \tag{4.2}$$

uniformly in $t \in [0, T]$*, as* $k \to +\infty$.

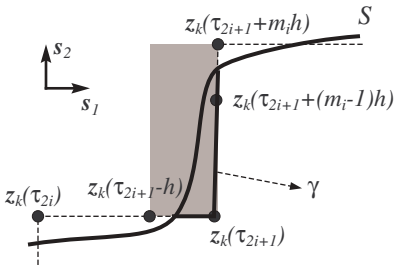


FIG. 4.2. *Proof of Proposition* 4.1.

Proof. We know that $V_1(\boldsymbol{x})$ and $V_2(\boldsymbol{x})$ are bounded in a neighborhood $\mathcal{U}$ of S, and we denote by $W = \max(\max_{\mathcal{U}} V_1, \max_{\mathcal{U}} V_2)$ the maximum speed in $\mathcal{U}$. Also, denote by $U = \min(\min_{\mathcal{U}} V_1, \min_{\mathcal{U}} V_2)$ the minimum speed in $\mathcal{U}$, and assume that $U > 0$.

Consider a typical portion γ of the polygonal $\boldsymbol{z}_k(t)$ lying in $\mathcal{U}$ between two successive intersections with S (Figure 4.2). Assume, to fix ideas, that $\gamma \in R(\boldsymbol{s}_2)$, so that the first intersection with S occurs in the portion of γ parallel to $\boldsymbol{s}_1$, with the second intersection occurring in the portion of γ parallel to $\boldsymbol{s}_2$. Let $[\tau_{2i}, \tau_{2i+1}]$ and $[\tau_{2i+1}, \tau_{2i+2}]$ be the time intervals corresponding to the rectilinear segments of $\boldsymbol{z}_k$ parallel to $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, respectively, and such that $\tau_{2i+1} - \tau_{2i} = n_i h$ and $\tau_{2i+2} - \tau_{2i-1} = m_i h$, with n_i and m_i integers (see (3.9) and (3.10)).

Notice now that (3.7) is equivalent to

$$\boldsymbol{s}_1 \cdot \boldsymbol{\tau}(\boldsymbol{x}) > 0, \qquad \boldsymbol{s}_2 \cdot \boldsymbol{\tau}(\boldsymbol{x}) > 0, \qquad \boldsymbol{x} \in S,$$

with $\boldsymbol{\tau}$ the tangent vector to S, and this in turn implies that, since S is bounded, there exists $\alpha > 0$ such that $\alpha < \boldsymbol{s}_1 \cdot \boldsymbol{\tau}(\boldsymbol{x})$ and $\boldsymbol{s}_2 \cdot \boldsymbol{\tau}(\boldsymbol{x}) < 1 - \alpha$.

Hence, the portion of S between two successive intersections with the polygonal lies in the shaded rectangle in Figure 4.2, and

$$\operatorname{dist}(\boldsymbol{x}, S) \le |\boldsymbol{z}_k(\tau_{2i+1} + m_i h) - \boldsymbol{z}_k(\tau_{2i+1} - h)|, \qquad \boldsymbol{x} \in \gamma.$$

But

$$\begin{aligned} |\boldsymbol{z}_k(\tau_{2i+1} + m_i h) - \boldsymbol{z}_k(\tau_{2i+1} - h)| &\le |\boldsymbol{z}_k(\tau_{2i+1}) - \boldsymbol{z}_k(\tau_{2i+1} - h)| \\ &+ |\boldsymbol{z}_k(\tau_{2i+1} + m_i h) - \boldsymbol{z}_k(\tau_{2i+1})| \le Wh + Wm_i h = Wh(1 + m_i). \end{aligned}$$

To prove the thesis it is enough to show that all m_i are bounded as $k \to +\infty$ (or equivalently $h \to 0$). But this is true since (see Figure 4.2)

$$\frac{U(m_i - 1)h}{Wh} \le \frac{y_k(\tau_{2i+1} + (m_i - 1)h) - y_k(\tau_{2i+1})}{x_k(\tau_{2i+1}) - x_k(\tau_{2i+1} - h)} \le \max_{\boldsymbol{x} \in S} \frac{\boldsymbol{s}_2 \cdot \boldsymbol{\tau}(\boldsymbol{x})}{\boldsymbol{s}_1 \cdot \boldsymbol{\tau}(\boldsymbol{x})} \le \frac{1 - \alpha}{\alpha},$$

an analogous assertion holding for n_i. This proves uniform convergence. ∎

We can now apply Corollary 3.3 to conclude that the sequence $\{\boldsymbol{z}_k(t)\}$ converges uniformly to the smooth motion $\boldsymbol{\xi}$ on S, and the discussion in the paragraph containing (3.27) and (3.28) can be used to show that $\{\boldsymbol{z}_k(t)\}$ maximizes the dissipation.

The following proposition is an independent proof of the maximizing property.

PROPOSITION 4.2. *The sequence of polygonals* $\{\boldsymbol{z}_k\}$ *defined by* (4.1) *maximizes the dissipation, i.e.,*

$$\lim_{k \to +\infty} E(\boldsymbol{z}_k) = 0. \tag{4.3}$$

Proof. Denote by

$$D_k(t) := D(\boldsymbol{z}_k(t), \dot{\boldsymbol{z}}_k(t))$$

the integrand of E. We shall prove that $D_k(t) \to 0$ uniformly in t (in particular, strongly in $L^\infty((0,T),\mathbb{R})$). Indeed, for each k, $D_k(t) = 0$ except when the maximum dissipation criterion is not satisfied, i.e., for those time intervals A_q, $q = 1, \dots, Q$, for which motion occurs in the $\boldsymbol{s}_1$ direction but $\boldsymbol{z}_k(t) \in R(\boldsymbol{s}_2)$, and for those time intervals B_q, $q = 1, \dots, Q$, for which motion occurs in the $\boldsymbol{s}_2$ direction but $\boldsymbol{z}_k(t) \in R(\boldsymbol{s}_1)$. Notice that $|A_q|, |B_q| < h$ by construction. Hence, we can write

$$D_k(t) = \begin{cases} 0, & t \in [0,T] \setminus (\cup_q (A_q \cup B_q)), \\ \boldsymbol{\jmath}(\boldsymbol{z}_k(t)) \cdot (V_2(\boldsymbol{z}_k(t))\boldsymbol{s}_2 - V_1(\boldsymbol{z}_k(t))\boldsymbol{s}_1), & t \in \cup_q A_q, \\ \boldsymbol{\jmath}(\boldsymbol{z}_k(t)) \cdot (V_1(\boldsymbol{z}_k(t))\boldsymbol{s}_1 - V_2(\boldsymbol{z}_k(t))\boldsymbol{s}_2), & t \in \cup_q B_q. \end{cases}$$

On the other hand, by (3.8),

$$\begin{aligned} &\boldsymbol{\jmath}(\boldsymbol{z}_k(t)) \cdot (V_2(\boldsymbol{z}_k(t))\boldsymbol{s}_2 - V_1(\boldsymbol{z}_k(t))\boldsymbol{s}_1) \\ &\quad = M\big[\boldsymbol{\jmath}(\boldsymbol{z}_k(t)) \cdot (\boldsymbol{s}_2 - \boldsymbol{s}_1)\big]\,\big[\boldsymbol{\jmath}(\boldsymbol{z}_k(t)) \cdot (\boldsymbol{s}_2 + \boldsymbol{s}_1) - F\big] = G(\boldsymbol{z}_k(t))\, Z(t) \end{aligned}$$

with $Z(t) = M[\boldsymbol{\jmath}(\boldsymbol{z}_k(t)) \cdot (\boldsymbol{s}_2 + \boldsymbol{s}_1) - F]$ and G defined by (3.5). Notice that $0 < Z(t) < C$ for some constant C. Hence we can write

$$D_k(t) = \begin{cases} 0, & t \in [0,T] \setminus (\cup_q (A_q \cup B_q)), \\ Z(t)|G(\boldsymbol{z}_k(t))|, & t \in \cup_q (A_q \cup B_q). \end{cases}$$

Write $A_q = [t_1, t_2]$, so that $\boldsymbol{z}_k(t_1) \in S$ and $\boldsymbol{z}_k(t) \in R(\boldsymbol{s}_2)$ for $t_1 < t \le t_2$. Since G is smooth and $\boldsymbol{z}_k(t)$ is also smooth for $t \in A_q$,

$$G(\boldsymbol{z}_k(t)) = G(\boldsymbol{z}_k(t_1)) + \big[\nabla G(\boldsymbol{z}_k(t_1)) \cdot \boldsymbol{s}_1\big]\,\big[V_1(\boldsymbol{z}_k(t_1))(t - t_1)\big] + o(h).$$

But ∇G and V_1 are bounded in a neighborhood of S, and $G(\boldsymbol{z}_k(t_1)) = 0$ since $\boldsymbol{z}_k(t_1) \in S$, and $t - t_1 < h$. Hence there exists a constant $K > 0$ such that

$$0 \le D_k(t) < Kh \quad \forall t \in [0,T],$$

and $D_k(t) \to 0$ uniformly in t as $k \to +\infty$ (i.e., as $h \to 0$), and the thesis is proved. ∎

5. Numerical simulations. In this section we present the results of numerical simulations of (2.1) and (2.5) for two different plane domains. The basic steps are as follows:

- We solve the Neumann problem (2.1) by a finite element method, for each position $\boldsymbol{z}$ of the dislocation varying on a square grid in Ω. The result is then inserted into (2.4) to determine the Peach–Köhler force $\boldsymbol{\jmath}$ on the same grid.
- Once the force field has been determined, the velocity field $\boldsymbol{V}$ in (2.6) results by projecting the force onto the crystallographic directions $\boldsymbol{s}_i \in \mathcal{C}$ and choosing the actual direction of motion by the maximum dissipation criterion. Plotting $\boldsymbol{V}$ in Ω gives some information about the phase portrait of the dynamical system (2.5), such as the form of the single slip regions, and the presence of cross-slip or attracting curves.

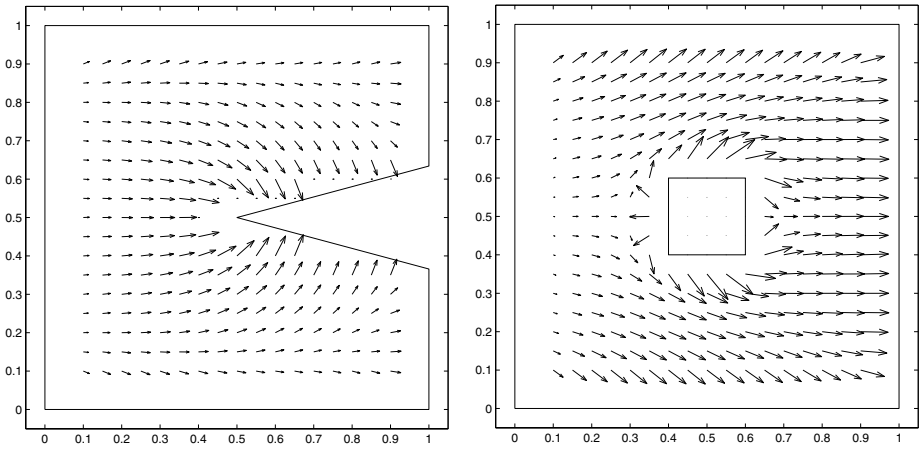


FIG. 5.1. *Plot of the vector field $\boldsymbol{\jmath}$ in Ω corresponding to shear stress boundary conditions $\sigma_0 = -1$ on the lower side of the square, $\sigma_0 = 1$ on the upper side, and $\sigma_0 = 0$ elsewhere. Left: cracked domain; right: domain clamped to a rigid inclusion.*

- Finally, we solve the dynamical system (2.5) numerically for initial conditions near an attracting curve S. We use Euler and Runge–Kutta methods in which the values of the velocity field $\boldsymbol{V}$ are determined again by solving (2.1) at each time step. Indeed, consistent with the discussion in section 4, the solution oscillates finely along the attracting curve.

In particular, we assume without loss of generality that[5]

$$M = 1, \quad F = 0,$$

and the involved domains have the following form:

(i) The first is a square domain without a wedge of angle ψ_0, i.e.,

$$\Omega = Q \backslash \left\{ (x,y) : \frac{1}{2} \le x \le 1, \frac{1}{2} - \left(\tan\frac{\psi_0}{2}\right)\left(x - \frac{1}{2}\right) \le y \le \frac{1}{2} + \left(\tan\frac{\psi_0}{2}\right)\left(x - \frac{1}{2}\right) \right\},$$

where $Q = [0,1] \times [0,1]$, and with boundary conditions

$$\sigma_0 = \begin{cases} +1 & \text{on the upper side of } Q : \{y = 1\}, \\ -1 & \text{on the lower side of } Q : \{y = 0\}, \\ 0 & \text{otherwise.} \end{cases}$$

In terms of the original three-dimensional cylindrical body, this corresponds to a shear force parallel to $\boldsymbol{e}_3$ applied to two opposite sides of the cylinder.

The field $\boldsymbol{\jmath}(\boldsymbol{x})$ representing the force on a screw dislocation in Ω is plotted in Figure 5.1 (left).

Assuming that the set $\mathcal{C}$ of crystallographic directions is as in (3.3), with $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ now given by

$$\boldsymbol{s}_1 = \frac{1}{\sqrt{2}}(\boldsymbol{e}_1 - \boldsymbol{e}_2), \qquad \boldsymbol{s}_2 = \frac{1}{\sqrt{2}}(\boldsymbol{e}_1 + \boldsymbol{e}_2), \tag{5.1}$$

[5] M is the mobility of the dislocation, which can be adjusted by rescaling time; F is the Peierls–Nabarro threshold: when $F > 0$ there are regions in which $\boldsymbol{V}(\boldsymbol{x}) = \boldsymbol{0}$ and the dislocation does not move, but elsewhere the phase portrait remains qualitatively the same as when $F = 0$.

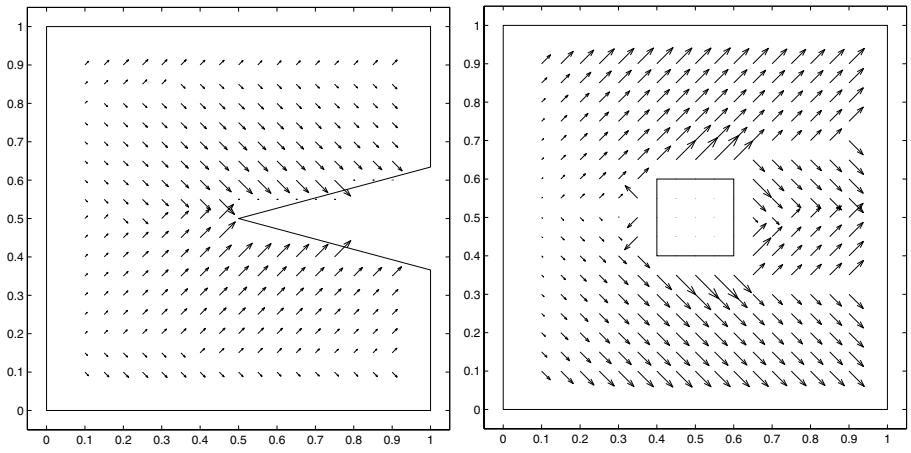


FIG. 5.2. *Plot of the vector field* $\boldsymbol{V}$ *in* Ω (*same boundary conditions as above*) *with crystallographic directions* $\boldsymbol{s}_1 = \frac{1}{\sqrt{2}}(\boldsymbol{e}_1 - \boldsymbol{e}_2)$ *and* $\boldsymbol{s}_2 = \frac{1}{\sqrt{2}}(\boldsymbol{e}_1 + \boldsymbol{e}_2)$. *Left: cracked domain; right: domain with a rigid inclusion.*

the resulting velocity field $\boldsymbol{V}(\boldsymbol{x})$ is plotted in Figure 5.2 (left). Notice that as $\psi_0 \to 0$ the domain becomes a square with a planar crack. Indeed, near the tip of the wedge, our results are consistent with the qualitative behavior of the analytical solution (6.2): the segment $S = \{0 < x < 1/2, y = 1/2\}$ is an attracting curve (see the next section).

Finally, we show in Figure 5.3 (left) some trajectories of the vector field (2.6). The trajectories have been computed by the Euler method. Notice that one of the trajectories is a fine polygonal oscillating near the attracting curve $y = 1/2$.

(ii) The second domain is a square clamped to a rigid inclusion:

$$\Omega = Q \setminus Q' \quad \text{with} \quad Q' = \left\{(x,y) : \frac{1}{2} - \frac{L}{2} \le x \le \frac{1}{2} + \frac{L}{2},\ \frac{1}{2} - \frac{L}{2} \le y \le \frac{1}{2} + \frac{L}{2}\right\},$$

and with Neumann conditions on the boundary of the outer square ∂Q,

$$\sigma_0 = \begin{cases} +1 & \text{on the upper side of } Q : \{y = 1\}, \\ -1 & \text{on the lower side of } Q : \{y = 0\}, \\ 0 & \text{on the lateral sides of } Q : \{x = 0\} \text{ and } \{x = 1\}, \end{cases}$$

and tangential-derivative conditions on the boundary of the inclusion Q', of the form

$$\boldsymbol{t} \cdot \nabla u = -\boldsymbol{g}_0 \cdot \boldsymbol{t} \quad \text{on} \quad \partial Q', \tag{5.2}$$

where $\boldsymbol{t}$ is the unit tangent vector to $\partial Q'$ and $\boldsymbol{g}_0$ is given by (2.2). The boundary conditions (5.2) may be interpreted as displacement boundary conditions, which correspond to clamping the cylinder to a rigid inclusion.

Notice in fact that, for any fixed $\boldsymbol{z}_0 \in \Omega$, the loop $\partial Q'$ does not encircle the singularity at $\boldsymbol{z}_0$. Hence, letting $\boldsymbol{g} = \boldsymbol{g}_0(\boldsymbol{x}, \boldsymbol{z}_0) + \nabla u(\boldsymbol{x}, \boldsymbol{z}_0)$, there exists a function w on $\partial Q'$ such that $\boldsymbol{g} \cdot \boldsymbol{t} = dw/ds$, with s the arc parameter on $\partial Q'$. Interpreting w as the total displacement at the boundary of Q', the condition $w = \text{const.}$ on $\partial Q'$ is equivalent to $dw/ds = 0$, which corresponds to (5.2).

The field $\boldsymbol{\jmath}(\boldsymbol{x})$ representing the force on a screw dislocation is plotted in Figure 5.1 (right). Assuming again that the set of crystallographic directions is as in (5.1), the resulting velocity field $\boldsymbol{V}(\boldsymbol{x})$ is plotted in Figure 5.2 (right).

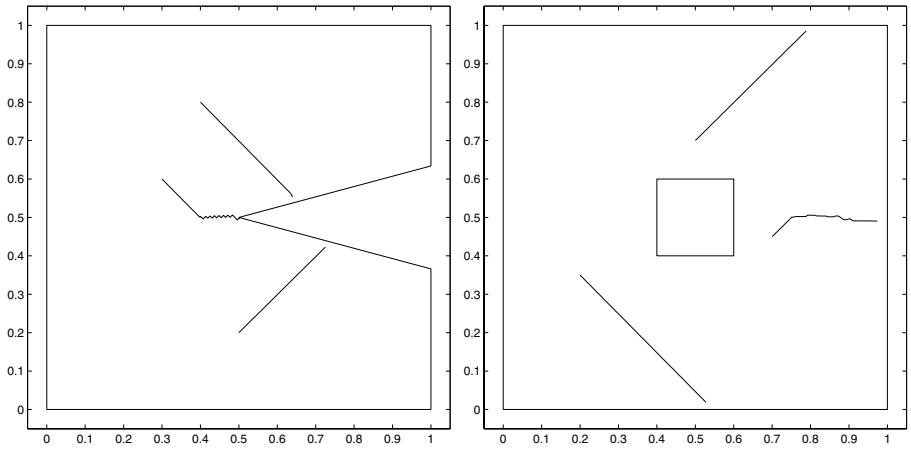


FIG. 5.3. *Some trajectories of the vector field* $\boldsymbol{V}$ *for a cracked domain and a domain with a rigid inclusion.*

Finally, we show in Figure 5.3 (right) some trajectories of the vector field (2.5). The trajectories have been computed by the Euler method. As before, notice that one of the trajectories is a fine polygonal oscillating near the attracting curve $y = 1/2$.

6. An explicit solution for a dislocation in an infinite domain with a crack. The dynamical system governing the motion of a single dislocation can be determined explicitly for some simple geometries, such as an unbounded domain with a rectilinear crack (corresponding to an unbounded cylinder with a plane crack parallel to the axis of the cylinder). Consider the plane domain

$$\Omega = \mathbb{R}^2 \setminus \{x \geq 0, y = 0\}$$

and rewrite (2.1), with $\sigma_0 = 0$, in the form

$$\begin{cases} \operatorname{Div} \boldsymbol{g} = 0 & \text{in } \Omega \setminus \{\boldsymbol{z}\}, \\ \boldsymbol{g} \cdot \boldsymbol{n} = 0 & \text{at } \partial\Omega \end{cases} \tag{6.1}$$

with $\boldsymbol{g} = \boldsymbol{g}_0 + \nabla u$ and $\boldsymbol{z} = (x_0, y_0) \in \Omega$. The explicit solution of (6.1) can be obtained using complex variables, by the procedure outlined in [1] as follows. Letting

$$\boldsymbol{g}(\boldsymbol{x}) = g_1(\boldsymbol{x})\boldsymbol{e}_1 + g_2(\boldsymbol{x})\boldsymbol{e}_2 \quad \text{with } \boldsymbol{x} = (x, y),$$

denote by $g = g(w)$ the complex function defined by

$$g(w) = g_1(w) - ig_2(w) \quad \text{with } w = x + iy.$$

Then it can be proved that [1]

$$g(w) = \frac{b}{2\pi i}\left(\frac{h'(w)}{h(w) - h(z)} + \frac{\bar{h}(z)h'(w)}{1 - \bar{h}(z)h(w)}\right),$$

where $h(w)$ is a conformal mapping from Ω into the unit disk and $z = x_0 + iy_0$. In particular, we may choose

$$h(w) = \frac{1 + i\sqrt{w}}{1 - i\sqrt{w}}$$

with $\sqrt{w}$ the principal determination of the complex square root, i.e., $\sqrt{w} = \sqrt{r}e^{i\varphi/2}$ for $w = re^{i\varphi}$ and $0 < \varphi < 2\pi$. We obtain

$$g(w) = \frac{b}{4\pi i\sqrt{w}} \frac{\sqrt{z} - \overline{\sqrt{z}}}{(\sqrt{w} - \sqrt{z})(\sqrt{w} - \overline{\sqrt{z}})}.$$

Notice that, near the tip of the crack, for $w = re^{i\varphi}$, the solution is singular:

$$|\boldsymbol{g}| \sim \frac{1}{\sqrt{r}} \quad \text{as} \quad r \to 0.$$

The same technique allows us to compute explicitly the force on the dislocation. Write

$$\boldsymbol{\jmath}(\boldsymbol{z}) = j_1(\boldsymbol{z})\boldsymbol{e}_1 + j_2(\boldsymbol{z})\boldsymbol{e}_2,$$

and consider the corresponding complex function

$$j(z) = j_1(z) - ij_2(z) \quad \text{with } z = \ x_0 + iy_0.$$

Then it can be proved that [1]

$$j(z) = \frac{b^2}{2\pi} \left\{ \frac{h''(z)}{2h'(z)} + \frac{\bar{h}(z)h'(z)}{1 - |h(z)|^2} \right\}$$

with $h(z)$ the conformal mapping from Ω into the unit disk introduced above. A lengthy calculation yields

$$\begin{aligned} j(z) &= \frac{b^2(-3\sqrt{z} + \overline{\sqrt{z}})}{8\pi z(\sqrt{z} - \overline{\sqrt{z}})} = \frac{ib^2}{16\pi r_0 \sin(\varphi_0/2)}(3e^{-i\varphi_0/2} - e^{-3i\varphi_0/2}) \\ &= \frac{b^2}{16\pi r_0 \sin(\varphi_0/2)} \left\{ \left(3\sin\frac{\varphi_0}{2} - \sin\frac{3\varphi_0}{2} \right) + i\left(3\cos\frac{\varphi_0}{2} - \cos\frac{3\varphi_0}{2} \right) \right\} \end{aligned}$$

for $z = r_0 e^{i\varphi_0}$, from which it follows that the Cartesian components of the force on a dislocation at $\boldsymbol{z} = r_0(\cos\varphi_0, \sin\varphi_0)$ are given by

(6.2)
$$\boldsymbol{\jmath}(\boldsymbol{z}) = \frac{b^2}{16\pi r_0 \sin(\varphi_0/2)} \left\{ \left(3\sin\frac{\varphi_0}{2} - \sin\frac{3\varphi_0}{2} \right) \boldsymbol{e}_1 + \left(-3\cos\frac{\varphi_0}{2} + \cos\frac{3\varphi_0}{2} \right) \boldsymbol{e}_2 \right\}.$$

Notice that $\boldsymbol{\jmath}$ is singular at the crack tip, as $r_0 \to 0$, and at the crack boundary, as $\varphi_0 \to 0, 2\pi$.

To determine the phase portrait, we must determine the single slip regions and their boundaries. Assuming that the set $\mathcal{C}$ of crystallographic directions is (3.3), with $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ given by (5.1), the single slip regions and the attracting curves can be determined explicitly. In fact, computing $\boldsymbol{\jmath} \cdot \boldsymbol{s}_1$ and $\boldsymbol{\jmath} \cdot \boldsymbol{s}_2$, we obtain

$$\boldsymbol{\jmath}(r_0, \varphi_0) \cdot \boldsymbol{s}_1 > \boldsymbol{\jmath}(r_0, \varphi_0) \cdot \boldsymbol{s}_2 \Leftrightarrow \cos\frac{\varphi_0}{2} \left(3 - 2\cos^2\frac{\varphi_0}{2} \right) > 0 \Leftrightarrow \varphi_0 \in (0, \pi),$$

so that the single slip regions $R(\boldsymbol{s}_1)$ and $R(\boldsymbol{s}_2)$ are the upper and lower half-plane, respectively.

The curve S separating $R(\boldsymbol{s}_1)$ and $R(\boldsymbol{s}_2)$ is an attracting curve, given by the relation $\boldsymbol{\jmath}(r_0, \varphi_0) \cdot \boldsymbol{s}_1 = \boldsymbol{\jmath}(r_0, \varphi_0) \cdot \boldsymbol{s}_2$, and is therefore the negative real half-line $\varphi_0 = \pi$.

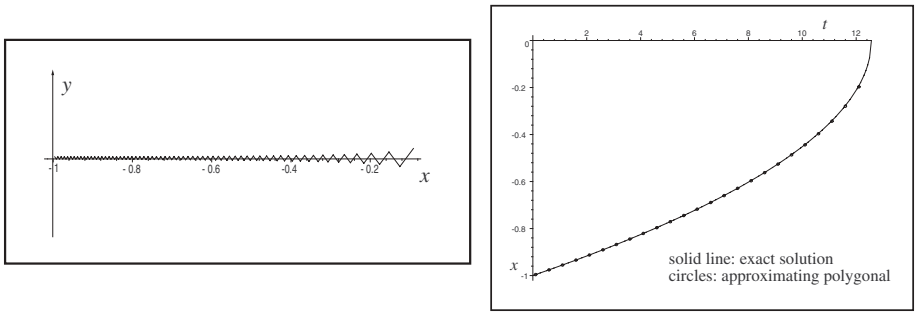


FIG. 6.1. *Left: trajectory of an approximating polygonal near the tip of the crack for* $h = 0.1$ (*Euler method*)*; right: comparison of the limit motion* $x(t)$ *and the horizontal component of the approximating polygonal.*

The smooth motion on S corresponding to fine cross-slip can be detemined explicitly: since

$$\boldsymbol{\jmath}(r_0, \pi) = \frac{1}{4\pi r_0}\boldsymbol{e}_1,$$

letting $\boldsymbol{\xi}(t) = x(t)\boldsymbol{e}_1$, the homogenized evolution equation (3.21) becomes

$$\dot{x}_0 = \frac{1}{8\pi|x_0|},$$

whose solution with the initial condition $x(0) = -1$ is

$$x(t) = -\frac{1}{2\pi}\sqrt{4\pi - t}. \tag{6.3}$$

Hence the dislocation reaches the tip of the crack at $x = 0$ in finite time.

In Figure 6.1 we compare the exact solution (6.3) with the horizontal component of the numerical solution (Euler method) of the dynamical system (2.5). The agreement is good already for the time step $h = 0.1$.

REFERENCES

[1] E. BUZANO AND P. CERMELLI, *A singular variational problem in dislocation theory*, Z. Angew. Math. Phys., 51 (2000), pp. 968–983.
[2] P. CERMELLI AND M. E. GURTIN, *The motion of screw dislocations in materials undergoing anti-plane shear: Glide, cross-slip, fine cross-slip*, Arch. Ration. Mech. Anal., 148 (1999), pp. 3–52.
[3] J. D. ESHELBY, *Energy relations and the energy-momentum tensor in continuum mechanics*, in Inelastic Behavior of Solids, M. Kanninen, W. Adler, A. Rosenfield, and R. Jaffee, eds., McGraw–Hill, New York, 1970, pp. 77–115.
[4] M. E. GURTIN, *Configurational Forces as Basic Concepts of Continuum Physics*, Springer, New York, 2001.
[5] D. KUHLMANN-WILSDORF, *Theory of plastic deformation: Properties of low energy dislocation structures*, Mater. Sci. Engrg., A113 (1989), pp. 1–41.
[6] P. HAASEN, *Physical Metallurgy*, Cambridge University Press, Cambridge, UK, 1996.
[7] J. P. HIRTH AND J. LOTHE, *Theory of Dislocations*, 2nd ed., McGraw–Hill, New York, 1982.
[8] F. R. N. NABARRO, *Theory of Crystal Dislocations*, Clarendon Press, Oxford, UK, 1967.
[9] M. PEACH AND J. S. KÖHLER, *The forces exerted on dislocations and the stress fields produced by them*, Phys. Rev., 80 (1950), pp. 436–439.

[10] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser, Basel, 1997.
[11] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, New York, 1966.
[12] V. Volterra, *Sur l'équilibre del corps élastiques multiplement connexes*, Ann. Ec. Norm., 24 (1907), pp. 401–51.
[13] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, Saunders, Philadelphia, 1969.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 2144–2166

RUPTURE OF A SURFACTANT-COVERED THIN LIQUID FILM ON A FLEXIBLE WALL*

OMAR K. MATAR† AND SATISH KUMAR‡

Abstract. The rupture of a surfactant-covered thin liquid film on a flexible wall is studied in this paper. Evolution equations for the deflections of the air-liquid and wall-liquid interfaces and surfactant surface concentration are derived using lubrication theory, and their linear and nonlinear stability characteristics are investigated. Our linear stability results indicate that increases in the level of damping, the longitudinal wall tension, and the relative magnitude of Marangoni stresses have a stabilizing influence. Numerical simulations of the evolution equations are used to investigate the nonlinear characteristics of the instability. In all cases considered, the surfactant concentration decreases in the rupture region as rupture is approached, and the resulting Marangoni flows retard but do not prevent rupture. Self-similar rupture is examined and power-law scalings are extracted for different parameter values. These appear to be unchanged from those for rigid substrates, evidently because the van der Waals forces that drive the instability dominate the rupture dynamics.

Key words. thin liquid films, flexible walls, surfactants, rupture

AMS subject classifications. 74F10, 76D08, 76D45, 76E17

DOI. 10.1137/S003613990242002X

1. Introduction.

Flow in thin viscous films has been the subject of numerous studies in the literature due to its obvious importance in a wide range of industrial and biomedical applications [1]. Many of these studies have addressed the problem of film rupture, where the film either rests on a solid horizontal support [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13] or is freely suspended [14, 15, 16, 17, 18, 19, 20, 21, 22]. The former situation, for instance, models the spreading of liquids and surfactant solutions on the surface of a solid or of another liquid that rests on a solid wall. The latter situation, on the other hand, represents the rupture of a soap film or of the continuous film which accompanies the coalescence of two droplets in an emulsion. A number of these studies have examined the evolution and self-similar nature of film thickness solutions as rupture is approached in the presence [9, 10, 11] and absence [4, 15, 13] of surfactant. In all cases, rupture is driven by intermolecular interactions such as van der Waals, hydration, depletion, or electrostatic forces, which come into operation for very small film thicknesses, typically on the order of 100 nm or less [23]. Studies of film rupture on solid walls generally assume that the solid is rigid. In this work, we relax that assumption and examine the rupture of a surfactant-covered thin liquid film on a flexible wall.

From a scientific standpoint, the rupture of thin liquid films on flexible walls is of interest because it is a problem of elastohydrodynamics: the study of the interaction between flowing fluids and flexible elastic structures. Such problems have

*Received by the editors December 19, 2002; accepted for publication (in revised form) February 19, 2004; published electronically September 14, 2004. This research was supported in part by the Petroleum Research Fund, administered by the American Chemical Society.

http://www.siam.org/journals/siap/64-6/42002.html

†Department of Chemical Engineering and Chemical Technology, Imperial College London, South Kensington Campus, London, SW7 2AZ, UK (o.matar@ic.ac.uk). The research of this author was supported in part by EPSRC grant GR/N 34895/01.

‡Department of Chemical Engineering and Materials Science, University of Minnesota, 151 Amundson Hall, 421 Washington Avenue SE, Minneapolis, MN 55455. The research of this author was supported in part by the Shell Oil Company Foundation through its Faculty Career Initiation Funds program and by a nontenured faculty award from 3M.

been examined by hydrodynamicists for years; examples include the use of flexible boundaries to delay the transition to turbulence [24], the modeling of airflow in the lungs [25] and blood flow in the heart [26], the use of rubber-covered rolls in coating processes to reduce defects [27], and flow instabilities near gel-like polymer interfaces [28]. Elastohydrodynamic problems are challenging because they involve two coupled media whose interface is often a free boundary. However, to our knowledge, elastohydrodynamic phenomena associated with film rupture have yet to be examined. From a practical standpoint, such a study may be relevant to several applications. The first involves the creation of textured or topographically patterned solids: if thin-film rupture leads to deformation of the underlying solid, then it may be possible to use rupture as a way to create surface features on the solid. The second application involves human lungs, where airways are flexible walls lined with a thin liquid film [29, 30]. If the film becomes thin enough, intermolecular forces may drive rupture and this in turn may deform the walls and lead to airway closure. Halpern and Grotberg have applied lubrication theory to study the instabilities associated with this system, but their analysis does not include intermolecular forces such as van der Waals interactions [29, 30]. Finally, studies of film rupture on flexible walls may be relevant in modeling the adhesion of cells and vesicles to solid substrates as discussed by Ramos de Souza, Gallez, and coworkers [20, 21, 22]. Here, the liquid-air interface would represent the cell or vesicle membrane, whereas the flexible wall would represent a soft substrate. If the wall is sufficiently compliant, it could become deformed during film rupture and this may modify the dynamics of cell or vesicle adhesion.

As a model system in the present work, we consider the linear stability and nonlinear evolution of a thin liquid film covered with an insoluble surfactant that rests on a flexible wall and is driven to rupture by van der Waals forces. Although other intermolecular forces may be present and lead to scenarios such as the formation of steady patterns [20, 21, 22], we restrict ourselves to van der Waals forces in this first study since our focus is on film rupture. Using lubrication theory, coupled evolution equations are derived for the deflection of the air-liquid and wall-liquid interfaces as well as for the surfactant concentration. A linear stability analysis and numerical simulations are conducted to determine the behavior of the film at the onset of the instability and as rupture is approached, respectively. This is carried out over a wide range of system parameters, which encompass the limits of weak and strong wall damping and longitudinal tension, and in the presence and absence of surfactant. Our results indicate that increasing the relative significance of wall damping retards (accelerates) the onset of van der Waals–driven film rupture for low (high) wall tensions; increasing wall tension was also found to delay rupture for all the damping coefficients examined. Moreover, the presence of Marangoni stresses [31], which arise due to gradients in the local surfactant concentration, exerts a stabilizing influence, acting to oppose the van der Waals–driven thinning process.

The rest of this paper is organized as follows. Details of the problem formulation are given in section 2, while results of the linear stability analysis are presented in section 3. A discussion of the numerical solutions together with a brief description of the numerical method employed to carry out the computations are provided in section 4. Finally, concluding remarks are given in section 5.

2. Formulation.

2.1. Governing equations. We consider a thin film of an incompressible Newtonian fluid of initial thickness $\mathcal{H}$, lateral extent $\mathcal{L}$, viscosity μ, and density ρ resting on a flexible wall. The air-liquid interface, which is bounded from above by an invis-

cid gas, is covered by an initially uniform dilute concentration of insoluble surfactant, $\Gamma_m \ll \Gamma_\infty$, where Γ_∞ represents the concentration at saturation. The analysis in the present work is restricted to planar geometry with coordinate system (x, y, z) and velocity field $\mathbf{u} = (u, 0, w)$, where x, y, and z denote the horizontal, transverse, and vertical coordinates, respectively, while $\mathbf{u}$ denotes the velocity field in which u and w represent the horizontal and vertical components of the velocity field, respectively. The origin of the z coordinate coincides with the midpoint of the initially undisturbed film thickness, such that the instantaneous location of the air-liquid interface is at $z = h + \mathcal{H}/2$, while the wall-liquid interface is located at $z = -\eta - \mathcal{H}/2$. Here, h and η denote deflections of the air-liquid and wall-liquid interfaces from their initially uniform states.

The film dynamics are governed by the equations of conservation of mass and momentum, respectively, given by

$$\nabla \cdot \mathbf{u} = 0, \tag{1}$$

$$\rho \left(\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} \right) = -\nabla \left(p + \phi \right) + \mu \nabla^2 \mathbf{u}, \tag{2}$$

where p represents pressure and $\phi = A/(\mathcal{H} + h + \eta)^3$ is an energy per unit volume corresponding to the van der Waals component of the disjoining pressure, in which A is the so-called Hamaker constant [23]; subscript notation denotes differentiation with respect to the spatial variables and time unless otherwise stated. Note that we have assumed the film to be sufficiently thin so that gravitational effects are negligible.

Equations (1) and (2) are complemented by an appropriate set of boundary conditions. At the air-liquid interface, $z = h + \mathcal{H}/2$, we have the normal and shear stress balances, respectively, given by

$$\mathbf{n} \cdot \mathbf{T} \cdot \mathbf{n} = \sigma \kappa, \tag{3}$$

$$\mathbf{n} \cdot \mathbf{T} \cdot \mathbf{t} = \nabla_s \sigma \cdot \mathbf{t}. \tag{4}$$

In (3) and (4), σ denotes the local value of the surface tension coefficient; $\mathbf{n} = (-h_x, 1)/\Delta$ and $\mathbf{t} = (1, h_x)/\Delta$, in which $\Delta \equiv (1 + h_x^2)^{1/2}$, denote the outward pointing unit normal and unit tangent to the air-liquid interface; and $\kappa = \nabla_s \cdot \mathbf{n}$ is the curvature where $\nabla_s = (\mathbf{I} - \mathbf{nn}) \cdot \nabla$ is the surface gradient operator. Also appearing in (3) and (4) is the film stress tensor:

$$\mathbf{T} = -p\mathbf{I} + \mu \left(\nabla \mathbf{u} + \nabla \mathbf{u}^T \right), \tag{5}$$

where $\mathbf{I}$ is the identity tensor. The kinematic boundary condition at $z = h + \mathcal{H}/2$ is expressed by

$$h_t + u_s h_x = w_s, \tag{6}$$

where the subscript s represents quantities evaluated at $z = h + \mathcal{H}/2$. The velocity field in the film must also satisfy the no-slip and kinematic boundary conditions at the wall-liquid interface, $z = -\eta - \mathcal{H}/2$:

$$u_w = 0, \tag{7}$$

$$\eta_t + u_w \eta_x = -w_w, \tag{8}$$

where the subscript w represents quantities evaluated at $z = -\eta - \mathcal{H}/2$.

The dynamics of the surfactant present at $z = h + \mathcal{H}/2$ are governed by the mass conservation equation [32]

$$\Gamma_t + (\nabla_s \cdot \mathbf{n})\Gamma(\mathbf{n} \cdot \mathbf{u}) + \nabla_s \cdot (\mathbf{u}_s \Gamma) = \mathcal{D}_s \nabla_s^2 \Gamma + J, \tag{9}$$

where $\mathcal{D}_s$ is a surface diffusion coefficient, which we will take to be constant, and J denotes a sorptive flux, which will be neglected in the present work since insoluble surfactants are considered. Note that σ depends on Γ through the equation of state [31],

$$\sigma = \sigma_o - (-\sigma_\Gamma)_{\Gamma=0}\Gamma, \tag{10}$$

where σ_o denotes the surface tension of the uncontaminated air-liquid interface while $-\sigma_\Gamma$ is a measure of surfactant activity.

Our model for the wall dynamics is very similar to that used by Halpern and Grotberg in their studies of liquid film dynamics inside flexible tubes [29, 30]. The wall is assumed to be infinitely long, isotropic, and impermeable with thickness δ, density ρ_w, and damping coefficient g. In addition, the wall is assumed to be sufficiently thin so that when a longitudinal (horizontal) tension of magnitude T_l is applied, it acts uniformly across the wall thickness. (Under these conditions, stresses due to bending may be neglected [33, 34].)

The film aspect ratio, $\epsilon \equiv \mathcal{H}/\mathcal{L}$, is taken to be small, which permits the use of lubrication theory, and in the limit $\epsilon \ll 1$, it can be shown that longitudinal deflections of the wall are much smaller than those in the vertical direction [35, 33, 29]. The relevant relation governing the wall deflections is given by

$$\frac{\rho_w \delta g \eta_t}{\Delta_w} - \frac{T_l \eta_{xx}}{\Delta_w^3} = -\mathbf{n}_w \cdot \mathbf{T} \cdot \mathbf{n}_w, \tag{11}$$

where $\mathbf{n}_w = (-\eta_x, 1)/\Delta_w$ is the unit normal to the wall-liquid interface in which $\Delta_w = (1 + \eta_x^2)^{1/2}$. The relevant scalings are presented next.

2.2. Scaling. The governing equations and boundary conditions are rendered dimensionless via the scalings for the hydrodynamic variables,

$$x = \mathcal{L}\tilde{x}, \quad z = \mathcal{H}(\tilde{z}, \tilde{h}, \tilde{\eta}), \quad u = \mathcal{U}\tilde{u}, \quad w = \epsilon\mathcal{U}\tilde{w}, \quad t = (\mathcal{L}/\mathcal{U})\tilde{t}, \quad p = \mathcal{P}\tilde{p}, \tag{12}$$

and for σ and Γ,

$$\Gamma = \Gamma_m \tilde{\Gamma}, \quad \sigma = \sigma_m + \mathcal{S}\tilde{\sigma}, \tag{13}$$

where $\mathcal{S} \equiv \sigma_o - \sigma_m$ is the spreading coefficient in which σ_m is the surface tension of the air-liquid interface when $\Gamma = \Gamma_m$. Here, $\mathcal{L} \equiv \mathcal{H}^2 (\sigma_m/A)^{1/2}$, $\mathcal{U} \equiv A/(\mu\mathcal{H}\mathcal{L})$, and $\mathcal{P} \equiv A/\mathcal{H}^3$, reflecting a balance between van der Waals, capillary, and viscous forces. Note that for typical values, $\mathcal{H} \sim 10^{-4}$ cm, $A \sim 10^{-13}$ erg, and $\sigma_m \sim 10$ dyne/cm, we have $\epsilon = \mathcal{H}/\mathcal{L} \sim O(10^{-3})$.

2.3. Lubrication theory. Substitution of the relevant scalings into the governing equations and boundary conditions yields the following set of dimensionless equations for the liquid film and surfactant concentration to leading order in ϵ (after suppressing the tilde):

$$u_x + w_z = 0, \tag{14}$$

$$0 = -(p+\phi)_x + u_{zz} + O(\epsilon^2, \epsilon \mathrm{Re}), \tag{15}$$

where $\phi = 1/(1+h+\eta)^3$,

$$0 = -p_z + O(\epsilon^2, \epsilon^3 \mathrm{Re}), \tag{16}$$

where $\mathrm{Re} \equiv \rho \mathcal{U}\mathcal{H}/\mu$ is the Reynolds number,

$$p = -h_{xx} + O(\epsilon^2) \qquad \text{at} \quad z = h + \frac{1}{2}, \tag{17}$$

$$u_z = \mathcal{M}\sigma_x + O(\epsilon^2) \qquad \text{at} \quad z = h + \frac{1}{2}, \tag{18}$$

where $\mathcal{M} \equiv \mathcal{S}\mathcal{H}^2/A$ is a Marangoni parameter, representing the magnitude of Marangoni stresses to van der Waals forces,

$$h_t + u_s h_x = w_s \qquad \text{at} \quad z = h + \frac{1}{2}, \tag{19}$$

$$u_w = 0 \qquad \text{at} \quad z = -\eta - \frac{1}{2}, \tag{20}$$

$$\eta_t = -w_w \qquad \text{at} \quad z = -\eta - \frac{1}{2}, \tag{21}$$

$$\Gamma_t + (u_s \Gamma)_x = \frac{\Gamma_{xx}}{\mathrm{Pe}} \qquad \text{at} \quad z = h + \frac{1}{2}. \tag{22}$$

Here, $\mathrm{Pe} \equiv \mathcal{U}\mathcal{L}/\mathcal{D}_s$ denotes the Peclet number, a ratio of the time scale for surfactant diffusion to that for surfactant convection.

Integration of (16) and application of (17) leads to the fact that the leading order pressure is independent of z and is given by

$$p = -h_{xx}. \tag{23}$$

Following the integration of (15) and application of (18) and (20), the leading order horizontal velocity is obtained:

$$u(x,z,t) = (p+\phi)_x \left[\frac{z^2}{2} - z\left(\frac{1}{2}+h\right) - \left(\frac{1}{2}+\eta\right)\left(\frac{3}{4}+h+\frac{\eta}{2}\right)\right] + \mathcal{M}\sigma_x \left(z + \frac{1}{2} + \eta\right). \tag{24}$$

Using continuity (and the Leibniz rule), (19) may be reexpressed as

$$h_t + \eta_t + Q_x = 0, \tag{25}$$

in which Q, the flow rate, is given by

$$Q = \int_{-(\frac{1}{2}+\eta)}^{\frac{1}{2}+h} u dz. \tag{26}$$

Substitution of (24) into (26) yields

$$Q = \frac{1}{2}(1+h+\eta)^2 \sigma_x - \frac{1}{3}(1+h+\eta)^3 (p+\phi)_x. \tag{27}$$

The dimensionless equation of state relating σ to Γ is given by

$$\sigma = 1 - \Gamma. \tag{28}$$

Hence, the evolution equation governing the dynamics of h becomes

$$h_t = -\eta_t + \left[\frac{1}{2}(1+h+\eta)^2 \mathcal{M}\Gamma_x - \frac{1}{3}(1+h+\eta)^3 \left(h_{xx} - \frac{1}{(1+h+\eta)^3}\right)_x\right]_x. \tag{29}$$

The velocity at $z = h + 1/2$, u_s, is given by

$$u_s = -\frac{1}{2}(1+h+\eta)^2 (p+\phi)_x + (1+h+\eta)\mathcal{M}\sigma_x. \tag{30}$$

Substitution of (30), along with (28), into (22) yields the following dimensionless evolution equation for Γ:

$$\Gamma_t = \left[(1+h+\eta)\mathcal{M}\Gamma\Gamma_x - \frac{1}{2}(1+h+\eta)^2 \Gamma \left(h_{xx} - \frac{1}{(1+h+\eta)^3}\right)_x\right]_x + \frac{\Gamma_{xx}}{\mathrm{Pe}}. \tag{31}$$

The dimensionless evolution equation governing the wall deflection is given by

$$\epsilon^2 \mathcal{R}\mathcal{G}\eta_t - \epsilon^2 \mathcal{T}\eta_{xx} + h_{xx} = 0 + O(\epsilon), \tag{32}$$

where $\mathcal{R} \equiv \rho_w \delta/(\rho\mathcal{H})$ represents the ratio of the mass of the wall to that of the fluid, $\mathcal{G} \equiv \rho g \mathcal{H}^2/\mu$ reflects the relative importance of wall damping to fluid damping, while $\mathcal{T} \equiv T_l \mathcal{H}^2/A$ corresponds to the ratio of the wall longitudinal tension to van der Waals forces.

Equation (29) is fully coupled to (31) and (32). Note that in the limit $\eta \to 0$ and $1 + h \to h$, (29), (31), and (32) reduce to the evolution equations for a surfactant covered thin liquid film resting on a rigid support (except for rescalings) [4].

For parameter values typical of terminal bronchioles [29], $\rho_w \sim 1$ g cm^{-3}, $\delta \sim 10^{-3}$ cm, $g \sim 10$ s^{-1}, $\mu \sim 10^{-2}$ poise, $T_l \sim 10$ dynes cm^{-1}, $\mathcal{R} \sim O(10)$, $\mathcal{G} \sim O(10^{-5})$, and $\mathcal{T} \sim O(10^6)$. This suggests that, for this setting, the wall is weakly damped and highly tensile longitudinally. It may therefore be possible to rescale $\mathcal{T}$ such that $\mathcal{T} = \hat{\mathcal{T}}/\epsilon^2$ with $\hat{\mathcal{T}} \sim O(1)$ and to eliminate the transient term in (32) to leading order:

$$-\hat{\mathcal{T}}\eta_{xx} + h_{xx} = 0 + O(\epsilon^2). \tag{33}$$

In fact, it is possible to consider different situations: $\epsilon^2\mathcal{R}\mathcal{G} \sim O(\epsilon^2)$ or $\epsilon^2\mathcal{R}\mathcal{G} \sim O(1)$, corresponding to either a weakly or strongly damped wall, respectively, and $\epsilon^2\mathcal{T} \sim O(\epsilon^2)$ or $\epsilon^2\mathcal{T} \sim O(1)$, which represent weak or strong longitudinal tension, respectively. These situations will be considered in the following sections. First, however, the linear stability characteristics of the system are analyzed.

3. Linear stability analysis. To gain insight into the problem and to provide a useful check on the performance of the numerical scheme utilized for the solution of the governing equations, a linear stability analysis is conducted. Equations (29), (31), and (32) are linearized using normal modes:

$$(h, \eta, \Gamma)(x,t) = (0, 0, \Gamma_0) + (H, F, G)\, e^{\omega t} e^{ikx}, \tag{34}$$

where $h = 0$, $\eta = 0$, and $\Gamma = \Gamma_0$ denote the spatially uniform base state and H, F, and G denote the amplitude of the initially infinitesimal applied disturbance of (real)

wavenumber k whose potentially complex growth rate is given by ω. Substitution of (34) into (29), (31), and (32) and neglecting quadratic and higher order terms in the perturbations leads to the following coupled algebraic eigenvalue equations:

$$\omega H = -\omega F - \frac{1}{2}\mathcal{M}k^2 G - \frac{1}{3}k^4 H + k^2\left(H + F\right), \tag{35}$$

$$\omega G = -\left(\mathcal{M}\Gamma_0 + \frac{1}{\mathrm{Pe}}\right)k^2 G - \frac{1}{2}\Gamma_0 k^4 H + \frac{3}{2}\Gamma_0 k^2\left(H + F\right), \tag{36}$$

$$\epsilon^2\mathcal{R}\mathcal{G}\omega F = -\epsilon^2\mathcal{T}k^2 F + k^2 H. \tag{37}$$

Equations (35)–(37) are then expressed in matrix form. Setting the determinant of this matrix to zero yields a characteristic equation for the growth rate:

$$\begin{aligned}(\epsilon^2\mathcal{R}\mathcal{G})\omega^3 + k^2\left(\epsilon^2\mathcal{T} + 1 + \epsilon^2\mathcal{R}\mathcal{G}\left[\frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 + \frac{k^2}{3} - 1\right]\right)\omega^2 \\ + k^4\left(\frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 - 1 + \epsilon^2\mathcal{R}\mathcal{G}\left[\frac{1}{\mathrm{Pe}} + \frac{\mathcal{M}\Gamma_0}{4}\right]\left[\frac{k^2}{3} - 1\right] + \epsilon^2\mathcal{T}\left[\frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 + \frac{k^2}{3} - 1\right]\right)\omega \\ - k^6\left(\frac{1}{\mathrm{Pe}} + \frac{\mathcal{M}\Gamma_0}{4}\right)\left(1 - \epsilon^2\mathcal{T}\left[\frac{k^2}{3} - 1\right]\right) = 0.\end{aligned} \tag{38}$$

Solution of this equation using Mathematica yields dispersion curves, which represent the dependence of $Re[\omega]$ on k as a function of system parameters. The existence of a band of wavenumbers for a given set of parameters over which $Re[\omega] > 0$ signifies the presence of a linear instability.

For a weakly damped wall, that is, for $\epsilon^2\mathcal{R}\mathcal{G} \ll 1$, (38) reduces to

$$\begin{aligned}(\epsilon^2\mathcal{T} + 1)\omega^2 + k^2\left(\epsilon^2\mathcal{T}\left[\frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 + \frac{k^2}{3} - 1\right] + \frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 - 1\right)\omega \\ - k^4\left[\frac{1}{\mathrm{Pe}} + \frac{\mathcal{M}\Gamma_0}{4}\right]\left[1 - \epsilon^2\mathcal{T}\left(\frac{k^2}{3} - 1\right)\right] = 0,\end{aligned} \tag{39}$$

while for a weakly (longitudinally) tensile wall, (38) reduces to

$$\begin{aligned}\left(\epsilon^2\mathcal{R}\mathcal{G}\right)\omega^3 + k^2\left(\epsilon^2\mathcal{R}\mathcal{G}\left[\frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 + \frac{k^2}{3} - 1\right] + 1\right)\omega^2 \\ + k^4\left(\epsilon^2\mathcal{R}\mathcal{G}\left[\frac{1}{\mathrm{Pe}} + \frac{\mathcal{M}\Gamma_0}{4}\right]\left[\frac{k^2}{3} - 1\right] + \frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 - 1\right)\omega - k^6\left(\frac{1}{\mathrm{Pe}} + \frac{\mathcal{M}\Gamma_0}{4}\right) = 0,\end{aligned} \tag{40}$$

and, in the absence of surfactant, we have

$$(\epsilon^2\mathcal{R}\mathcal{G})\omega^2 + k^2\left(\epsilon^2\mathcal{T} + 1 + \epsilon^2\mathcal{R}\mathcal{G}\left[\frac{k^2}{3} - 1\right]\right)\omega + k^4\left(-1 + \epsilon^2\mathcal{T}\left[\frac{k^2}{3} - 1\right]\right) = 0. \tag{41}$$

For the case of a very highly damped or a very highly (longitudinally) tensile wall, corresponding to $\epsilon^2\mathcal{R}\mathcal{G} \gg 1$ and $\epsilon^2\mathcal{T} \gg 1$, respectively, (38) reduces to

$$\omega^2 + k^2\left(\frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 + \frac{k^2}{3} - 1\right)\omega + k^4\left(\frac{k^2}{3} - 1\right)\left(\frac{1}{\mathrm{Pe}} + \frac{\mathcal{M}\Gamma_0}{4}\right) = 0, \tag{42}$$

which corresponds to the equation describing the linear stability characteristics of a thin viscous film resting on a rigid horizontal support undergoing van der Waals driven rupture in the presence of surfactant-induced Marangoni stresses [12].

From (38), the wavenumber corresponding to the cutoff mode, k_c, for which $\omega = 0$, is expressed by

$$k_c = \sqrt{3\left(\frac{1+\epsilon^2\mathcal{T}}{\epsilon^2\mathcal{T}}\right)}. \tag{43}$$

Inspection of (43), which represents a generalization of k_c for a rupturing thin film resting on a flexible wall, reveals that the cut-off mode is unaffected either by the presence of surfactant, in agreement with previous work [3, 4, 12, 36], or the effect of wall damping. Note that for $\epsilon^2\mathcal{T} \gg 1$, $k_c = \sqrt{3}$ [37, 4, 12, 36], which reflects a balance between van der Waals and capillary forces only. Marangoni stresses and wall damping effects will, however, have an influence on the magnitude of the wavenumber corresponding to the so-called most dangerous mode and its associated maximal growth rate. Also note that for $\epsilon^2\mathcal{T} \ll 1$, $k_c \sim \sqrt{3/(\epsilon^2\mathcal{T})}$, which indicates that the equations may become ill posed in the limit of very weak longitudinal tension; this can also be ascertained on inspection of (40). We leave this case aside and concentrate in the present section on film rupture on a highly tensile wall in the limits of weak and significant damping, which we consider next.

3.1. Weakly damped, highly tensile wall. Here, we set $\mathcal{T} = \hat{\mathcal{T}}/\epsilon^2$ and $\mathcal{RG} \sim O(1)$; (38) becomes

$$\begin{aligned}(\hat{\mathcal{T}}+1)\omega^2 + k^2\left(\hat{\mathcal{T}}\left[\frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 + \frac{k^2}{3} - 1\right] + \frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 - 1\right)\omega \\ - k^4\left[\frac{1}{\mathrm{Pe}} + \frac{\mathcal{M}\Gamma_0}{4}\right]\left[1 - \hat{\mathcal{T}}\left(\frac{k^2}{3} - 1\right)\right] = 0.\end{aligned} \tag{44}$$

Thus, ω is given by

$$\omega_{\pm} = \frac{\left(-k^4(\hat{\mathcal{T}}[Q_1 + \frac{k^2}{3} - 1] + Q_1 - 1)\left[1 \pm \left(1 + \frac{4(\hat{\mathcal{T}}+1)Q_2\left(1-\hat{\mathcal{T}}\left(\frac{k^2}{3}-1\right)\right)}{\left(\hat{\mathcal{T}}\left(Q_1 - 1 + \frac{k^2}{3}\right) + Q_1 - 1\right)^2}\right)^{1/2}\right]\right)}{2(\hat{\mathcal{T}}+1)}, \tag{45}$$

in which $Q_1 \equiv \mathcal{M}\Gamma_0 + 1/\mathrm{Pe}$ and $Q_2 \equiv (\mathcal{M}\Gamma_0/4) + 1/\mathrm{Pe}$; here, k_c is given by (43). In Figure 1, we show the effect of varying $\hat{\mathcal{T}}$ on the behavior of the dispersion curves with $\Gamma_0 = 1$, $\mathcal{M} = 1$, and $\mathrm{Pe} = 100$. (We consider only ω_+, the dominant mode.) Clearly, increasing $\hat{\mathcal{T}}$ decreases k_c and the magnitudes of the growth rate and most dangerous mode. Thus, longitudinal tension exerts a stabilizing influence.

Next we investigate the effect of Marangoni stresses on the dispersion curves. Since the parameters $\mathcal{M}$ and Γ_0 arise as a product in (45), only one of these parameters will be varied. Inspection of Figure 2, which shows the effect of varying $\mathcal{M}$ on the linear stability characteristics, reveals that increasing $\mathcal{M}$ stabilizes the flow, decreasing the magnitude of ω_+, in agreement with previous studies of film rupture on rigid solid substrates [3, 4, 12, 36]. It is worthy of mention that the magnitude of k_c remains unaltered, in agreement with (43), and the changes in ω_+ rapidly saturate with increasing $\mathcal{M}$. We have also found that the dispersion curves exhibit a very weak dependence on Pe (not shown) for physically realizable Pe values.

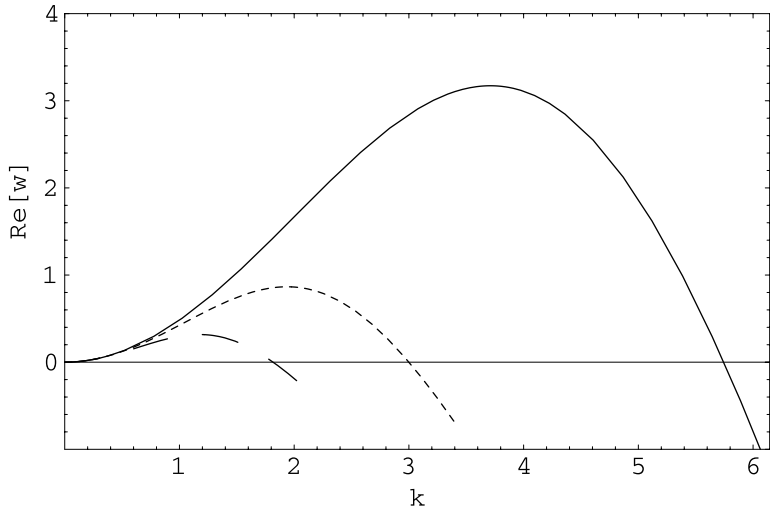


FIG. 1. *The effect of varying* $\hat{\mathcal{T}}$ *on the dispersion curves obtained via the solution of* (45) *with* $\hat{\mathcal{T}} = 0.1$ *(solid lines),* $\hat{\mathcal{T}} = 0.5$ *(short-dashed lines), and* $\hat{\mathcal{T}} = 10$ *(dashed lines),* $\Gamma_0 = 1$, $\mathcal{M} = 1$, *and* $\mathrm{Pe} = 100$.

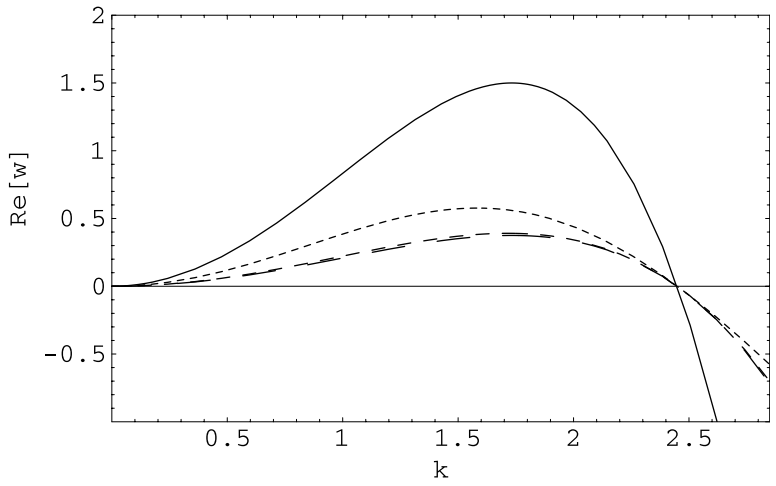


FIG. 2. *The effect of varying* $\mathcal{M}$ *on the dispersion curves obtained via the solution of* (45) *with* $\mathcal{M} = 0$ *(solid lines),* $\mathcal{M} = 1$ *(short-dashed lines),* $\mathcal{M} = 10$ *(dashed lines), and* $\mathcal{M} = 10^5$ *(long-dashed lines),* $\hat{\mathcal{T}} = 1$, $\Gamma_0 = 1$, *and* $\mathrm{Pe} = 100$.

3.2. Strongly damped, highly tensile wall. We turn our attention to the case of a highly tensile, strongly damped wall. Here, $\mathcal{T}$ remains as $\mathcal{T} = \hat{\mathcal{T}}/\epsilon^2$ and we set $\mathcal{RG} = \hat{\mathcal{B}}/\epsilon^2$ with $\hat{\mathcal{B}} \sim O(1)$. We note that inclusion of significant wall damping reintroduces temporal variations of wall deflections, which corresponds to the term proportional to ω^3 in (38); this term had been omitted from the previous section. Equation (38) then becomes

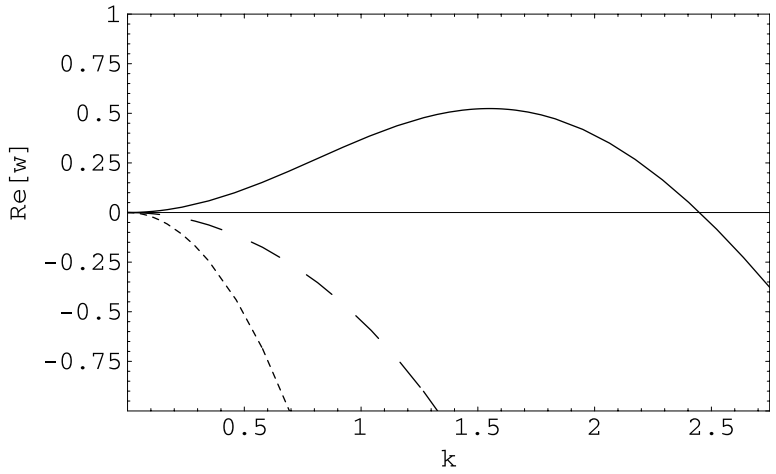


FIG. 3. *Dispersion curves obtained via the solution of* (46) *with* $\hat{\mathcal{B}} = \hat{\mathcal{T}} = \mathcal{M} = \Gamma_0 = 1$ *and* $\mathrm{Pe} = 100$*:* ω_1 *(solid line),* ω_2 *(short-dashed line), and* ω_3 *(dashed line), which are the three roots of* (46)*. Here,* ω_1 *and* ω_2 *are complex conjugates while* ω_3 *is a real root of* (46).

$$\begin{aligned}\hat{\mathcal{B}}\omega^3 &+ k^2\left(\hat{\mathcal{T}} + 1 + \hat{\mathcal{B}}\left[\frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 + \frac{k^2}{3} - 1\right]\right)\omega^2 \\ &+ k^4\left(\frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 - 1 + \hat{\mathcal{B}}\left[\frac{1}{\mathrm{Pe}} + \frac{\mathcal{M}\Gamma_0}{4}\right]\left[\frac{k^2}{3} - 1\right]\right. \\ &\left. + \hat{\mathcal{T}}\left[\frac{1}{\mathrm{Pe}} + \mathcal{M}\Gamma_0 + \frac{k^2}{3} - 1\right]\right)\omega - k^6\left(\frac{1}{\mathrm{Pe}} + \frac{\mathcal{M}\Gamma_0}{4}\right)\left(1 - \hat{\mathcal{T}}\left[\frac{k^2}{3} - 1\right]\right) = 0. \end{aligned} \tag{46}$$

Equation (46), which is a third-order polynomial in ω, was solved using standard Mathematica routines for different parameters. Figure 3 shows the three solutions, ω_i $(i = 1, 2, 3)$, as a function of k for a case where all physical parameters are represented: $\hat{\mathcal{B}} = \hat{\mathcal{T}} = \mathcal{M} = \Gamma_0 = 1$ and $\mathrm{Pe} = 100$. In subsequent plots in this subsection only the dominant mode will be shown.

In Figure 4, we show the effect of varying $\hat{\mathcal{B}}$ on the dispersion curves with the rest of the parameter values remaining unaltered from Figure 3. Increasing $\hat{\mathcal{B}}$, which corresponds to an increase in the magnitude of wall damping, results in a decrease of the growth rate and a shift of the most dangerous mode toward smaller wavenumbers. Increasing the relative magnitude of longitudinal tension by increasing $\hat{\mathcal{T}}$ has a stabilizing effect, which is similar to that of increasing $\hat{\mathcal{B}}$; this is shown in Figure 5. We have also investigated the effect of Marangoni stresses on the linear stability characteristics in the presence of wall damping. Our results (not shown) indicate that increasing $\mathcal{M}$ (or, equivalently, Γ_0) and Pe stabilizes the flow in this case as well.

Although linear theory provides insight into the destabilizing mechanisms and a reasonable estimate of the rupture time, this theory breaks down as rupture is approached since the magnitude of the disturbances is no longer negligibly small and nonlinearities become significant. Numerical solution of the fully nonlinear evolution equations is necessary in that case; this is considered in the following section.

4. Numerical simulations. In this section, we present results from numerical simulations of the governing equations. We begin with a concise description of the

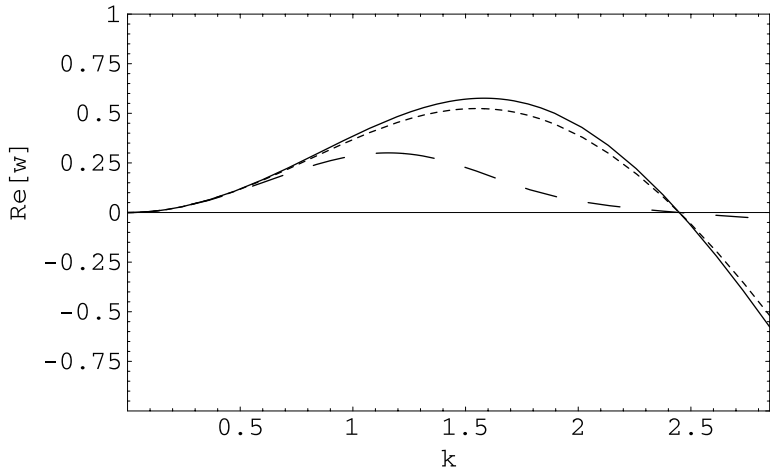


FIG. 4. *The effect of varying* $\hat{\mathcal{B}}$ *on the dispersion curves obtained via the solution of* (46): $\hat{\mathcal{B}} = 0.01$ *(solid line),* $\hat{\mathcal{B}} = 1$ *(short-dashed line), and* $\hat{\mathcal{B}} = 100$ *(dashed line). The rest of the parameter values are the same as those used to generate Figure* 3.

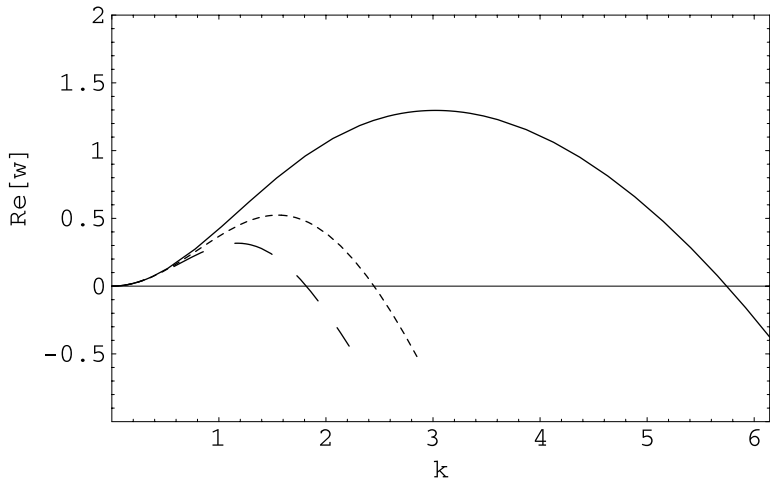


FIG. 5. *The effect of varying* $\hat{\mathcal{T}}$ *on the dispersion curves obtained via the solution of* (46): $\hat{\mathcal{T}} = 0.1$ *(solid line),* $\hat{\mathcal{T}} = 1$ *(short-dashed line), and* $\hat{\mathcal{T}} = 10$ *(dashed line). The rest of the parameter values are the same as those used to generate Figure* 4.

methods employed in the numerical solution of these equations, and then we examine in detail the cases of uncontaminated and surfactant-covered films.

4.1. Numerical procedure. We have used an efficient solver, EPDCOL [38, 39], which has been used for the solution of nonlinear parabolic partial differential equations in related problems involving film rupture and thread breakup [13, 12, 40, 36, 19]. This routine uses finite element collocation to discretize the spatial derivatives and Gear's method in time. Typically 6,000 grid points were used on a computational grid of length $L = 3.6$ dimensionless units. Convergence was achieved on refinement of the spatial mesh by increasing the number of grid points up to 10,000.

Numerical solutions are obtained starting from the initial conditions

$$h(x,0) = \eta(x,0) = A\cos(kx), \qquad \Gamma(x,0) = \Gamma_0, \tag{47}$$

which correspond to a periodic disturbance to $h = 0$ with $A\ \varepsilon\ (10^{-3}, 0.2)$. These solutions are subject to the boundary conditions on h and Γ,

$$h_x = h_{xxx} = 0 \quad \text{at} \quad x = 0, L, \tag{48}$$

$$\Gamma_x = 0 \quad \text{at} \quad x = 0, L, \tag{49}$$

and the conditions on η,

$$\text{either} \qquad \eta_x = 0 \qquad \text{or} \qquad \eta = 0 \quad \text{at} \quad x = 0, L. \tag{50}$$

Physically, the Neumann conditions correspond to a flow which is even about the origin; these conditions, which will give rise to spatially periodic solutions, are appropriate when considering a thin film resting on a wall of infinite lateral extent. The Dirichlet conditions, on the other hand, correspond to a wall of finite lateral extent, fixed at both ends. We shall focus mainly on the former case and examine the latter case only briefly.

We have checked that the numerical solutions are consistent with the predictions of linear theory. This can be confirmed on inspection of Figure 6, in which the natural logarithm of $\delta h(t)$, which is half the difference between the maximal and minimal values of h at time t, is plotted normalized by its initial value, $\delta h(0)$. The parameter values used correspond to the wavenumber associated with the most dangerous mode with all the relevant mechanisms represented: $k = 1.75$, $A = 10^{-3}$, $\hat{\mathcal{B}} = \hat{\mathcal{T}} = \mathcal{M} = \Gamma_0 = 1$, $\mathrm{Pe} = 100$, and the number of grid points used is 6,000. Inspection of Figure 6(a) reveals excellent agreement between the numerical solution and linear theory over the initial stages of the rupture process with deviations occurring at later stages when the amplitude of the perturbations is large and the nonlinearities no longer negligible. These deviations coincide with the onset of rapid thinning of the film, as shown in Figure 6(b), which depicts the evolution of the minimal total film thickness toward rupture for the same parameter values as in Figure 6(a). The wavenumber of the initial perturbation used in the validation studies is also used to generate the results shown in all subsequent figures but with $A = 0.1$ except for Figure 7. Note that the most dangerous mode associated with each set of parameters could have been used. However, to conduct a parametric study in which the effect of various physical mechanisms is investigated, the same initial conditions were used and only the relevant parameter is varied. As in the study of other problems involving film rupture, we have found that the details of the numerical solutions in the initial stages of the breakup process depend on the choice of initial conditions. The solutions as rupture is approached in the vicinity of the rupture region, however, are very weakly dependent on initial conditions.

In the following section, solutions of the evolution equations, (29), (31), and (32), are presented for a wide range of parameter values with particular attention paid to the structure of the film near rupture and the dependence of the estimated rupture time, t_r, on system parameters; here, t_r corresponds to the time at which computations were halted since spatial derivatives could not be resolved accurately. We shall begin by considering the evolution to rupture of a surfactant-covered film. We then explore the effect of varying $\hat{B}$ and $\hat{T}$ on the time to rupture as well as the structure of contaminated and uncontaminated films near rupture. This is then followed by a brief examination of self-similar rupture in the presence of surfactant and wall flexibility.

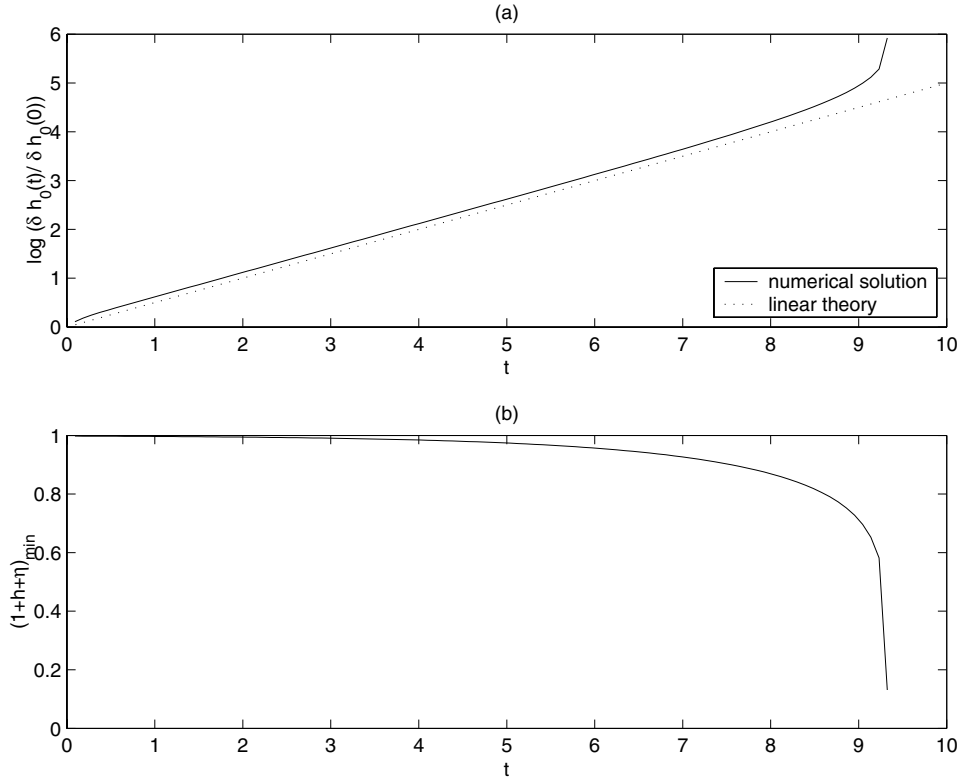


FIG. 6. *Validation of the numerical procedure.* (a) *Comparison of the numerical solutions with the predictions of linear theory;* (b) *variation of the minimum total film thickness,* $(1 + h + \eta)_{\min}$*, with time. The parameter values are* $A = 10^{-3}$*,* $\hat{\mathcal{B}} = \hat{\mathcal{T}} = \mathcal{M} = \Gamma_0 = 1$*, and* $\mathrm{Pe} = 100$*;* 6,000 *grid points were used.*

4.2. Parametric study. We consider first the system in the presence of surfactant and all relevant mechanisms, such as wall damping and longitudinal tension, in order to examine a typical development of the film thickness and surfactant concentration toward rupture; this is shown in Figure 7 with $\hat{\mathcal{B}} = \hat{\mathcal{T}} = \mathcal{M} = \Gamma_0 = 1$ and $\mathrm{Pe} = 100$. The magnitude of van der Waals forces rises beneath the depression in the film, driving fluid away from this region, causing further thinning and, eventually, film rupture. The thinning of the film is transmitted to the underlying flexible wall resulting in its deformation; the wall-liquid interface assumes a similar shape to that of the air-liquid interface. For this choice of parameters, surfactant is also driven away from the rupture region, resulting in the surfactant concentration increasing on either side of the thinning region. This then drives a reverse Marangoni flow that opposes the van der Waals–driven thinning. This flow, however, succeeds only in retarding rather than preventing film rupture. The parametric dependence of the surfactant concentration profile on $\hat{\mathcal{T}}$ and $\hat{\mathcal{B}}$ as rupture is approached will be examined below.

Next, we concentrate on uncontaminated films and study the effect of $\hat{\mathcal{B}}$ and $\hat{\mathcal{T}}$ on the estimated rupture time, t_{r}. Figure 8 is a semilog plot of the dependence of t_{r} on $\hat{\mathcal{T}}$ with $\hat{\mathcal{B}}$ varying parametrically. Inspection of this figure shows clearly that for a fixed value of $\hat{\mathcal{B}}$, t_{r} is largely independent of $\hat{\mathcal{T}}$ for small $\hat{\mathcal{T}}$ values and increases

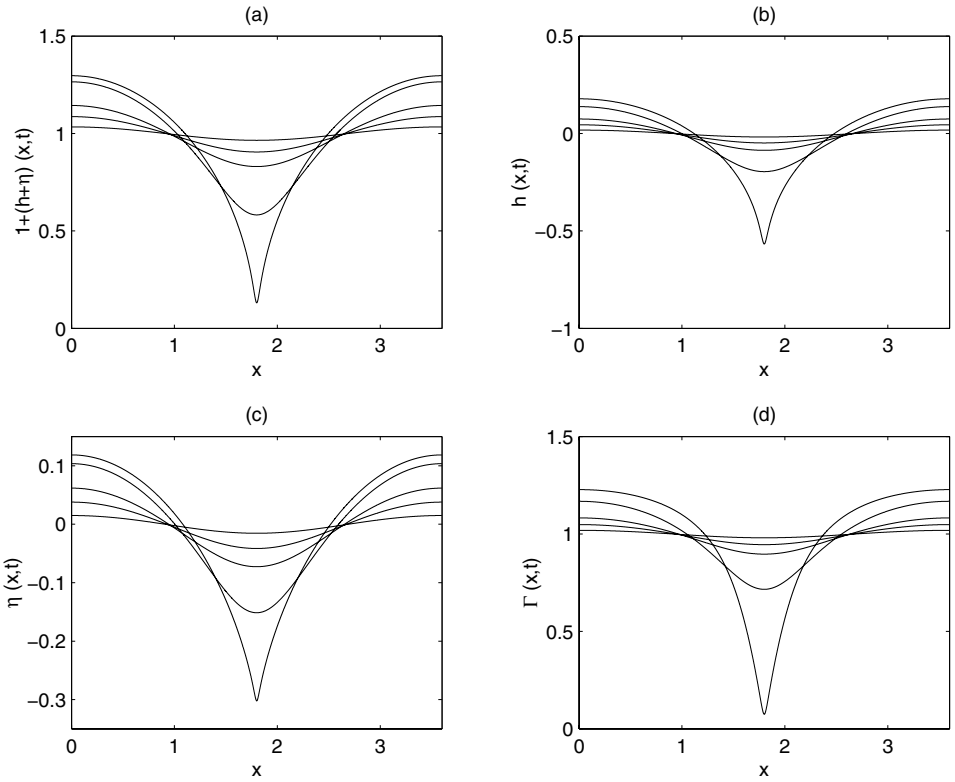


FIG. 7. *Evolution of the total film thickness,* $1+h+\eta$*, shown in* (a)*, the deflections of the air-liquid and liquid-solid interfaces, shown in* (b) *and* (c)*, respectively, and the surfactant concentration,* Γ*, shown in* (d)*, toward rupture. The parameter values are* $A = 10^{-3}$*,* $\hat{\mathcal{B}} = \hat{\mathcal{T}} = \mathcal{M} = \Gamma_0 = 1$*, and* $\mathrm{Pe} = 100$*.*

significantly beyond $\hat{\mathcal{T}} \sim O(1)$. This suggests that increasing the magnitude of the wall longitudinal tension promotes film stability. Furthermore, for a given value of $\hat{\mathcal{T}}$, increasing $\hat{\mathcal{B}}$ results in an increase (decrease) in t_{r} for small (large) $\hat{\mathcal{T}}$ values. Thus, an increase in wall damping exerts a stabilizing influence on the dynamics in the case of weak wall tension and accelerates rupture slightly in the limit of significant wall tension. Note that in the limit of large longitudinal tension and significant wall damping, the rigid wall dynamics are recovered, as will be shown in section 4.3.

We now examine the effect of $\hat{\mathcal{B}}$ and $\hat{\mathcal{T}}$ on the structure of the film as rupture is approached. As can be ascertained on inspection of Figure 9, which shows the variation of the film profile before rupture with $\hat{\mathcal{T}}$ and $\hat{\mathcal{B}}$, small values of $\hat{\mathcal{T}}$ and $\hat{\mathcal{B}}$ give rise to highly localized film rupture accompanied by damped oscillations away from the rupture location. The amplitude of these oscillations, however, is dampened further following an increase in either $\hat{\mathcal{T}}$ or $\hat{\mathcal{B}}$. Thus, increasing the magnitude of either wall damping or longitudinal tension results in localized thinning and eventual rupture. Moreover, comparison of Figure 9(a) and Figure 10(a) reveals that the overall structure of the film thickness profile remains qualitatively very similar despite the absence of surfactant.

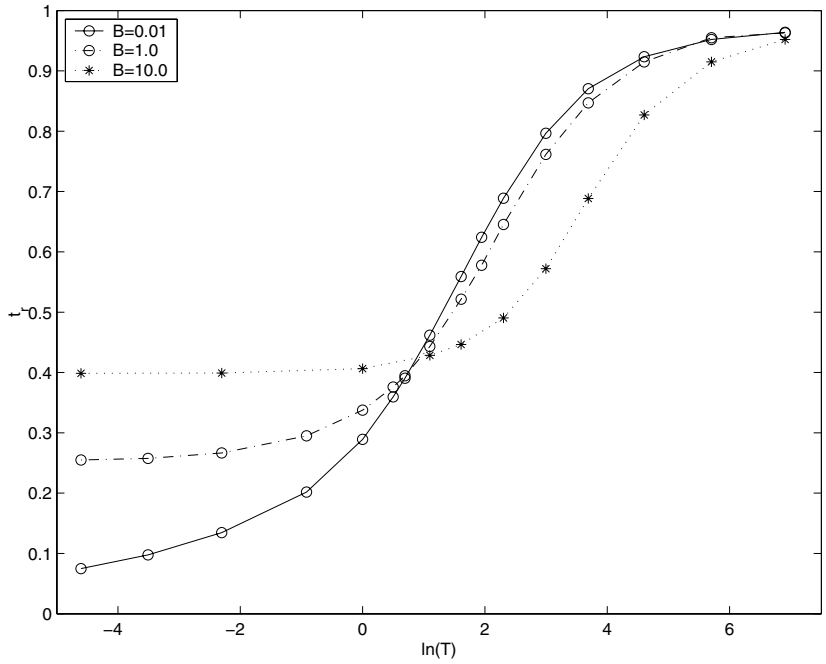


FIG. 8. *Dependence of the estimated rupture time on $\hat{\mathcal{T}}$ with $\hat{\mathcal{B}}$ varying parametrically for the uncontaminated case with $A = 0.1$.*

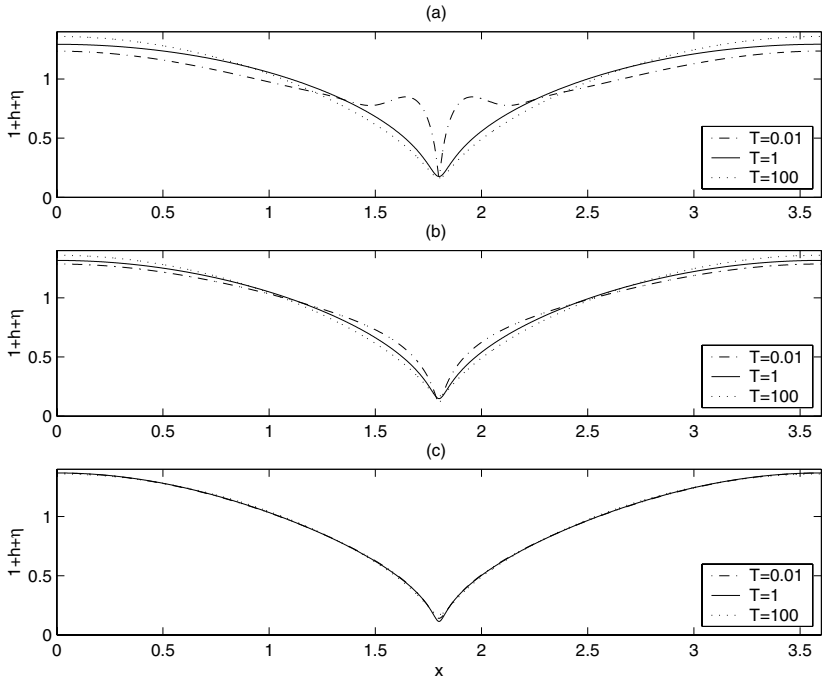


FIG. 9. *Dependence of the total thickness profile of a surfactant-free film before rupture on $\hat{\mathcal{T}}$ with $\hat{\mathcal{B}}$ varying parametrically and $A = 0.1$.* (a) $\hat{\mathcal{B}} = 0.01$, (b) $\hat{\mathcal{B}} = 1$, *and* (c) $\hat{\mathcal{B}} = 10$.

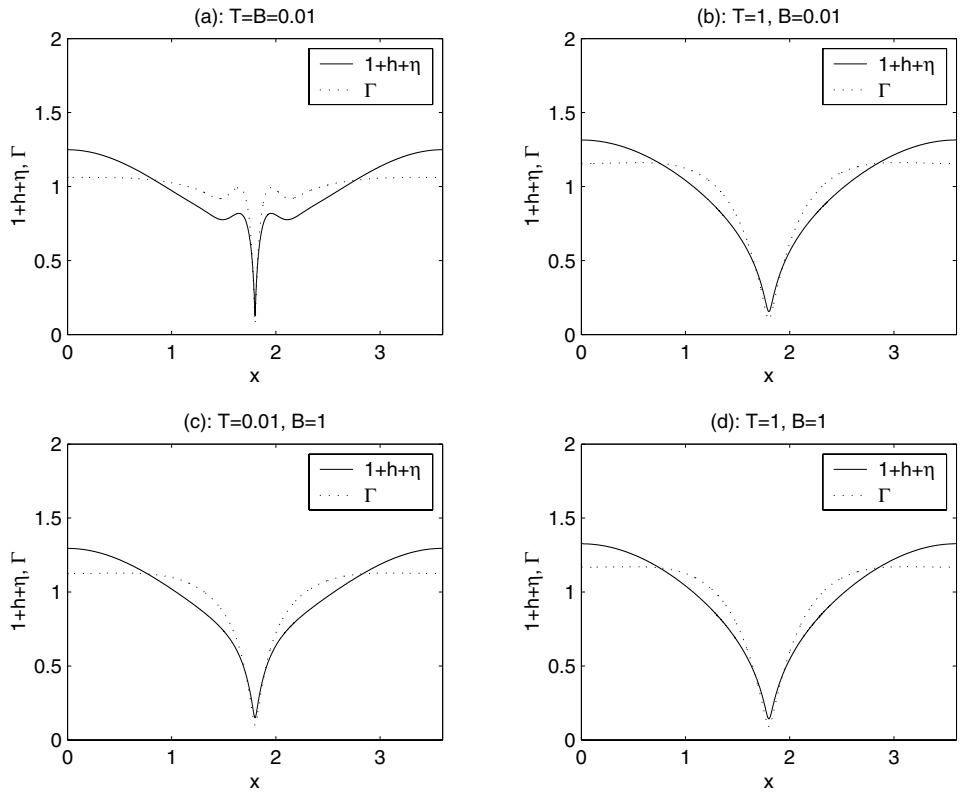


FIG. 10. *Dependence of the total thickness and surfactant concentration profiles before rupture on* $\hat{\mathcal{T}}$ *and* $\hat{\mathcal{B}}$ *with* $A = 0.1$, $\Gamma_0 = 1$, $\mathcal{M} = 1$, *and* $\mathrm{Pe} = 100$.

The effect of changing $\hat{\mathcal{T}}$ and $\hat{\mathcal{B}}$ on the total film thickness and surfactant concentration as rupture is approached is also interesting to explore. In all the cases considered, the behavior of the surfactant concentration, Γ, appears to mimic closely that of the total film thickness, as shown in Figure 10. Van der Waals–driven thinning leads to the advection of fluid and surfactant away from the thinning region, producing a rapid decrease in the thickness and surfactant concentration in that region of very similar rate (Figure 11). An increase in the value of $\hat{\mathcal{T}}$ from 0.01 to 1 while keeping $\hat{\mathcal{B}} = 0.01$ results in considerable damping of the oscillations in the film thickness and surfactant concentration as rupture is approached (see Figures 10(a) and (b)). Finally, either increasing $\hat{\mathcal{B}}$ from 0.01 to 1, while leaving $\hat{\mathcal{T}} = 0.01$ (Figure 10(c)), or having $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 1$ (Figure 10(d)) results in a very similar behavior.

We also explore the effect of $\mathcal{M}$ on the structure of the total film thickness profile in two limits: weak damping and longitudinal tension, and significant damping and longitudinal tension. Figures 12(a) and (b) show that an increase in $\mathcal{M}$ with $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 1$ and $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 0.01$, respectively, $\Gamma_0 = 1$, and $\mathrm{Pe} = 100$, results in retardation of the thinning process. This is due to interface rigidification, which is brought about via a Marangoni-driven reverse flow that counteracts the van der Waals–driven thinning process. A similar effect is brought about via an increase in Γ_0 or Pe (not shown). These findings are in line with the predictions of linear theory.

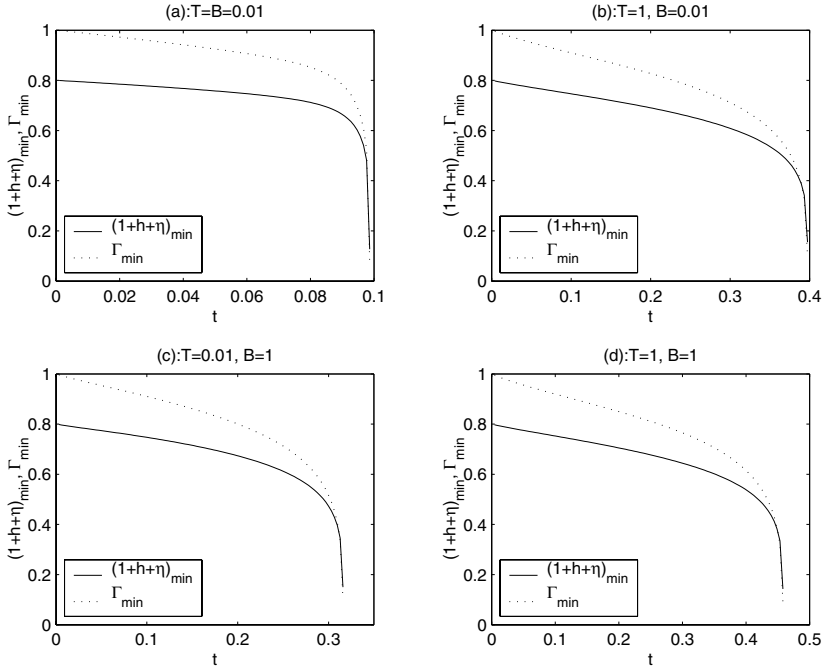


FIG. 11. *Dependence of the total thickness and surfactant concentration at the rupture location on* $\hat{\mathcal{T}}$ *and* $\hat{\mathcal{B}}$*; the rest of the parameter values remain unchanged from Figure* 10.

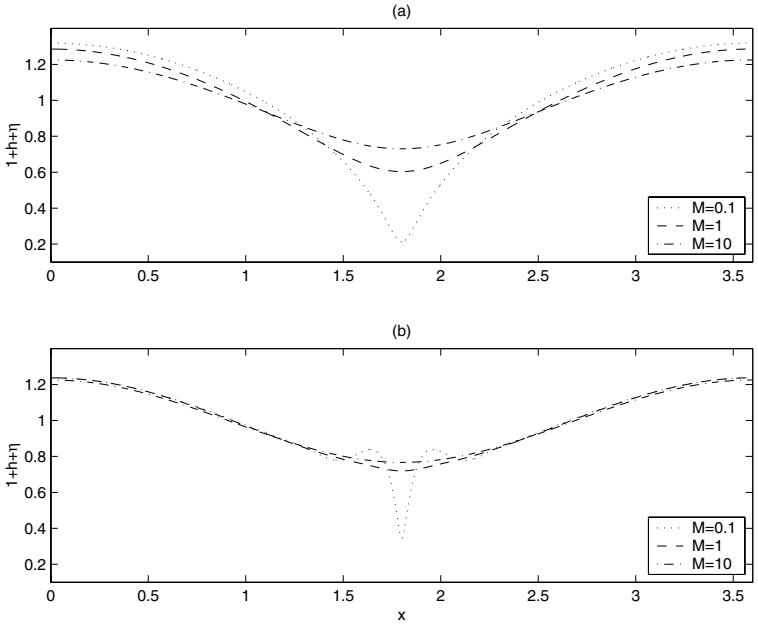


FIG. 12. *Structural dependence of the total film thickness profile on* $\mathcal{M}$ *with* $\Gamma_0 = 1$*, and* $\mathrm{Pe} = 100$. (a) $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 1$ *and* $t = 0.348$*;* (b) $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 0.01$ *and* $t = 0.077$*; the value of* $A = 0.1$ *was used in all cases shown.*

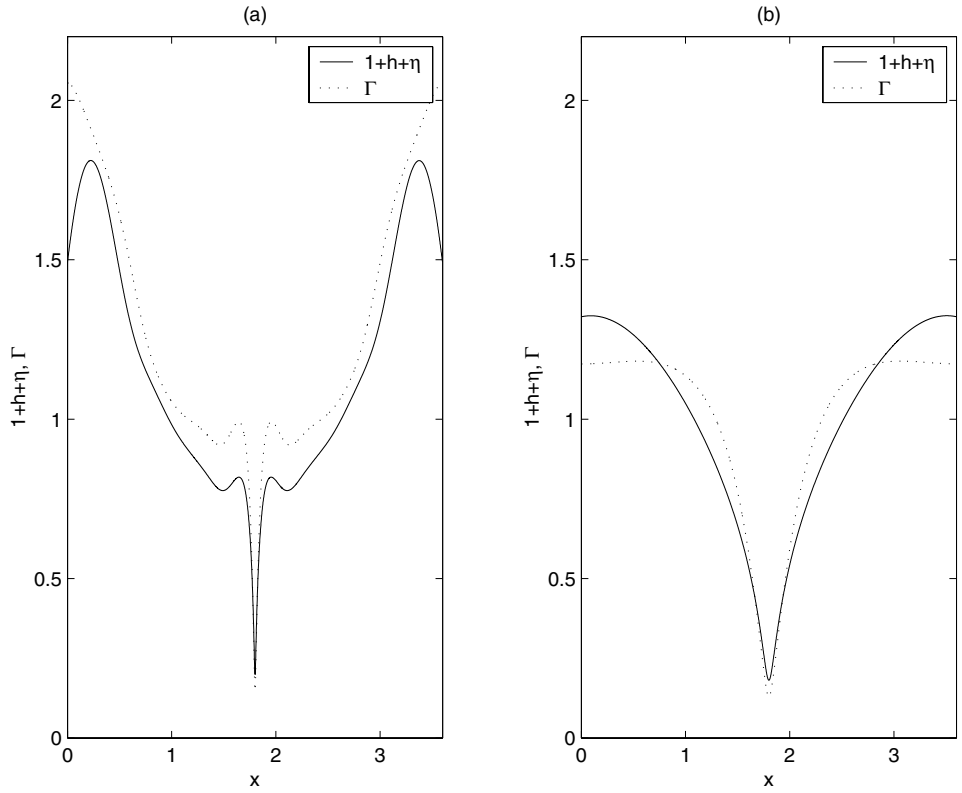


FIG. 13. *Effect of imposing fixed boundary conditions: the total film thickness and surfactant concentration at* $t = 9.869 \times 10^{-2}$ *with* $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 0.01$ *are shown in* (a) *and at* $t = 0.45515$ *with* $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 1$ *in* (b). *The rest of the parameter values are* $A = 0.1$, $\Gamma_0 = 1$, $\mathcal{M} = 1$, *and* $\mathrm{Pe} = 100$.

Finally, we examine briefly the effect of altering the boundary conditions on the dynamics. Figure 13 shows the total film thickness and surfactant concentration before rupture in the limits of weak tension and wall damping ($\hat{\mathcal{T}} = \hat{\mathcal{B}} = 0.01$) in (a) and significant tension and damping ($\hat{\mathcal{T}} = \hat{\mathcal{B}} = 1$) in (b); the rest of the parameter values used are $\Gamma_0 = \mathcal{M} = 1$ and $\mathrm{Pe} = 100$. In the weak tension and damping limit, imposition of fixed boundary conditions on η, $\eta = 0$ at $x = 0$ and L, results in a relatively large buildup of fluid and surfactant near the boundaries while leaving the film thickness and surfactant concentration relatively unaltered in the vicinity of the rupture location. In the limit of significant wall tension and damping, the profiles bear a close resemblance to those shown in Figure 7, which have been generated subject to Neumann boundary conditions in the vicinity of the rupture location. Small discrepancies arise, however, near the edges of the spatial domain due to differences in the imposed boundary conditions.

4.3. Self-similar rupture. Here, we examine the possibility of self-similar rupturing solutions for a thin liquid film resting on a flexible support. Similar studies have been conducted for the case of thin films on a rigid support [9, 10, 11, 18, 13, 12] and free films [18, 19] in the presence [13, 12, 19] and absence [9, 10, 11, 18] of surfactant. The growth rate of a perturbation to the interface increases under the action of van der Waals forces and decreases due to capillarity resulting in a dominant balance,

which yields $H_{xx} \sim 1/H^3$; here, H is the total film thickness. The rate of change of the total film thickness is then proportional to $H/\tau \sim (H/x)^4 \sim (1/x^2)$, from which it then follows that $x \sim \tau^{2/5}$ and $H \sim \tau^{1/5}$, where $\tau = t_r - t$ in which t_r represents the estimated rupture time. This dominant balance between van der Waals, capillary, and viscous forces renders the Marangoni terms in the evolution equations subdominant. As a result, the evolution equations for the surfactant concentration and film thickness become structurally similar. Consistency of these equations then dictates that $\Gamma \sim \tau^{3/10}$, which is similar to previous findings involving studies of thin-film stability on a rigid substrate [12, 13].

We compare the above scaling arguments against scalings extracted directly from the numerical solutions of the evolution equations. In Figure 14 we follow the approach previously adopted in the literature [9, 18, 19] and show log-log plots of the film curvature and the surfactant concentration evaluated at the rupture location against the minimal film thickness; this approach removes the need for an accurate estimate of the rupture time. Here, the slopes of the curvature and concentration curves provide an estimate of the ratio of the self-similar exponent of the film thickness and surfactant concentration, respectively. In all cases considered, the surfactant concentration decreases sharply as rupture is approached, as shown in Figure 11. Inspection of Figure 14(b) reveals that the slopes of the curves are in agreement with the predicted scalings for H and Γ. This, in turn, provides some evidence for consistency of our numerical simulations with the predicted power-law scalings as rupture is approached up to the point where the computations were halted due to the increasingly singular nature of the spatial derivatives.

Finally, we compare our findings in the limit of large wall damping and longitudinal tension with the case of a thin surfactant covered film resting on a rigid substrate [4, 12]. Inspection of Figure 15, which depicts the temporal variation of the minimal total film thickness, the total film thickness and concentration profiles before rupture, and a log-log plot of the curvature and concentration against the minimal thickness with $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 10^3$, $\Gamma_0 = \mathcal{M} = 1$, and $\mathrm{Pe} = 100$, reveals close agreement with the rupture dynamics involving a rigid substrate. In fact, the curves shown in Figure 15 are virtually indistinguishable. These results, which are to be expected, provide a further check on the accuracy of the numerical procedure used to carry out the computations.

The results presented in this section indicate that as rupture is approached, the van der Waals–driven thinning leads to a surfactant-free region. Marangoni stresses become progressively weaker as rupture is approached, leading to a balance between van der Waals, capillary, and viscous forces. Similar results were obtained in related problems involving thin films and slender threads (see, for example, [12, 13, 19, 40]). Furthermore, due to the $1/h^3$ dependence of the van der Waals interactions, it is plausible that these forces would overwhelm wall effects. Thus, it appears that the wall becomes enslaved to the rupturing film, leaving the self-similar scalings unaltered.

5. Conclusions. In this paper, we have investigated the nonlinear evolution and rupture of a thin liquid film covered with insoluble surfactant that rests on a flexible support and is bounded from above by air (inviscid gas). Evolution equations for the deflection of the air-liquid and wall-liquid interfaces and the surfactant concentration were derived using lubrication theory. These equations are parameterized by dimensionless groups which reflect the relative importance of wall to fluid damping, wall longitudinal tension to van der Waals forces, Marangoni stresses to van der Waals forces, and surface diffusion time to convection time.

Both the linear and the nonlinear stability characteristics of these equations have

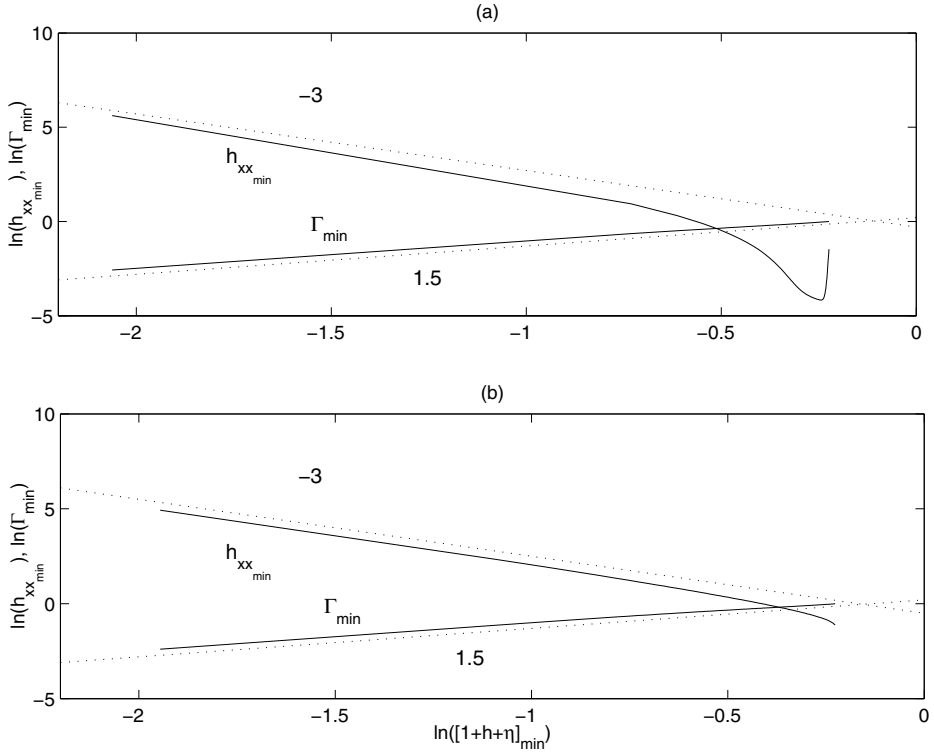


FIG. 14. *Log-log plots of the film curvature and surfactant concentration at the rupture location as a function of the minimal total thickness: parametric dependence of self-similar scalings on* $\hat{\mathcal{T}}$ *and* $\hat{\mathcal{B}}$. (a) $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 0.01$; (b) $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 1$. *The scalings are shown by the dotted lines of constant slope. The rest of the parameter values remain unchanged from Figure* 13.

been investigated. In the linear regime, increasing the level of wall damping exerted a stabilizing influence, as did an increase in the wall longitudinal tension and relative magnitude of Marangoni stresses. It is worth noting that the wavenumber corresponding to the cutoff mode was found to be independent of wall damping and physicochemical parameters, depending only on the longitudinal tension.

In the nonlinear regime, van der Waals forces grow beneath perturbations and drive flow away from that region, giving rise to rapid thinning and driving the film thickness toward rupture. The behavior of the total film thickness and surfactant concentration as rupture is approached was similar in all cases considered: the surfactant concentration decreases in the rupture location as rupture is approached, which drives a Marangoni reverse flow from the adjoining regions of relatively higher concentration to the rupturing region that retards but does not prevent rupture. When both wall longitudinal tension and damping are weak, damped oscillations in the film thickness are observed. For weak longitudinal tension, increasing the damping retards the rupture time, whereas for strong longitudinal tension, the opposite effect in observed.

The self-similar nature of film rupture on a flexible wall was also briefly examined. The total film thickness and surfactant concentration exhibit power-law behavior as rupture is approached with power-law scalings which are consistent with a dominant balance of van der Waals, capillary, and viscous forces. These scalings are identical to those obtained by previous investigators who studied the rupture of thin clean and

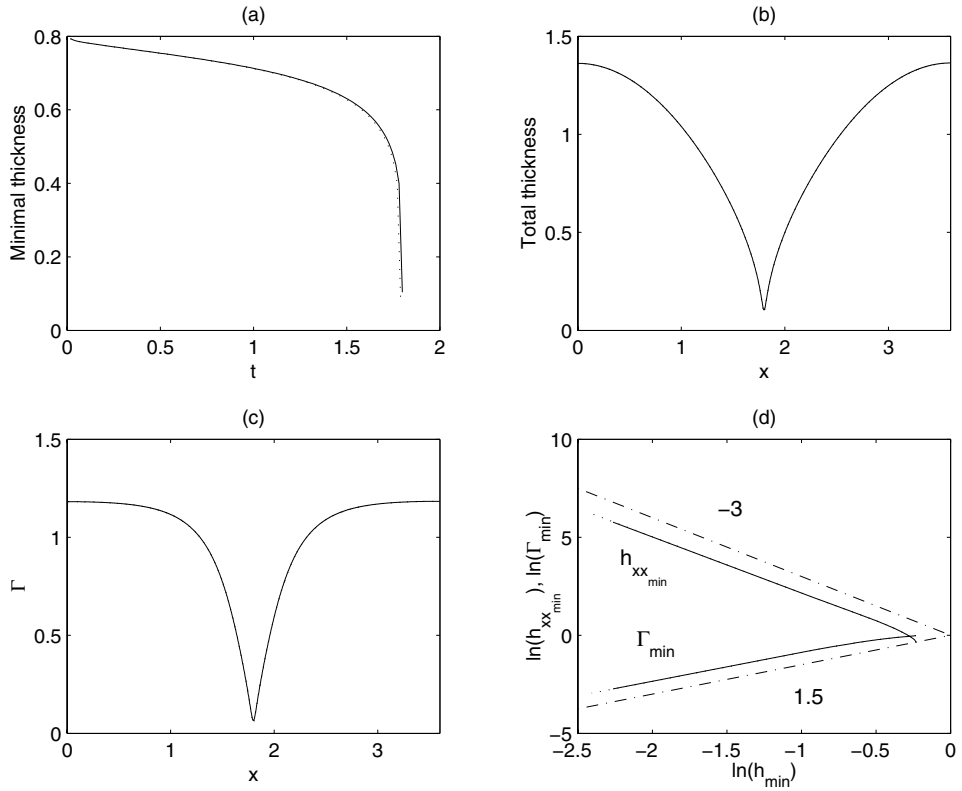


FIG. 15. *Comparison of the rupture dynamics of a thin film resting on a flexible wall using* $\hat{\mathcal{T}} = \hat{\mathcal{B}} = 10^3$ *(dotted lines) with that of the rigid wall case (solid lines).* (a) *Temporal variation of the minimal total film thickness;* (b), (c) *total film thickness and surfactant concentration profiles before rupture at* $t = 1.78929$ *and* $t = 1.79906$ *for the flexible and rigid cases, respectively;* (d) *log-log plots of the total film curvature and surfactant concentration at the rupture location as a function of the minimal total thickness. The relevant scalings are shown by the dot-dashed lines of constant slope, and the rest of the parameter values are* $\Gamma_0 = \mathcal{M} = 1$, $\mathrm{Pe} = 100$, *and* $A = 0.2$.

contaminated liquid films resting on rigid substrates [9, 10, 11, 18, 13, 12]. The reason for this may be related to the $1/h^3$ dependence of the van der Waals forces, which dominates the dynamics as rupture is approached.

Our results may also have consequences for the practical applications that motivated this work. Based on our nonlinear simulations, it appears plausible that film rupture could be used to create patterns on compliant substrates provided that the damping and tension in the substrate are not too strong. It is unclear whether film rupture would promote airway closure in the lungs, given that there is a circumferential tension in cylindrical airways that does not appear in our planar model; this tension may have a significant impact on the problem dynamics. Nevertheless, given that the wall does deform on film rupture, our results raise the question of whether such a deformation will lead to airway damage. To the extent that film rupture is important in cell and vesicle adhesion to solid surfaces, it also appears plausible that adhesion dynamics would be different near a compliant substrate than a rigid one since the former can undergo significant deformation. Of course, verification of such a conjecture will require a more detailed model and corresponding experiments.

Acknowledgment. We thank the referees for a number of helpful comments.

REFERENCES

[1] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Modern Phys., 69 (1997), pp. 931–980.
[2] E. Ruckenstein and R. K. Jain, *Spontaneous rupture of thin liquid films*, Chem. Soc. Farad. Trans., 70 (1974), pp. 132–147.
[3] A. Sharma and E. Ruckenstein, *An analytical nonlinear theory of thin film rupture and its applications to wetting films*, J. Coll. Int. Sci., 113 (1986), pp. 456–479.
[4] O. E. Jensen and J. B. Grotberg, *Insoluble surfactant spreading on a thin viscous film: Shock evolution and film rupture*, J. Fluid Mech., 240 (1992), pp. 259–288.
[5] A. L. Bertozzi, M. P. Brenner, T. F. Dupont, and L. P. Kadanoff, *Singularities and similarities in interface flows*, in Trends and Perspectives in Applied Mathematics, Appl. Math. Sci. 100, Springer-Verlag, New York, 1994, pp. 155–208.
[6] A. L. Bertozzi and M. Pugh, *The lubrication approximation for thin viscous films: The moving contact line with a "porous media" cut-off of van der Waals interactions*, Nonlinearity, 7 (1994), pp. 1535–1564.
[7] A. L. Bertozzi and M. Pugh, *Long-wave instabilities and saturation in thin film equations*, Comm. Pure Appl. Math., 51 (1998), pp. 625–661.
[8] C.-C. Hwang, C.-K. Lin, and W.-Y. Uen, *A nonlinear three-dimensional rupture theory of thin liquid films*, J. Coll. Int. Sci., 190 (1997), pp. 250–252.
[9] W. W. Zhang and J. R. Lister, *Similarity solutions for van der Waals rupture of a thin film on a solid substrate*, Phys. Fluids A, 11 (1999), pp. 2454–2462.
[10] T. P. Witelski and A. J. Bernoff, *Stability of self-similar solutions for van der Waals driven thin film rupture*, Phys. Fluids A, 11 (1999), pp. 2443–2445.
[11] T. P. Witelski and A. J. Bernoff, *Dynamics of three-dimensional thin film rupture*, Phys. D, 147 (2000), pp. 155–176.
[12] M. R. E. Warner, R. V. Craster, and O. K. Matar, *Unstable van der Waals driven line rupture in Marangoni driven thin liquid films*, Phys. Fluids, 14 (2002), pp. 1642–1654.
[13] O. K. Matar, R. V. Craster, and M. R. E. Warner, *Surfactant transport on highly viscous surface films*, J. Fluid Mech., 466 (2002), pp. 85–111.
[14] B. T. Erneux and S. H. Davis, *Nonlinear rupture of free films*, Phys. Fluids A, 5 (1993), pp. 1117–1122.
[15] A. DeWit, D. Gallez, and C. Christov, *Nonlinear evolution equations for thin liquid films with insoluble surfactants*, Phys. Fluids, 6 (1994), pp. 3256–3266.
[16] J. R. Lister and H. A. Stone, *Capillary breakup of a viscous thread surrounded by another viscous fluid*, Phys. Fluids A, 10 (1998), pp. 2758–2764.
[17] S. Naire, R. J. Braun, and S. A. Snow, *Limiting cases of gravitational drainage of a vertical free film for evaluating surfactants*, SIAM J. Appl. Math., 61 (2000), pp. 889–913.
[18] D. Vaynblat, J. R. Lister, and T. P. Witelski, *Rupture of thin viscous films by van der Waals forces: Evolution and self-similarity*, Phys. Fluids A, 13 (2001), pp. 1130–1140.
[19] O. K. Matar, *Nonlinear evolution of a thin free viscous film in the presence of soluble surfactant*, Phys. Fluids, 14 (2002), pp. 4216–4234.
[20] E. Ramos de Souza and D. Gallez, *Pattern formation in thin liquid films with insoluble surfactant*, Phys. Fluids, 10 (1998), pp. 1804–1820.
[21] W. T. Coakley, D. Gallez, E. Ramos de Souza, and H. Gauci, *Ionic strength dependence of localized contact formation between membranes: Nonlinear theory and experiment*, Biophys. J., 77 (1999), pp. 817–828.
[22] E. Ramos de Souza, C. Anteneodo, D. Gallez, and P. M. Bisch, *Long-scale evolution of thin liquid films bounded by a viscous phase with diffusing charged surfactants*, J. Coll. Int. Sci., 244 (2001), pp. 303–312.
[23] J. N. Israelachvili, *Intermolecular and Surface Forces with Applications to Colloidal and Biological Systems*, Academic Press, London, 1985.
[24] J. J. Riley, M. G. el Hak, and R. W. Metcalfe, *Compliant coatings*, Ann. Rev. Fluid Mech., 20 (1988), pp. 393–420.
[25] J. B. Grotberg, *Pulmonary flow and transport phenomena*, Ann. Rev. Fluid Mech., 26 (1994), pp. 529–571.
[26] S. A. Berger and L.-D. Jou, *Flows in stenotic vessels*, Ann. Rev. Fluid Mech., 32 (2000), pp. 347–382.

[27] M. S. Carvalho and L. E. Scriven, *Deformable roll coating flows: Steady state and linear perturbation analysis*, J. Fluid Mech., 339 (1997), pp. 143–172.
[28] V. Kumaran and R. Muralikrishnan, *Spontaneous growth of fluctuations in the viscous flow of a fluid past a soft interface*, Phys. Rev. Lett., 84 (2000), pp. 3310–3313.
[29] D. Halpern and J. B. Grotberg, *Fluid-elastic instabilities of liquid-lined flexible tubes*, J. Fluid Mech., 244 (1992), pp. 615–632.
[30] D. Halpern and J. B. Grotberg, *Surfactant effects on fluid-elastic instabilities of liquid-lined flexible tubes: A model of airway closure*, J. Biomech. Engrg., 115 (1993), pp. 271–277.
[31] D. A. Edwards, H. Brenner, and D. T. Wasan, *Interfacial Transport Processes and Rheology*, Butterworth-Heinemann, New York, 1991.
[32] H. A. Stone, *A simple derivation of the time-dependent convective-diffusion equation for surfactant transport along a deforming interface*, Phys. Fluids A, 2 (1990), pp. 111–112.
[33] H. B. Atabek and S. H. Lew, *Wave propagation through a viscous incompressible fluid contained in an initially stressed elastic tube*, Biophys. J., 6 (1966), pp. 481–503.
[34] L. D. Landau and E. M. Lifshitz, *Theory of Elasticity*, 3rd ed., Butterworth-Heinemann, New York, 1986.
[35] A. C. Goldenveizer, *Theory of Elastic Thin Shells*, Pergamon, New York, 1961.
[36] M. R. E. Warner, R. V. Craster, and O. K. Matar, *Dewetting of surfactant covered ultra thin films*, Phys. Fluids, 14 (2002), pp. 4040–4054.
[37] M. B. Williams and S. H. Davis, *Nonlinear theory of film rupture*, J. Coll. Int. Sci., 90 (1982), pp. 220–228.
[38] R. F. Sincovec and N. K. Madsen, *Algorithm* 540 *PDECOL*, ACM Trans. Math. Software, 5 (1979), pp. 326–351.
[39] P. Keast and P. H. Muir, *Algorithm* 688 *EPDCOL—A more efficient PDECOL code*, ACM Trans. Math. Software, 17 (1991), pp. 153–166.
[40] R. V. Craster, O. K. Matar, and D. T. Papageorgiou, *Pinchoff and satellite formation in surfactant covered viscous threads*, Phys. Fluids, 14 (2002), pp. 1364–1376.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 2167–2185

FORWARD SCATTERING SERIES AND SEISMIC EVENTS: FAR FIELD APPROXIMATIONS, CRITICAL AND POSTCRITICAL EVENTS*

BOGDAN G. NITA†, KENNETH H. MATSON‡, AND ARTHUR B. WEGLEIN†

Abstract. Inverse scattering series is the only nonlinear, direct inversion method for the multidimensional, acoustic or elastic equation. Recently developed techniques for inverse problems based on the inverse scattering series [Weglein et al., *Geophys.*, 62 (1997), pp. 1975–1989; *Top. Rev. Inverse Problems*, 19 (2003), pp. R27–R83] were shown to require two mappings, one associating nonperturbative description of seismic events with their forward scattering series description and a second relating the construction of events in the forward to their treatment in the inverse scattering series. This paper extends and further analyzes the first of these two mappings, introduced, for 1D normal incidence, in Matson [*J. Seismic Exploration*, 5 (1996), pp. 63–78] and later extended to two dimensions in Matson [*An Inverse Scattering Series for Attenuating Elastic Multiples from Multicomponent Land and Ocean Bottom Seismic Data*, Ph.D. thesis, Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC, Canada, 1997]. It brings a new and more rigorous understanding of the mathematics and physics underlying the calculation of terms in the forward scattering series and the events in the seismic model. The convergence of the series for 1D acoustic models is examined, and the earlier precritical analysis is extended to critical and postcritical reflections. An explanation is proposed for the divergence of the series for postcritical incident planewaves.

Key words. scattering theory, forward problem, critical reflections, postcritical reflections

AMS subject classifications. 34L25, 47A40

DOI. 10.1137/S0036139903435619

1. Introduction.

Scattering theory is a form of perturbation theory. In seismic exploration, it relates the propagation of a wave in an actual medium with the propagation of the wave in a reference medium and a perturbation operator which describes the difference between the two media. The forward problem (or forward modeling) is to construct the actual wave-field, given the reference wave-field and the perturbation operator; the inverse problem is to construct the perturbation operator, given the reference wave-field everywhere and the actual wave-field on a measurement surface. The relation between these three quantities is nonlinear and cannot be given, at least so far, in a closed form in either the forward or the inverse problem. This relationship takes the form of a series which, when convergent, constructs the actual wave-field and the perturbation operator.

Inverse scattering series is the only nonlinear, direct inversion method for the multidimensional, acoustic or elastic equation. Early tests on the convergence of the entire series for an acoustic medium by Carvalho [4] were not favorable for real world application. Weglein and collaborators then developed the "subseries method" for the inverse problem (for a description and a complete history, see Weglein et al. [13] and references therein). The overall undertaking of the inverse scattering series was broken up into four tasks, which otherwise would be performed simultaneously by the series acting upon the input data. The four tasks are 1. elimination of the free surface

*Received by the editors September 25, 2003; accepted for publication (in revised form) February 27, 2004; published electronically September 14, 2004.
http://www.siam.org/journals/siap/64-6/43561.html
†Department of Physics, University of Houston, 617 Science and Research Bldg. 1, Houston, TX 77204-5005 (bnita@uh.edu, aweglein@uh.edu).
‡British Petroleum, 200 Westlake Park Blvd., Houston, TX 77079 (matsonkh@bp.com).

multiples, 2. elimination of the internal multiples, 3. locating where rapid changes in the medium properties occur (imaging), and 4. determining the changes at those locations (inversion). These tasks were associated with subseries of the full series, subseries which, if identified, would perform their job as if no other task existed in the series. Two immediate advantages of this separation of tasks are the favorable convergence properties of the subseries and the ability to judge the effectiveness of each step before proceeding on to the next. To facilitate the identification of the task-specific subseries in the inverse series, two maps have to be constructed (see [12]): one map associates seismic events with their forward scattering series description, while the second relates the construction of events in the forward to their usage in the inverse scattering series. In this paper we advance the analysis of the first of these mappings, introduced, for 1D (one-dimensional) normal incidence, by Matson [6] and later extended by Matson to two dimensions [7].

The forward series takes as input the information about the wave-field propagating through the reference medium and about the perturbation operator and outputs the wave-field everywhere in the actual medium. This process can be regarded as creating data (primaries, free surface multiples, internal multiples) for a given model; in practice, the forward series is never used for this purpose due to its inefficiency: it takes an infinite number of terms to create any single event. The events recorded in a seismic experiment are used by the inverse series to find the perturbation and, although the relation between their creation in the forward and their exploitation in the inverse series is not one-to-one, certain analogies could provide useful hints or at least point to where various activities reside in the inverse series. The forward series does not hint at whether events will be signal or noise in the full inverse series; it only suggests where one might look for that answer in the subseries. Take multiples, for example: it turns out that the inverse scattering subseries made of terms that mimic the diagrams for multiples in the forward series is responsible for attenuating/removing such multiples from the data [1].

The forward scattering series models seismic events in a fundamentally different way from conventional nonperturbative theory, where seismic waves propagate through the medium with different velocities and are reflected and transmitted at media boundaries. To construct one event alone, the forward series needs a sequence of terms which can be viewed as a succession of propagations in the reference medium separated by different orders of scattering interactions with a point scatterer; the different terms in the perturbation series correspond to the number of scattering interactions a wave experiences. Even with these differences taken into account, the wave-field output by the forward scattering series has to agree, when the series converges, with the well-known nonperturbative results for any given seismic experiment. Precritical data has been studied by Matson [7], who showed that the expected (from wave-theory) reflected wave-field is constructed by the convergent forward scattering series in a 2D (two-dimensional) experiment. This study brings new understanding about the physical interpretation of these previous results; it also shows that the same forward series converges for critical angles and diverges for postcritical, and an explanation of this divergence is proposed.

The plan for this paper is as follows. In section 2 we present the mathematical description for the forward scattering series for a 3D (three-dimensional) earth, both in operator and nonoperator form; in section 3, following Matson [7], we apply this description to a specific, 2D seismic model, and discuss the convergence of the forward scattering series for that model. Section 4 presents an alternative method for

solving for the terms in the series using saddle point analysis, which, in this setting, is equivalent to far field approximation. Section 5 presents the physical interpretation of the approximations performed in section 4. Section 6 shows the convergence of the forward scattering series for this model at the critical angle, and section 7 proposes an explanation for the divergence of the series for postcritical events. Some conclusions are given in section 8. Although in this paper we mainly focus on application of the scattering theory to seismic exploration, we mention that the same methods and discussions apply to other areas of explorative sciences like medical imaging, whole earth exploration, etc.

2. Forward scattering series. In operator form, the differential equations describing wave propagation in an actual and a reference medium can be written as

$$\mathbf{LG} = -\mathbf{I} \tag{2.1}$$

and

$$\mathbf{L}_0\mathbf{G}_0 = -\mathbf{I}, \tag{2.2}$$

where $\mathbf{L}$, $\mathbf{L}_0$ and $\mathbf{G}$, $\mathbf{G}_0$ are the actual and reference differential and Green's operators, respectively, for a single temporal frequency and $\mathbf{I}$ is the identity operator. The above equations (2.1) and (2.2) assume that the source and receiver signatures have been deconvolved. The perturbation, $\mathbf{V}$, and the scattered field operator, ψ_s, are defined as

$$\mathbf{V} = \mathbf{L} - \mathbf{L}_0, \tag{2.3}$$

$$\psi_{\mathbf{s}} = \mathbf{G}-\mathbf{G}_0. \tag{2.4}$$

The fundamental equation of scattering theory, the Lippmann–Schwinger equation, relates ψ_s, $\mathbf{G}_0$, $\mathbf{V}$, and $\mathbf{G}$ (see, e.g., [10]):

$$\psi_s = \mathbf{G}-\mathbf{G}_0 = \mathbf{G}_0\mathbf{V}\mathbf{G}. \tag{2.5}$$

When $\mathbf{G}$ corresponds to the pressure field in an inhomogeneous acoustic medium, an example of $\mathbf{L}$, $\mathbf{L}_0$, and $\mathbf{V}$ is (see, e.g., [5])

$$\mathbf{L} = \frac{\omega^2}{\kappa} + \nabla\cdot\left(\frac{1}{\rho}\nabla\right), \tag{2.6}$$

$$\mathbf{L}_0 = \frac{\omega^2}{\kappa_0} + \nabla\cdot\left(\frac{1}{\rho_0}\nabla\right), \tag{2.7}$$

and

$$\mathbf{V} = \omega^2\left(\frac{1}{\kappa} - \frac{1}{\kappa_0}\right) + \nabla\cdot\left[\left(\frac{1}{\rho} - \frac{1}{\rho_0}\right)\nabla\right], \tag{2.8}$$

where κ, κ_0, ρ, and ρ_0 are the actual and reference bulk moduli and densities, respectively. If the density is constant $(\rho = \rho_0 = const.)$, the above expressions become

$$\mathbf{L} = \frac{\omega^2}{\kappa}, \tag{2.9}$$

$$\mathbf{L}_0 = \frac{\omega^2}{\kappa_0}, \tag{2.10}$$

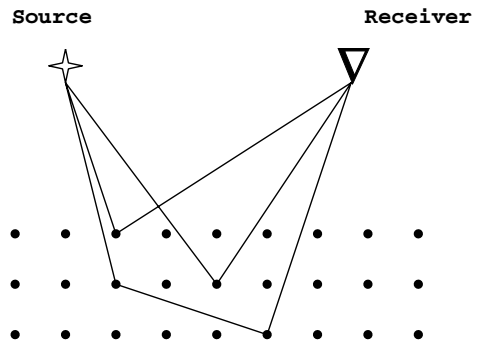


FIG. 2.1. *Graphical representation of the terms in the forward scattering series: the first term is an integral over all* 1*-interaction events, the second term is an integral over all* 2*-interactions events, etc.*

and

$$\mathbf{V} = \omega^2 \left(\frac{1}{\kappa} - \frac{1}{\kappa_0} \right). \tag{2.11}$$

For an elastic isotropic actual and a homogeneous reference medium, the expressions for $\mathbf{L}$, $\mathbf{L}_0$, and $\mathbf{V}$ are different and given, e.g., in [9].

Equation (2.5) can be expanded in an infinite series by repeatedly substituting $\mathbf{G} = \mathbf{G}_0 - \mathbf{G}_0\mathbf{V}\mathbf{G}$ into the right-hand side to obtain

$$\psi_s \equiv \mathbf{G} - \mathbf{G}_0 = \mathbf{G}_0\mathbf{V}\mathbf{G}_0 + \mathbf{G}_0\mathbf{V}\mathbf{G}_0\mathbf{V}\mathbf{G}_0 + \cdots \quad . \tag{2.12}$$

This series constructs the scattered field operator ψ_s as a series of terms formed as propagations in the reference medium ($\mathbf{G}_0$) and interactions with the inhomogeneity ($\mathbf{V}$). Note that the nth term in this series is of order n in the perturbation operator $\mathbf{V}$ and, in fact, can be written as $(\psi_s)_n \equiv \mathbf{G}_0 (\mathbf{V}\mathbf{G}_0)^n$.

For the previous example (constant density case), define $k_0 = \frac{\omega}{c_0}$ and $\alpha = \left(1 - \frac{c_1^2}{c_0^2}\right)$, where c_1 and c_0 are the actual and the reference medium velocities, respectively; the series becomes

$$\begin{aligned} \psi_s\left(\mathbf{r}_g|\mathbf{r}_s;\omega\right) = & \int_{\mathbf{V}} G_0\left(\mathbf{r}_g|\mathbf{r}';\omega\right) k_0^2 \alpha\left(\mathbf{r}'\right) G_0\left(\mathbf{r}'|\mathbf{r}_s;\omega\right) d\mathbf{r}' \\ & + \int_{\mathbf{V}} G_0\left(\mathbf{r}_g|\mathbf{r}';\omega\right) k_0^2 \alpha\left(\mathbf{r}'\right) \int_{\mathbf{V}} G_0\left(\mathbf{r}'|\mathbf{r}'';\omega\right) k_0^2 \alpha\left(\mathbf{r}''\right) G_0\left(\mathbf{r}''|\mathbf{r}_s;\omega\right) d\mathbf{r}'' d\mathbf{r}' \\ & + \cdots, \end{aligned} \tag{2.13}$$

where the integrals are 3D volume integrals taken over the inhomogeneity $\mathbf{V}$. For an easy physical interpretation of this series, consider the perturbation $\mathbf{V}$ to be composed of point scatterers separated by the reference medium. The first term in the series for the scattered field (2.13) represents a summation over all 1-interaction events, i.e., events formed from a wave propagating from the source location $\mathbf{r}_s$ to the scatterer location at $\mathbf{r}'$, $G_0\left(\mathbf{r}'|\mathbf{r}_s;\omega\right)$, interacting with the scatterer at $\mathbf{r}'$, $k_0^2\alpha\left(\mathbf{r}'\right)$, and propagating to the receiver location at $\mathbf{r}_g$, $G_0(\mathbf{r}_g|\mathbf{r}';\omega)$. The second term represents a summation over all 2-interaction events and so on. Note that, as stated before, the propagations between source, receiver, and scatterers occur only in the reference medium, i.e., with the Green's function G_0, even though the speed of the wave in the actual medium is different from the speed of the wave in the reference medium. A picture of the physical interpretation of these terms is shown in Figure 2.1.

3. A 2D seismic profile. Matson [6, 7] describes the propagation of a wave-field in a given 1D or 2D medium, using the forward scattering series. We use the same 2D model in this paper to give an alternate derivation, and a physical interpretation for that derivation, for the Matson [7] result. The model is a half-space earth with no lateral variance and an interface at z_1; the scattering perturbation for this model is, therefore,

$$V(z') = k_0^2 \alpha H(z' - z_1), \tag{3.1}$$

where, as before, $\alpha = 1 - c_0^2/c_1^2$, c_1 is the velocity in the second medium, c_0 is the velocity in the reference medium, and H is the Heaviside step function.

The propagations in the reference medium are described by the 2D Green's function (see, e.g., [2])

$$G_0(x_g, z_g|x_s, z_s; \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{ik_s(x_g - x_s)} e^{i\nu_{0s}|z_g - z_s|}}{2i\nu_{0s}} dk_s, \tag{3.2}$$

where k_s and ν_{0s} are the horizontal and the vertical wavenumber, respectively, of the reference medium ($\nu_{0s}^2 + k_s^2 = \omega^2/c_0^2$). Rewriting G_0 as

$$G_0(x_g, z_g|x_s, z_s; \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ik_s x_s}}{2i\nu_{0s}} \phi_0(x_g, z_g|k_s, z_s; \omega) dk_s \tag{3.3}$$

with $\phi_0(x_g, z_g|k_s, z_s; \omega) = e^{i(k_s x_g + \nu_{0s}|z_g - z_s|)}$, it is apparent that G_0 represents a superposition of weighted planewaves. This motivates the use of a planewave component as the incident wave with the remark that one can construct solutions for point sources from planewave solutions by performing the above-mentioned weighted integration. Denote by P the actual wave-field and by P_0, P_1, etc., the corresponding term in the forward scattering series. For simplicity consider the source location to be (0, 0); the Born series takes the form

$$\begin{aligned} P(x_g, z_g|k; \omega) = {} & e^{i(kx_g + \nu_0 z_g)} \\ & + \int_{z_1}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{ik_g(x_g - x')} e^{i\nu_{0g}|z_g - z'|}}{2i\nu_0} dk_g k_0^2 \alpha P_0(x', z'|k; \omega) dx' dz' \\ & + \int_{z_1}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{ik_g(x_g - x')} e^{i\nu_{0g}|z_g - z'|}}{2i\nu_0} dk_g k_0^2 \alpha P_1(x', z' > z_1|k; \omega) dx' dz' \\ & + \cdots . \end{aligned} \tag{3.4}$$

Note that the incoming wave hits all the scatterers at once; each scatterer then emits a cylindrical wave which propagates to the receiver or to another scatterer. Each term in the forward series represents the response, at the receiver, after a certain number of interactions: the zeroth term represents the direct arrival, the first term represents the wave-field after one interaction with a point scatterer, and so on. To construct even the simplest event, one needs an infinite number of terms in the forward series. To obtain the total wave-field at the receiver we have to solve the integrals in the previous expression. Following Matson [7], we solve for the first term in the series

$$P_1(x_g, z_g|k; \omega) = \int_{z_1}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{ik_g(x_g - x')} e^{i\nu_{0g}|z_g - z'|}}{2i\nu_0} dk_g k_0^2 \alpha e^{i(kx' + \nu_0 z')} dx' dz'. \tag{3.5}$$

Begin by switching the order of integration so that the integration with respect to dx' is performed first. Hence

(3.6)
$$P_1(x_g, z_g|k;\omega) = \frac{1}{2\pi}\int_{z_1}^{\infty}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} e^{i(k-k_g)x'}dx'\right)e^{ikgx_g}e^{i\nu_{0g}|z_g-z'|}e^{i\nu_0 z'}\frac{k_0^2\alpha}{2i\nu_{0g}}dk_g dz'.$$

Using

$$\int_{-\infty}^{\infty} e^{i(k-k_g)x'}dx' = 2\pi\delta(k_g-k), \tag{3.7}$$

P_1 becomes

$$P_1(x_g, z_g|k;\omega) = \int_{z_1}^{\infty}\int_{-\infty}^{\infty}\delta(k_g-k)e^{ikgx_g}e^{i\nu_{0g}|z_g-z'|}e^{i\nu_0 z'}\frac{k_0^2\alpha}{2i\nu_{0g}}dk_g dz'. \tag{3.8}$$

Using the properties of the delta function, we see that the inside integral switches $k_g \to k$ and hence $\nu_{0g} \to \nu_0$, and so the expression becomes

$$P_1(x_g, z_g|k;\omega) = \frac{k_0^2\alpha}{2i\nu_0}e^{ikx_g}\int_{z_1}^{\infty} e^{i\nu_0|z_g-z'|}e^{i\nu_0 z'}dz'. \tag{3.9}$$

There are two cases to be considered at this point: $z_g < z_1$ for the reflected P_1 and $z_g > z_1$ for the transmitted part. The first enters into the series for the total reflected field, while the second is used either in the series for transmitted wave-field or for the calculation of P_2 (reflected or transmitted). We have

$$P_1(x_g, z_g < z_1|k;\omega) = \frac{k_0^2\alpha}{2i\nu_0}e^{ikx_g}e^{-i\nu_0 z_g}\int_{z_1}^{\infty} e^{i\nu_0 2z'}dz'. \tag{3.10}$$

The last integral,

$$\int_{z_1}^{\infty} e^{i\nu_0 2z'}dz', \tag{3.11}$$

is not defined in the Riemannian sense because the integrand oscillates, preserving its amplitude towards infinity. We are going to define this integral to be the value of the antiderivative of the integrand calculated at its finite boundary z_1, i.e.,

$$\int_{z_1}^{\infty} e^{i\nu_0 2z'}dz' = -\frac{e^{i\nu_0 2z_1}}{2i\nu_0}. \tag{3.12}$$

This definition is consistent with considering that the reference medium is attenuating the wave-field which will vanish at infinity. The attenuation is introduced in the equations through an imaginary part in the velocity c_0 (see [2, Chapter 5, equations 5.87 and 5.88]) so that the new velocity c_0^{new} is

$$\frac{1}{c_0^{new}} = \frac{1}{c_0} + i\varepsilon,$$

with ε being a small parameter such that $\varepsilon > 0$ for $\omega > 0$. It is easy to see that, with this new effective velocity, the value of the integral is indeed the one defined in (3.12).

The final expression for P_1 is hence

$$P_1(x_g, z_g < z_1|k;\omega) = \frac{k_0^2\alpha}{4\nu_0^2} e^{ikx_g} e^{i\nu_0(2z_1 - z_g)}. \tag{3.13}$$

The same integration procedure is used for the calculation of P_2, P_3, etc. The calculated series for the scattered field (also denoted by P) is

$$P(x_g, z_g < z_1|k;\omega) = e^{ikx_g} e^{i\nu_0(2z_1 - z_g)} \left[\frac{1}{4}\frac{k_0^2\alpha}{\nu_0^2} + \frac{1}{8}\left(\frac{k_0^2\alpha}{\nu_0^2}\right)^2 + \frac{5}{64}\left(\frac{k_0^2\alpha}{\nu_0^2}\right)^3 + \cdots \right] \tag{3.14}$$

and indicates a certain regularity after some algebraic operations: the series is recognized to be the Taylor series of $\sqrt{1 - \frac{k_0^2\alpha}{\nu_0^2}}$ about $\frac{k_0^2\alpha}{\nu_0^2} = 0$ (a rigorous proof is given in the appendix). The ratio test indicates that the series converges for $\left|\frac{k_0^2\alpha}{\nu_0^2}\right| < 1$. By writing $\nu_0 = k_0 \cos\theta$, with θ being the incidence angle of the incoming planewave, this condition becomes

$$\sin\theta < \frac{c_0}{c_1} < \left(1 + \cos^2\theta\right)^{1/2}. \tag{3.15}$$

This last relation can be viewed in the following two ways:

1. First, for a fixed incidence angle θ, this is a restriction on the velocity contrast between the reference and the actual medium. In particular, for $\theta = 0$ (normal incidence) the left inequality is satisfied for any two velocities; the right inequality becomes $c_0 < \sqrt{2}c_1$, a result obtained in Matson [6].
2. Second, for a fixed velocity model, the restriction is on the incident angle. Note that, given any two velocities c_0 and c_1, one of the two inequalities is automatically satisfied. For $c_0 > c_1$, the condition reads $\frac{c_0}{c_1} < (1 + \cos^2\theta)^{1/2}$ or $\sin^2\theta < 1 + \alpha$ with $\alpha < 0$.

For $c_0 < c_1$, the condition becomes $\sin\theta < \frac{c_0}{c_1}$ or $\theta < \theta_c$, where θ_c is the critical angle $\theta_c = \sin^{-1}(c_0/c_1)$. When the series converges, the limit is

$$2\frac{\nu_0^2}{k_0^2\alpha}\left[1 - \sqrt{1 - \frac{k_0^2\alpha}{\nu_0^2}}\right] - 1 = \frac{\nu_0 - \nu_1}{\nu_0 + \nu_1}, \tag{3.16}$$

and so the final expression for the reflected field is

$$P(x_g, z_g < z_1|k;\omega) = \frac{\nu_0 - \nu_1}{\nu_0 + \nu_1} e^{ikx_g} e^{i\nu_0(2z_1 - z_g)}, \tag{3.17}$$

which is the expected result from nonperturbative theory (see, e.g., [2]).

4. An alternative derivation using saddle point approximations. The calculation of

$$P_1(x_g, z_g|k;\omega) = \int_{z_1}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi}\left(\int_{-\infty}^{\infty} \frac{e^{ik_g(x_g - x')} e^{i\nu_{0g}|z_g - z'|}}{2i\nu_0} dk_g\right) k_0^2\alpha e^{i(kx' + \nu_0 z')} dx'dz' \tag{4.1}$$

contains a reordering of integrals: in the original expression the dk_g integral should be solved first, then the dx', and finally the dz' integral. As we saw in the previous

section, the calculations are greatly simplified if the dx' integration is performed first, then the dk_g, and finally the dz' integration. However, this kind of operation has to be performed with great care since it might impose some restrictions, which might change the result obtained from solving the integrals in the original order.

The theorem which deals with interchanging integrals is Fubini's theorem. It states that when a function f is integrable on $R^n = R^k \times R^m$, the iterated integrals of f over R^k and R^m exist and

$$\int_{R^n} f = \int_{R^k}\int_{R^m} f(x,y)dydx = \int_{R^m}\int_{R^k} f(x,y)dxdy. \tag{4.2}$$

The theorem gives sufficient conditions for interchanging the order of integrals, but those conditions are not necessary. For example, you can have a function non-integrable over R^n for which the integration in both directions would yield the same result. The only way to show that the interchange of integrals does not hold is to calculate the integrals in both direction and obtain different results. However, to calculate the dk_g integral first in the expression (4.1) means to find a closed form for the Green's function (3.2), which is not possible. For an in-depth analysis of the cylindrical functions, see [11].

In this section we show that the interchange of integrals yields the same result as the far field approximation of the integrals in question. The Fubini theorem does not apply here because the function to be doubly integrated is not integrable. To be more specific, the integral representation of the Dirac delta function,

$$\int_{-\infty}^{\infty} e^{i(k-k_g)x'}dx', \tag{4.3}$$

is meaningless in the strict Riemannian sense.

Recalculate P_1 using saddle point approximations for the two integrals involved without switching the order of integration, and show that the result is the one obtained in Matson [7]. Saddle point or stationary phase approximation gives the leading asymptotic behavior of generalized Fourier integrals, i.e., of the form $\int_{-\infty}^{\infty} F(p)e^{\omega f(p)}dp$, having stationary points, i.e., points p_s such that $f'(p_s) = 0$. The idea of the method is to use the analyticity of the integrand to justify deforming the path of integration to a new path on which $f(p)$ has a constant imaginary path. How the contour is deformed depends on the singularities and branch cuts of the integrand. Once this has been done, the integral may be found asymptotically ($\omega \to \infty$) to be

$$\int_{-\infty}^{\infty} F(p)e^{\omega f(p)}dp \ \sim \left|\frac{2\pi}{\omega f''(p_s)}\right|^{1/2} F(p_s)e^{isign\left(f''(p_s)\right)\frac{\pi}{4}} \exp\left[\omega f(p_s)\right]. \tag{4.4}$$

To calculate P_1 in (4.1), start by rewriting

$$G_0 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{ik_g(x_g-x')}e^{i\nu_{0g}\left|z_g-z'\right|}}{2i\nu_0}dk_g \tag{4.5}$$

as

$$G_0 = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(p)e^{\omega f(p)}dp, \tag{4.6}$$

where

$$F(p) = \frac{1}{2i\sqrt{1/c_0^2 - p^2}} \tag{4.7}$$

and

$$f(p) = i\left[p(x_g - x') + |z_g - z'|\sqrt{\frac{1}{c_0^2} - p^2}\right], \tag{4.8}$$

p being the horizontal slowness $p = \frac{k_g}{\omega}$. Note that, due to the square root, $F(p)$ defines two branch cuts in the complex p plane; the branch cuts are hyperbolas in the first and third quadrant and are running very close to the coordinate axis. (For a full discussion of the branch cuts of F, see [2, Box 6.2].) By definition, branch cuts are lines of discontinuities for $F(p)$ and here are given by $Im\sqrt{1/c_0^2 - p^2} = 0$. This means that when the new integration path (see Figure 6.6 in [2]) intersects these branch cuts, $F(p)$ is discontinuous and hence not analytic. This apparent problem can be avoided if we relax the condition $Im\sqrt{1/c_0^2 - p^2} \geq 0$ along the integration path. Instead we allow $Im\sqrt{1/c_0^2 - p^2}$ to change sign at each branch cut intersection which, for the integration path, is equivalent to a transition to a different Riemann sheet. The integrand looses physical interpretation while on another Riemann sheet but gains analyticity. However, the two intersections with the branch cut insure two sign changes and the emergence of the integrand with the correct sign at the saddle point. (Eventually the integrand is going to be expanded in a Taylor series at that point, and the rest of the path is going to be discarded.) To calculate the location of the saddle point, equate the derivative of f with zero; this gives

$$p_s = \frac{x_g - x'}{c_0 d'}, \tag{4.9}$$

with $d' = \sqrt{(z_g - z')^2 + (x_g - x')^2}$. Calculate

$$f(p_s) = i\frac{d'}{c_0}, \tag{4.10}$$

$$f''(p_s) = -\frac{ic_0 d'^3}{|z_g - z'|^2}, \tag{4.11}$$

$$F(p_s) = \frac{c_0 d'}{2i\,|z_g - z'|}, \tag{4.12}$$

and plug them into the above formula (4.4) to obtain

$$G_0 \sim \frac{1}{4\pi i}\left(\frac{2\pi c_0}{i\omega d'}\right)^{1/2} e^{ik_0 d'}. \tag{4.13}$$

(Compare with the approximation for $i\pi H_0^{(1)}(\omega/c_0 d')$, the Green's function for the 2D Helmholtz equation, where $H_0^{(1)}$ is the Hankel function of the first kind, given by formula (5.3.69) in [8].) With this approximation, expression (4.1) for P_1 becomes

$$P_1(x_g, z_g|k;\omega) = \frac{1}{4\pi i}\int_{z_1}^{\infty}\int_{-\infty}^{\infty} e^{ik_0 d'}\left(\frac{2\pi c_0}{i\omega d'}\right)^{1/2} k_0^2 \alpha e^{i(kx' + \nu_0 z')} dx' dz' \tag{4.14}$$

or

$$P_1(x_g, z_g|k;\omega) = \frac{k_0^{3/2}\alpha}{2\pi i}\sqrt{\frac{\pi}{2i}}\int_{z_1}^{\infty} e^{i\nu_0 z'}\int_{-\infty}^{\infty}\frac{e^{i\omega\left(\frac{d'}{c_0}+\frac{k}{\omega}x'\right)}}{\sqrt{d'}}dx'dz'. \tag{4.15}$$

Again, the innermost integral has the form

$$I = \int_{-\infty}^{\infty} F(x')e^{\omega f(x')}dx' \tag{4.16}$$

with $F(x') = \frac{1}{\sqrt{d'}}$ and $f(x') = i(\frac{d'}{c_0} + \frac{k}{\omega}x')$. Note that the integrand has no branch cuts this time since $d' = \sqrt{(z_g - z')^2 + (x_g - x')^2}$ is always positive; the saddle point is x'_s such that

$$x_g - x'_s = |z_g - z'|\,\frac{k}{\nu_0}, \tag{4.17}$$

and so we have

$$f(x'_s) = i\left(\frac{\nu_0}{\omega}\,|z_g - z'| + \frac{k}{\omega}x_g\right), \tag{4.18}$$

$$f''(x'_s) = \frac{ic_0^2\nu_0^3}{\omega^3\,|z_g - z'|}, \tag{4.19}$$

and

$$F(x'_s) = \frac{1}{\sqrt{|z_g - z'|}}\sqrt{\frac{c_0\nu_0}{\omega}}. \tag{4.20}$$

Using the same high frequency approximation (4.4), we find

$$\int_{-\infty}^{\infty}\frac{e^{i\omega\left(\frac{d'}{c_0}+\frac{k}{\omega}x'\right)}}{\sqrt{d'}}dx' \sim \frac{1}{\nu_0}\sqrt{\frac{2\pi i\omega}{c_0}}e^{i\left(\nu_0|z_g - z'| + kx_g\right)}. \tag{4.21}$$

Substituting this into the expression (4.15) for P_1, we obtain

$$P_1(x_g, z_g|k;\omega) = \frac{k_0^2\alpha}{2i\nu_0}e^{ikx_g}\int_{z_1}^{\infty} e^{i\nu_0|z_g - z'|}e^{i\nu_0 z'}dz', \tag{4.22}$$

which is the same result as that obtained before by switching the order of integration. The rest of the terms in the series for P can be similarly shown to resemble the expressions given by Matson [7].

5. Physical interpretation of the approximations. The two far field approximations performed in the previous derivation have an easily understandable physical interpretation. The approximation of the first integral in the expression of P_1 represents the most important contribution arriving at the receiver from each point scatterer (see Figure 5.1).

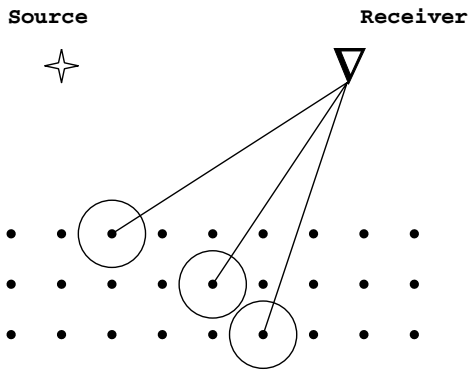


FIG. 5.1. *The physical interpretation of the approximation of the first integral in the calculation of* P_1.

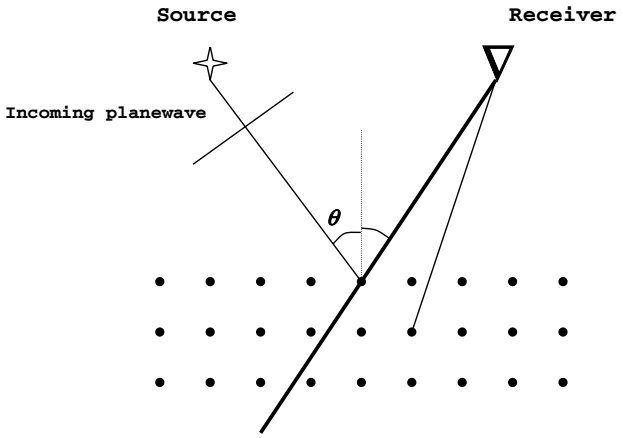


FIG. 5.2. *The physical interpretation of the approximation of the second integral in the calculation of* P_1.

As the figure shows, each scatterer behaves as a point source producing a wave propagating in all directions described by the Green's function given by (3.2). However, when the integral is approximated using saddle point techniques, only the direction of propagation bringing in the highest contribution is kept. The result given by (4.13),

$$G_0 \sim \frac{1}{4\pi i}\left(\frac{2\pi c_0}{i\omega d'}\right)^{1/2} e^{ik_0 d'}, \tag{5.1}$$

represents the part arriving from the scatterer to the receiver along the straight line connecting them, multiplied by a coefficient which accounts for the dismissal of all the other directions.

The approximation of the second integral in the expression of P_1 picks out the most important contribution arriving at the receiver from the totality of incoming rays. Here, the main contribution is found to be the one from the rays that make an angle equal to the incident's planewave angle with the vertical (see Figure 5.2); this

can be seen from the expression of the saddle point for the x' integration:

$$x_g - x'_s = |z_g - z'| \frac{k'}{\nu_0}. \tag{5.2}$$

The last integral in the expression of P_1 is a 1D integral along the thick line shown in Figure 5.2. Even though the parameter of integration is z', there is a certain relation between z' and x', given by (5.2), such that the direction of integration is tilted at an angle equal to the incident angle rather than vertical. The lack of symmetry in this last integral is expected since the model is not symmetric: the discussion here is for a planewave component of a line source and a line receiver. It is anticipated that the symmetry would be recovered in the line source–line receiver case.

6. Convergence at the critical angle. The forward scattering series for the reflected wave-field for the model discussed in this paper is (see Matson [7])

$$P(x_g, z_g < z_1|k;\omega) = e^{ikx_g} e^{i\nu_0(2z_1 - z_g)} \left[\frac{1}{4}\frac{k_0^2\alpha}{\nu_0^2} + \frac{1}{8}\left(\frac{k_0^2\alpha}{\nu_0^2}\right)^2 + \frac{5}{64}\left(\frac{k_0^2\alpha}{\nu_0^2}\right)^3 + \cdots \right]. \tag{6.1}$$

The ratio test shows convergence for $|\frac{k_0^2\alpha}{\nu_0^2}| < 1$, divergence for $|\frac{k_0^2\alpha}{\nu_0^2}| > 1$, and is inconclusive for $|\frac{k_0^2\alpha}{\nu_0^2}| = 1$. When $c_0 < c_1$, this last condition is equivalent to $\frac{k_0^2\alpha}{\nu_0^2} = 1$, which in turn is equivalent to $\theta = \theta_c$; i.e., the incident angle is the critical angle. In other words, the forward series is convergent for precritical incidence and divergent for postcritical incidence; no information is found about the critical incidence. For a critical incident planewave, the series becomes

$$P(x_g, z_g < z_1|k;\omega) = e^{ikx_g} e^{i\nu_0(2z_1 - z_g)} \left[\frac{1}{4} + \frac{1}{8} + \frac{5}{64} + \frac{7}{128} + \cdots \right]. \tag{6.2}$$

Rewrite

$$R = \frac{1}{4} + \frac{1}{8} + \frac{5}{64} + \frac{7}{128} + \cdots = \sum_{n=2}^{\infty} \frac{1}{n!} \frac{1\cdot 3\cdot 5 \ldots (2n-3)}{2^{n-1}} = \sum_{n=1}^{\infty} \frac{\Gamma(n+1/2)}{(n+1)!\Gamma(1/2)}. \tag{6.3}$$

Note that the series has the form $\sum_{n=2}^{\infty} a_n$ with $a_n = \frac{1}{n!}\frac{1\cdot 3\cdot 5\ldots(2n-3)}{2^{n-1}}$, and so

$$\lim_{n\to\infty} n\left(\frac{a_n}{a_{n+1}} - 1\right) = \lim_{n\to\infty} n\left(\frac{2n+2}{2n-1} - 1\right) = \frac{3}{2} > 1. \tag{6.4}$$

Hence Raabe's convergence test shows convergence. (For a full discussion of this convergence test, see [3].) The conclusion is that the forward scattering series for this model converges at the critical angle as well. Note that, in this case, the sum of the series, which corresponds to the reflection coefficient, is $R = 1$.

7. Postcritical divergence. For a $c_0 < c_1$ model, the forward series converges for precritical and critical incidence and diverges for postcritical incidence. From wave nonperturbative theory, the reflection coefficient R, which should be constructed by the forward scattering series, is

- $R = \frac{\nu_0 - \nu_1}{\nu_0 + \nu_1} < 1$ for precritical incidence. In this case both ν_0 and ν_1 are real.

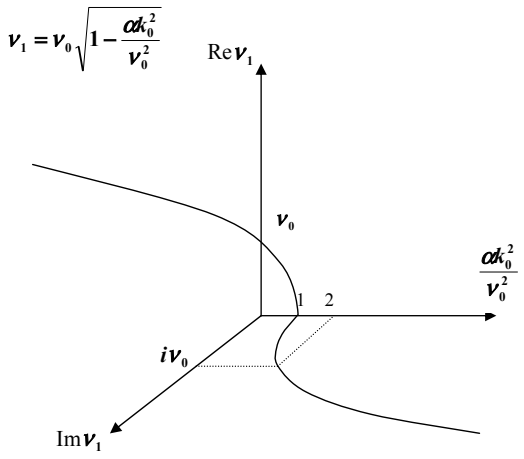


FIG. 7.1. *The graph of* ν_1 *as a function of* $\frac{\alpha k_0^2}{\nu_0^2}$.

- $R = 1$ for critical incidence. In this case $\nu_1 = 0$.
- $R = \frac{\nu_0 - \nu_1}{\nu_0 + \nu_1}$ for postcritical incidence. In this case ν_1 is purely imaginary, and hence R is complex. However, $|R| = 1$, and the complexity of R is attributed to a phase-shift of the emerging wave after hitting the interface due to the evanescent waves created in the second medium.

The term $\alpha k_0^2/\nu_0^2 = 1 - \nu_1^2/\nu_0^2$ is > 1 exactly when ν_1 becomes imaginary. In fact, if for this case one writes $R = e^{i\varepsilon}$, where ε is the phase-shift of the wave-field, then $\alpha k_0^2/\nu_0^2 = 1 + \tan^2 \varepsilon/2$, enforcing the earlier statement that the divergence is due to the phase-shift of the reflected wave. In other words, it is the impossibility of constructing a complex number ν_1 as a series of real numbers (powers of ν_0) which leads to the divergence of the series. The graph of ν_1 as a function of $\alpha k_0^2/\nu_0^2$ is shown in Figure 7.1.

For $c_0 < c_1$we have that $\alpha > 0$, so we are looking at the positive x-axis of the graph; if the velocity model is fixed, αk_0^2 is a constant. The vertical wavenumber of the propagating wave in the actual medium, ν_1, is equal to ν_0 when $\alpha k_0^2/\nu_0^2 = 0$, i.e., at normal incidence. When $\alpha k_0^2/\nu_0^2 = 1$ (at critical incidence), ν_1 is zero, showing that there is no propagation into the second medium. When $\alpha k_0^2/\nu_0^2 > 1$ (postcritical incidence), ν_1 is complex, and it becomes unrecoverable by a Taylor series written at $\alpha k_0^2/\nu_0^2 = 0$; the series is now divergent. For $c_0 > c_1$ it seems like this problem does not exist. In this case there is no critical angle, and so the vertical wavenumber ν_1 never becomes complex. However, the series inherits the divergent behavior for $\alpha k_0^2/\nu_0^2 < -1$ due to the singularity at $\alpha k_0^2/\nu_0^2 = 1$. For any value of $\alpha k_0^2/\nu_0^2$ outside the unit sphere centered at the origin the series will diverge due to that same singularity.

8. Conclusion. We have shown that the interchange of certain integrals in the calculation of terms in the forward scattering series yields the same result as the far field approximations of those integrals. The later approach allows the study of the restrictions imposed on the model by the former approach and provides new insights and physical interpretations for the terms in the forward scattering series. It is also anticipated that the new method would be more practical in the study of more complicated models (e.g., line source and receiver).

We have also proved the convergence of the forward scattering series at critical angle for the model of Matson [7] and provided an explanation for the divergence of the series for postcritical incident angles. The divergence is due to the inability of the forward scattering series to construct a complex vertical wavenumber from a series of real terms. Several possibilities for extending this result exist. First, one could introduce imaginary terms in the calculated series by using more than just the leading asymptotic behavior of the integral representation of the Hankel function, or of the dx' integral involved in the calculations. Second, one could try to make use of the evanescent part of the wave-field emanating from the scatterers to construct a complex vertical wavenumber. The evanescent part is always discarded when the asymptotic behavior of the integral representation of the Hankel function is considered; using it is attractive because it makes sense intuitively to construct an evanescent wave in the actual medium using evanescent waves in the reference medium. Third, an imaginary term in the reference velocity, and hence complex terms in the forward scattering series, could be brought in by the introduction of an absorptive reference medium. These ideas will be considered in future research.

Appendix. In section 3 we indicated how to calculate the first few terms in the forward scattering series for the reflected wave-field in a 2D vertically varying medium. We stated there that the calculated series for the scattered field for that specific model is (see (3.14))

(A.1)
$$P(x_g, z_g|\, k;\omega) = e^{ikx_g} e^{i\nu_0(2z_1 - z_g)} \left[\frac{1}{4}\frac{k_0^2\alpha}{\nu_0^2} + \frac{1}{8}\left(\frac{k_0^2\alpha}{\nu_0^2}\right)^2 + \frac{5}{64}\left(\frac{k_0^2\alpha}{\nu_0^2}\right)^3 + \cdots \right],$$

which is recognized to be the Taylor series for $\sqrt{1 - k_0^2\alpha/\nu_0^2}$ about $\frac{k_0^2\alpha}{\nu_0^2} = 0$ after some algebraic operations are performed on it. In this section we provide a rigorous proof of this statement. The proof will proceed as follows: first we will write down the general term for the transmitted wave-field and show by induction that the expression is correct; then we will use it to calculate the general term for the reflected wave-field and show that it corresponds to the general term in the aforementioned Taylor series. The need for the general term for the transmitted field is obvious since the iteration step occurs in the transmitted wave rather than the reflected one. Once the general term for the transmitted wave-field, P_n^T, is obtained, the general term for the reflected wave-field, P_n^R, is obtained by calculating

(A.2)
$$\begin{aligned} &P_{n+1}^R\,(x_g, z_g < z_1|\, k;\omega) \\ &= \int_{z_1}^{\infty} dz' \int_{-\infty}^{\infty} dx' \frac{1}{2\pi} \left(\int_{-\infty}^{\infty} \frac{e^{ik_g(x_g - x')} e^{i\nu_{0g}|z_g - z'|}}{2i\nu_0} dk_g \right) k_0^2\alpha P_n^T\,(x', z'|\, k;\omega)\,. \end{aligned}$$

To simplify the writing we introduce the notation

(A.3)
$$\frac{k_0^2\alpha}{\nu_0^2} = X$$

and

(A.4)
$$S_n = \frac{X^n}{2^n n!}\,(1+R)^{n+1}\,,$$

where

$$(1+R) = \frac{2}{X}\left[1 - Taylor\left(\sqrt{1-X}\right)\right] \tag{A.5}$$

and $Taylor\left(\sqrt{1-X}\right)$ stands for the Taylor series of $\sqrt{1-X}$ about $X=0$. Notice that S_n is a series in X of lowest order n. Also denote by S_n^j the coefficient of the jth order in S_n, and notice that all these coefficients are zero for $j<n$.

We will prove by induction that the general term for the transmitted wave-field P_n^T for $n \geq 1$ is

$$P_n^T\left(x_g, z_g > z_1 \middle| \, k;\omega\right) = e^{ikx_g} e^{i\nu_0 z_g} X^n \sum_{l=0}^{n} \left[-i\nu_0\left(z_g - z_1\right)\right]^l S_l^n. \tag{A.6}$$

The first step of the induction is to verify this relation for $n=1$, i.e., to check that

$$P_1^T = e^{ikx_g} e^{i\nu_0 z_g} X\left\{S_0^1 + \left[-i\nu_0\left(z_g - z_1\right)\right] S_1^1\right\}. \tag{A.7}$$

Note that $S_0^1 = 1/4$ and $S_1^1 = 1/2$, and hence this is the expression (2.25) found in Matson [7]. For the second step of the induction we assume that the relation (A.6) for P_n^T is true, and we calculate P_{n+1}^T and show that it has the same form; i.e., we want to prove that

$$P_{n+1}^T\left(x_g, z_g > z_1 \middle| \, k;\omega\right) = e^{ikx_g} e^{i\nu_0 z_g} X^{n+1} \sum_{l=0}^{n+1} \left[-i\nu_0\left(z_g - z_1\right)\right]^l S_l^{n+1}. \tag{A.8}$$

We have

$$\begin{aligned} P_{n+1}^T &= \int_{z_1}^{\infty} dz' \int_{-\infty}^{\infty} dx' \frac{1}{2\pi}\left(\int_{-\infty}^{\infty} \frac{e^{ik_g\left(x_g - x'\right)} e^{i\nu_{0g}\left|z_g - z'\right|}}{2i\nu_0} dk_g\right) k_0^2 \alpha P_n^T\left(x', z' \middle| \, k;\omega\right) \\ &= \int_{z_1}^{\infty} dz' \int_{-\infty}^{\infty} dx' \frac{1}{2\pi}\left(\int_{-\infty}^{\infty} \frac{e^{ik_g\left(x_g - x'\right)} e^{i\nu_{0g}\left|z_g - z'\right|}}{2i\nu_0} dk_g\right) \\ &\qquad \times k_0^2 \alpha e^{ikx'} e^{i\nu_0 z'} X^n \sum_{l=0}^{n} \left[-i\nu_0\left(z' - z_1\right)\right]^l S_l^n. \end{aligned} \tag{A.9}$$

We now solve the dk_g and the dx' by either one of the two methods described in the text and obtain

$$\begin{aligned} P_{n+1}^T &= \int_{z_1}^{\infty} dz' \frac{k_0^2 \alpha}{2i\nu_0} e^{ikx_g} e^{i\nu_0\left|z_g - z'\right|} e^{i\nu_0 z'} X^n \sum_{l=0}^{n} \left[-i\nu_0\left(z' - z_1\right)\right]^l S_l^n \\ &= e^{ikx_g} X^{n+1} \frac{\nu_0}{2i} \int_{z_1}^{\infty} dz' e^{i\nu_0\left|z_g - z'\right|} e^{i\nu_0 z'} \sum_{l=0}^{n} \left[-i\nu_0\left(z' - z_1\right)\right]^l S_l^n \\ &= e^{ikx_g} X^{n+1} \frac{(-i\nu_0)}{2} \sum\nolimits_{l=0}^{n} \int_{z_1}^{\infty} dz' e^{i\nu_0\left|z_g - z'\right|} e^{i\nu_0 z'} \left[-i\nu_0\left(z' - z_1\right)\right]^l S_l^n. \end{aligned} \tag{A.10}$$

We split the integral into two integrals in order to be able to evaluate the absolute

value and get

$$P_{n+1}^T = e^{ikx_g} X^{n+1} \frac{(-i\nu_0)}{2} \sum_{l=0}^{n} \left\{ \int_{z_1}^{z_g} dz' e^{i\nu_0 z_g} \left[-i\nu_0 (z'-z_1)\right]^l S_l^n + \int_{z_g}^{\infty} dz' e^{i\nu_0(2z'-z_g)} \left[-i\nu_0 (z'-z_1)\right]^l S_l^n \right\}. \tag{A.11}$$

The first integral has an easy solution; the second is a bit more tedious since it involves integration by parts. After solving the two integrals, we find

(A.12)

$$\begin{aligned} P_{n+1}^T = e^{ikx_g} e^{i\nu_0 z_g} X^{n+1} \Bigg\{ & \sum_{l=0}^{n} \frac{S_l^n}{2(l+1)} \left[-i\nu_0 (z_g - z_1)\right]^{l+1} \\ & + \frac{(-i\nu_0)}{2} \Bigg[S_0^n (-i\nu_0)^0 \left(-\frac{1}{2i\nu_0} \right) \\ & \quad + S_1^n (-i\nu_0)^1 \left(-\frac{1}{2i\nu_0}(z_g - z_1) + \frac{1}{(2i\nu_0)^2} \right) \\ & \quad + S_2^n (-i\nu_0)^2 \left(-\frac{1}{2i\nu_0}(z_g - z_1)^2 + \frac{2}{(2i\nu_0)^2}(z_g - z_1) - \frac{2}{(2i\nu_0)^3} \right) \\ & \quad \vdots \\ & \quad + S_n^n (-i\nu_0)^n \left(\frac{-1}{2i\nu_0}(z_g - z_1)^n + \frac{n}{(2i\nu_0)^2}(z_g - z_1)^{n-1} + \cdots + \frac{(-1)^{n+1} n!}{(2i\nu_0)^{n+1}} \right) \Bigg] \Bigg\}. \end{aligned}$$

Grouping together the terms with like powers of $[-i\nu_0 (z' - z_1)]$ in the expression above, we find

(A.13)

$$\begin{aligned} P_{n+1}^T = e^{ikx_g} e^{i\nu_0 z_g} X^{n+1} \Bigg\{ & \left[-i\nu_0 (z_g - z_1)\right]^{n+1} \frac{S_n^n}{2(n+1)} \\ & + \left[-i\nu_0 (z_g - z_1)\right]^{n} \left(\frac{S_{n-1}^n}{2n} + \frac{S_n^n}{2} \frac{1}{2} \right) \\ & + \left[-i\nu_0 (z_g - z_1)\right]^{n-1} \left(\frac{S_{n-2}^n}{2(n-1)} + \frac{S_n^n}{2} \frac{n}{2^2} + \frac{S_{n-1}^n}{2} \frac{1}{2} \right) \\ & + \left[-i\nu_0 (z_g - z_1)\right]^{n-2} \left(\frac{S_{n-3}^n}{2(n-2)} + \frac{S_n^n}{2} \frac{n(n-1)}{2^3} + \frac{S_{n-1}^n}{2} \frac{n-1}{2^2} + \frac{S_{n-2}^n}{2} \frac{1}{2} \right) \\ & \vdots \\ & + \left[-i\nu_0 (z_g - z_1)\right]^{1} \left(\frac{S_0^n}{2} + \frac{S_n^n}{2} \frac{n(n-1)\ldots 2}{2^n} + \frac{S_{n-1}^n}{2} \frac{(n-1)\ldots 2}{2^{n-1}} + \cdots + \frac{S_1^n}{2} \frac{1}{2} \right) \\ & + \left[-i\nu_0 (z_g - z_1)\right]^{0} \left(0 + \frac{S_n^n}{2} \frac{n!}{2^{n+1}} + \frac{S_{n-1}^n}{2} \frac{(n-1)!}{2^n} + \cdots + \frac{S_1^n}{2} \frac{1!}{2^2} + \frac{S_0^n}{2} \frac{1}{2} \right) \Bigg\}. \end{aligned}$$

We next show that the coefficients of $[-i\nu_0 (z_g - z_1)]^j$ in the above expression are exactly equal to S_j^{n+1}, and hence this last expression is the one required for the second step of the induction (see (A.8)).

For the first coefficient recall that, by definition, we have

$$S_n = \frac{X^n}{2^n n!}\left(1+R\right)^{n+1}, \tag{A.14}$$

and hence we can write

$$S_{n+1} = \frac{X^{n+1}}{2^{n+1}(n+1)!}\left(1+R\right)^{n+2} = \frac{X}{2(n+1)}\left(1+R\right)S_n. \tag{A.15}$$

This is an equality of two series, which implies that the coefficients of identical powers from both sides are equal. By equating the coefficients of the $n+1$ power, we obtain

$$S_{n+1}^{n+1} = \frac{1}{2(n+1)}S_n^n. \tag{A.16}$$

For the second coefficient we start with the identity

$$S_n = \frac{X^n}{2^n n!}\left(1+R\right)^{n+1} \tag{A.17}$$

and rewrite it as

$$S_n = \frac{X^n}{2^n n!}\left(1+R\right)^n\left(1+R\right) = \frac{X}{2n}S_{n-1} + \frac{X^n}{2^n n!}R\left(1+R\right)^n. \tag{A.18}$$

By equating the coefficients of the $n+1$ power from both sides, we find

$$S_n^{n+1} = \frac{1}{2n}S_{n-1}^n + \frac{1}{4}S_n^n, \tag{A.19}$$

where we have used that the coefficient of the first power of X in the expression for R is $1/4$.

For the third coefficient we start with the identity

$$S_{n-1} = \frac{X^{n-1}}{2^{n-1}(n-1)!}\left(1+R\right)^n \tag{A.20}$$

and rewrite it as

$$S_{n-1} = \frac{X}{2(n-1)}S_{n-2} + \frac{X^{n-1}}{2^{n-1}(n-1)!}R\left(1+R\right)^{n-2} + \frac{X^{n-1}}{2^{n-1}(n-1)!}R^2\left(1+R\right)^{n-2}. \tag{A.21}$$

By equating the coefficients of the $n+1$ power from both sides, we find

$$S_{n-1}^{n+1} = \frac{1}{2(n-1)}S_{n-2}^n + \frac{1}{4}S_{n-1}^n + \frac{n}{8}S_n^n. \tag{A.22}$$

For this last expression we have used again the fact that the coefficient of the first power of X in the expression for R is $1/4$.

The procedure outlined for these first three coefficient can be continued without difficulty to show that all the coefficients in the expression (A.13) coincide with those in (A.8). This concludes the second step of the induction and hence the proof that the expression for the transmitted wave-field P_n^T is

$$P_n^T\left(x_g, z_g > z_1 \middle| \, k;\omega\right) = e^{ikx_g}e^{i\nu_0 z_g}X^n\sum_{l=0}^{n}\left[-i\nu_0\left(z_g - z_1\right)\right]^l S_l^n. \tag{A.23}$$

The general term in the forward scattering series representation (3.14) for the reflected wave-field can hence be calculated using the following formula:

(A.24)

$$P_{n+1}^{R}\left(x_g, z_g < z_1 \middle| \, k;\omega\right) = \int_{z_1}^{\infty} dz' \int_{-\infty}^{\infty} dx' \frac{1}{2\pi} \left(\int_{-\infty}^{\infty} \frac{e^{ik_g(x_g-x')} e^{i\nu_{0g}\left|z_g-z'\right|}}{2i\nu_0} dk_g \right) k_0^2 \alpha P_n^T\left(x', z' \middle| \, k;\omega\right).$$

Introducing the expression for P_n^T, we find

$$P_{n+1}^{R} = \int_{z_1}^{\infty} dz' \int_{-\infty}^{\infty} dx' \frac{1}{2\pi} \left(\int_{-\infty}^{\infty} \frac{e^{ik_g(x_g-x')} e^{i\nu_{0g}\left|z_g-z'\right|}}{2i\nu_0} dk_g \right) k_0^2 \alpha e^{ikx'} e^{i\nu_0 z'} X^n$$

(A.25)
$$\times \sum_{l=0}^{n} \left[-i\nu_0\left(z'-z_1\right)\right]^l S_l^n.$$

Solving for the dx' and the dk_g integrals gives

(A.26)
$$P_{n+1}^{R} = e^{ikx_g} X^{n+1} \frac{(-i\nu_0)}{2} \int_{z_1}^{\infty} dz' \sum_{l=0}^{n} \left[-i\nu_0\left(z'-z_1\right)\right]^l S_l^n e^{i\nu_0(2z'-z_g)}.$$

Notice that this integral has been dealt with before: it is the integral appearing in the second part of (A.11), and its solution is given in the second part of (A.12). However, the limits of integration are different: the solution for our integral may be obtained from the second part of (A.12) by replacing z_g with z_1. This substitution cancels most of the terms, and the result is

(A.27)
$$P_{n+1}^{R} = e^{ikx_g} e^{i\nu_0 z_g} X^{n+1} \frac{(-i\nu_0)}{2} \left[-S_0^n \frac{1}{2i\nu_0} - S_1^n \frac{1}{2^2 i\nu_0} - \cdots - S_n^n \frac{n!}{2^{n+1} i\nu_0} \right]$$

or

(A.28)
$$P_{n+1}^{R} = e^{ikx_g} e^{i\nu_0 z_g} X^{n+1} \frac{1}{4} \left[S_0^n + \frac{S_1^n}{2^1} 1! + \frac{S_2^n}{2^2} 2! + \cdots + \frac{S_n^n}{2^n} n! \right].$$

Again, the sum inside the square brackets is an expression that we have already analyzed before: it is the coefficient of $\left[-i\nu_0\left(z_g-z_1\right)\right]^0$ in (A.13). It was shown there that

(A.29)
$$\frac{1}{4} \left[S_0^n + \frac{S_1^n}{2^1} 1! + \frac{S_2^n}{2^2} 2! + \cdots + \frac{S_n^n}{2^n} n! \right] = S_0^{n+1},$$

and hence the expression for P_{n+1}^{R} becomes

(A.30)
$$P_{n+1}^{R}\left(x_g, z_g < z_1 \middle| \, k;\omega\right) = e^{ikx_g} e^{i\nu_0 z_g} X^{n+1} S_0^{n+1}.$$

Recall from (A.4) and (A.5) that S_0^{n+1} represents the coefficient of the $n+1$ degree in the series for $1+R$, and hence it is the coefficient of the $n+1$ degree in the Taylor series for $\sqrt{1-k_0^2\alpha/\nu_0^2}$ about $\frac{k_0^2\alpha}{\nu_0^2}=0$ after some algebraic operations are performed on it. The total scattered field P is the summation of all P_n^R and hence it represents the full Taylor series for $\sqrt{1-k_0^2\alpha/\nu_0^2}$ about $\frac{k_0^2\alpha}{\nu_0^2}=0$ after some algebraic operations are performed on it.

REFERENCES

[1] F. V. Araujo, *Linear and Nonlinear Methods Derived from Scattering Theory: Backscattered Tomography and Internal Multiple Attenuation,* Ph.D. thesis, Department of Geophysics, Universidad Federal de Bahia, Salvador-Bahia, Brazil, 1994 (in Portuguese).

[2] K. Aki and P. G. Richards, *Quantitative Seismology,* W. H. Freeman, San Francisco, CA, 1980.

[3] T. J. Bromwich, *An Introduction to the Theory of Infinite Series,* Macmillian, London, 1965.

[4] P. M. Carvalho, *Free Surface Multiple Reflection Elimination Method Based on Nonlinear Inversion of Seismic Data*, Ph.D. thesis, Department of Geophysics, Universidad Federal de Bahia, Salvador-Bahia, Brazil, 1992 (in Portuguese).

[5] R. W. Clayton and R. H. Stolt, *A Born-WKBJ inversion method for acoustic reflection data,* Geophys., 46 (1981), pp. 1559–1567.

[6] K. H. Matson, *The relationship between scattering theory and the primaries and multiples of reflection seismic data,* J. Seismic Exploration, 5 (1996), pp. 63–78.

[7] K. H. Matson, *An Inverse Scattering Series Method for Attenuating Elastic Multiples from Multicomponent Land and Ocean Bottom Seismic Data,* Ph.D. thesis, Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC, Canada, 1997.

[8] P. M. Morse and H. Feschbach, *Methods of Theoretical Physics,* McGraw–Hill, New York, 1953.

[9] R. H. Stolt and A. B. Weglein, *Migration and inversion of seismic data,* Geophys., 50 (1985), pp. 2458–2472.

[10] J. R. Taylor, *Scattering Theory*, John Wiley and Sons, New York, 1972.

[11] G. N. Watson, *A Treatise on the Theory of Bessel Functions,* 2nd ed., Cambridge University Press, Cambridge, UK, 1962.

[12] A. B. Weglein, F. A. Gasparotto, P. M. Carvalho, and R. H. Stolt , *An inverse scattering series method for attenuating multiples in seismic reflection data*, Geophys., 62 (1997), pp. 1975–1989.

[13] A. B. Weglein, F. V. Araujo, P. M. Carvalho, R. H. Stolt, K. H. Matson, R. Coates, D. Corrigan, D. J. Foster, S. A. Shaw, and H. Zhang, *Inverse scattering series and seismic exploration,* Top. Rev. Inverse Problems, 19 (2003), pp. R27–R83.

SIAM J. APPL. MATH.
Vol. 64, No. 6, pp. 2186–2204

GINZBURG–LANDAU MODEL IN THIN LOOPS WITH NARROW CONSTRICTIONS*

JACOB RUBINSTEIN†, MICHELLE SCHATZMAN‡, AND PETER STERNBERG†

Abstract. We consider the Ginzburg–Landau model for a superconducting thin ring in the presence of an applied field. The ring is constricted, and we derive an asymptotic form for the energy as the ring thickness tends to zero. The constriction leads in the limit to a jump condition for the order parameter, yielding a transmission condition across the weak link of the type postulated by de Gennes for superconducting/normal/superconducting junctions.

Key words. Ginzburg–Landau energy, asymptotic analysis, constricted rings

AMS subject classifications. 35Q60, 78M30, 78M35

DOI. 10.1137/S0036139903434456

1. Introduction.

The Josephson effect models the peculiar flow of currents through a normal thin layer, called a junction, separating two bulk superconducting samples. The fundamental feature of this effect is an expression for the current as a function of the phase difference across the junction:

$$J = J_M \sin[\phi]. \tag{1.1}$$

The parameter J_M is the maximal current that the junction can transmit, and $[\phi]$ denotes the difference between the phase of the superconducting wave function on the two ends of the junction.

Josephson predicted relation (1.1) from the microscopic Bardeen–Cooper–Schrieffer (BCS) theory [14]. Soon thereafter, it was realized that a similar expression, as well as some other features of the junction, can also be derived by ad hoc models that are coupled to the macroscopic Ginzburg–Landau (GL) model of superconductivity. In particular de Gennes [7] modeled the junction through a set of linear conditions relating the wave function and its derivatives on the two sides of the thin normal layer. Similar equations were written also in [1]. Alternatively, other authors, and, in fact, most of the physics literature on the subject (e.g., [21], [2]) use (1.1) as a basic paradigm and supplement it with equations and arguments based on the GL theory and classical electromagnetism.

It is therefore desirable to develop a theory for Josephson junctions that is built up coherently and directly upon the GL model. One way to do so is to model the normal layer into the GL energy functional. This was done in [5], [9], [10], [12], [11], and [20]. In particular it was shown in [20] that a large variety of junctions can be modeled in this way, leading to different types of current flow patterns.

The purpose of the paper is to construct, directly from the GL equations, a "geometrical" Josephson junction. Such junctions, called "weak links" in the literature,

*Received by the editors September 4, 2003; accepted for publication (in revised form) April 6, 2004; published electronically September 14, 2004.
http://www.siam.org/journals/siap/64-6/43445.html
†Department of Mathematics, Indiana University, Bloomington, IN 47405 (jrubinst@indiana.edu, sternber@indiana.edu). The research of the first author was partially supported by NSF DMS-0203312. The research of the third author was partially supported by NSF DMS-0100540.
‡Laboratoire de Mathématiques Appliquées de Lyon, CNRS et Université Claude Bernard, 69622 Villeurbanne Cedex, France (schatz@maply.univ-lyon1.fr).

are characterized by a sharp constriction in the thickness of the sample [16]. We shall show below that under appropriate selection of the sample geometry and its scaling, the GL model converges to a new model that provides in a natural way the linear relations postulated by de Gennes. The convergence is established rigorously in section 2, with the proof inspired in part by an analogous convergence result in the field of elasticity [6]. In section 3 we discuss some implications of the convergence result. Finally, we should mention the work in [15], as it also relates to variational problems in constricted domains, though in the context of micromagnetics. While in our problem, the asymptotic limit leads us to a one-dimensional problem, the limit in [15] is in general two-dimensional.

2. Formulation. We begin with a description of the geometry of the region Ω_ε to be occupied by the sample. To this end, we introduce a continuous, piecewise linear function $g_\varepsilon : [-\pi, \pi] \to \mathbb{R}^1$ that will govern the thickness of the ring. Fixing any positive number $p < 1$, we define g_ε via

$$g_\varepsilon(y_1) = (\varepsilon^{1-p} - 2\varepsilon)\,|y_1| + 2\varepsilon^{1+p} \quad \text{for} \quad |y_1| \le \varepsilon^p, \tag{2.1}$$

$$g_\varepsilon(y_1) = \varepsilon \quad \text{for} \quad \varepsilon^p \le |y_1| \le \pi. \tag{2.2}$$

We then define a ring-shaped region $\Omega_\varepsilon \subset \mathbb{R}^3$ of thickness g_ε as the image of the cylinder

$$\mathcal{C} = \{(y_1, y_2, y_3) : -\pi \le y_1 \le \pi,\ 0 \le y_2^2 + y_3^2 < 1\} \tag{2.3}$$

under the mapping $T_\varepsilon : \mathbb{R}^3 \to \mathbb{R}^3$ given by

$$x = T_\varepsilon(y_1, y_2, y_3) = \big((1 + g_\varepsilon(y_1)y_2)\cos y_1, (1 + g_\varepsilon(y_1)y_2)\sin y_1, g_\varepsilon(y_1)y_3\big). \tag{2.4}$$

That is, $\Omega_\varepsilon \equiv T_\varepsilon(\mathcal{C})$. Note, in particular, that in (x_1, x_2, x_3)-space, the variable y_1 corresponds to the polar angle in the x_1x_2-plane and that the ring Ω_ε has uniform thickness ε except near the constriction at $y_1 = 0$. See Figure 1.

We will use the following nondimensional version of the GL energy functional

$$G_\epsilon(u, \mathbf{A}) = \frac{1}{\varepsilon^2}\int_{\Omega_\epsilon}\left(|(i\nabla + \mathbf{A})u|^2 + \frac{\nu^2}{2}\left(|u|^2 - \mu^2\right)^2\right)dx + \frac{1}{\varepsilon^2}\int_{\mathbb{R}^3}|\nabla \times \mathbf{A} - \mathbf{H}^e|^2\,dx. \tag{2.5}$$

Here $u : \Omega_\varepsilon \to \mathbb{C}$ is the order parameter, $\mathbf{A} : \mathbb{R}^3 \to \mathbb{R}^3$ is the magnetic potential associated with the magnetic field $\mathbf{H}$ through $\nabla \times \mathbf{A} = \mathbf{H}$, and $\mathbf{H}^e$ is a given, smoothly varying, applied magnetic field directed along the x_3-axis and taken to be independent of the coordinate x_3. The quantities ν and μ are material parameters with μ^2 proportional to the difference between the critical temperature T_c and the temperature of the sample [20]. We assume we are in the superconducting temperature regime where this difference is positive. One could scale μ out by setting it to be one, but we retain it in order to use it later on as a bifurcation parameter. The energy G_ε has been scaled so that the minimum energy remains uniformly bounded away from both zero and infinity for small ε.

We would like to investigate the asymptotic behavior of minimizers to (2.5), and this will require a precise description of function spaces over which the minimization

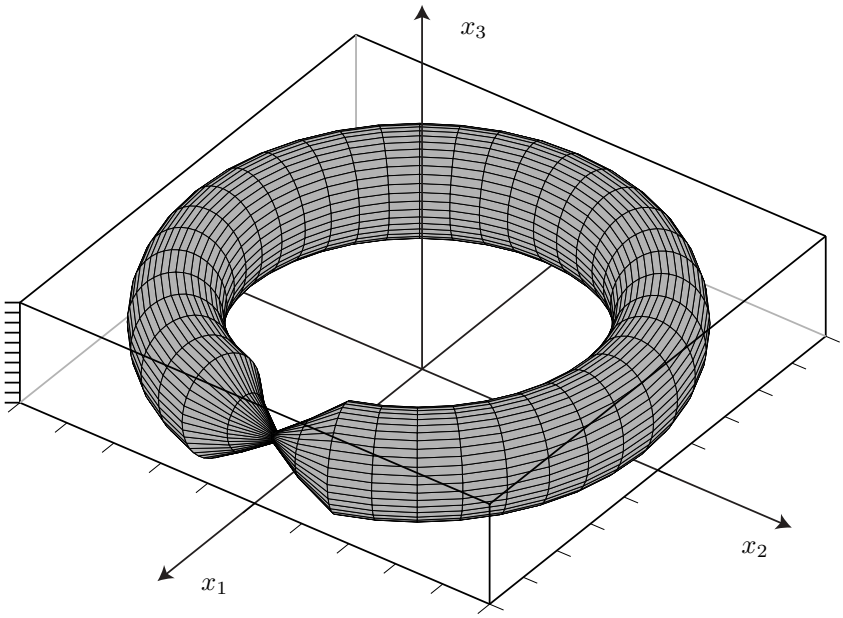


FIG. 1. *The constricted ring* Ω_ε.

is to take place. For the order parameter u we shall take competitors in the standard Sobolev space $W^{1,2}(\Omega_\varepsilon;\mathbb{C})$ consisting of square-integrable functions with square-integrable first derivatives. For the magnetic potential $\mathbf{A}$ we introduce the space $\mathcal{H}$ as the completion of the set

$$\{\phi \in C^\infty(\mathbb{R}^3;\mathbb{R}^3) : \phi \text{ compactly supported}\}$$

with respect to the norm $\|\nabla\phi\|_{L^2(\mathbb{R}^3;\mathbb{R}^3)} = (\int_{\mathbb{R}^3} |\nabla\phi|^2\, dx)^{1/2}$. Then we define $\mathcal{H}_0$ to be

$$\mathcal{H}_0 = \{\phi \in \mathcal{H} : \operatorname{div}\phi = 0\}$$

and consider competitors $\mathbf{A}$ satisfying $\mathbf{A} - \mathbf{A}^e \in \mathcal{H}_0$, where $\mathbf{A}^e = \mathbf{A}^e(x_1, x_2)$ is the applied magnetic potential satisfying

$$\nabla \times \mathbf{A}^e = \mathbf{H}^e \quad \text{and } \operatorname{div}\mathbf{A}^e = 0 \quad \text{in } \mathbb{R}^3, \tag{2.6}$$

$$\mathbf{A}^e \cdot (0,0,1) = 0 \quad \text{in } \mathbb{R}^3. \tag{2.7}$$

Condition (2.7) holds by our assumptions on $\mathbf{H}^e$, while we can arrange for the zero divergence condition by a suitable choice of gauge.

Through a rather standard application of the direct method in the calculus of variations, along with standard elliptic regularity theory, one obtains the following.

THEOREM 2.1. *For all positive* $\varepsilon < 1$*, there exists a pair* $(u^\varepsilon, \mathbf{A}^\varepsilon)$ *solving the variational problem*

$$\inf_{\{u\in W^{1,2}(\Omega_\varepsilon;\mathbb{C}),\mathbf{A}-\mathbf{A}^e\in\mathcal{H}_0\}} G_\varepsilon(u,\mathbf{A}). \tag{2.8}$$

The function u^ε is smooth in Ω_ε, while the function $\mathbf{A}^\varepsilon$ is smooth in $\mathbb{R}^3 \setminus \partial\Omega_\varepsilon$ and continuously differentiable across $\partial\Omega$. Furthermore, the minimizers satisfy the GL system

$$(i\nabla + \mathbf{A}^\varepsilon)^2 u^\varepsilon = \nu^2(|u^\varepsilon|^2 - \mu^2)u^\varepsilon \quad \text{in} \quad \Omega_\varepsilon, \tag{2.9}$$

$$\begin{aligned} &\nabla \times \nabla \times (\mathbf{A}^\varepsilon - \mathbf{A}^e) = -\Delta(\mathbf{A}^\varepsilon - \mathbf{A}^e) \\ &= \begin{cases} \frac{i}{2}(\overline{u^\varepsilon}\nabla u^\varepsilon - u^\varepsilon \nabla \overline{u^\varepsilon}) - |u^\varepsilon|^2 \mathbf{A}^\varepsilon & \text{for } x \in \Omega_\varepsilon, \\ 0 & \text{for } x \in \mathbb{R}^3 \setminus \overline{\Omega_\varepsilon} \end{cases} \end{aligned} \tag{2.10}$$

and the boundary condition

$$(i\nabla + \mathbf{A}^\varepsilon)u^\varepsilon \cdot \mathbf{n} = 0 \quad \text{on} \quad \partial\Omega_\varepsilon. \tag{2.11}$$

Here $\bar{\cdot}$ denotes complex conjugation and $\mathbf{n}$ denotes the outer unit normal along $\partial\Omega_\varepsilon$.

Finally, the order parameter u^ε satisfies the condition

$$|u^\varepsilon| \le \mu \quad \text{in} \quad \overline{\Omega_\varepsilon}. \tag{2.12}$$

Application of the direct method in establishing the existence of minimizers to the GL energy can be found, for instance, in [8] or [19]. The regularity theory in this context can be found, for instance, in [13]. Inequality (2.12) is an easy consequence of the maximum principle; see, e.g., [8].

PROPOSITION 2.2. *There exist positive constants C_1 and C_2 independent of ε such that*

$$G_\varepsilon(u^\varepsilon, \mathbf{A}^\varepsilon) \le C_1 \quad \text{and} \tag{2.13}$$

$$\int_{\Omega_\varepsilon} |\nabla u^\varepsilon|^2 \, dx \le C_2 \varepsilon^2. \tag{2.14}$$

Furthermore, one has the uniform convergence

$$\|\mathbf{A}^\varepsilon - \mathbf{A}^e\|_{L^\infty(B_R(0);\mathbb{R}^3)} \to 0 \quad \text{as} \quad \varepsilon \to 0 \quad \text{for every } R > 0, \tag{2.15}$$

where $B_R(0) = \{x \in \mathbb{R}^3 : |x| < R\}$. Condition (2.15) in particular implies that

$$\sup_{y \in \mathcal{C}} |\mathbf{A}^\varepsilon(T_\varepsilon(y)) - \mathbf{A}^e(\cos y_1, \sin y_1, 0)| \to 0 \quad \text{as} \quad \varepsilon \to 0. \tag{2.16}$$

Proof. The bound (2.13) follows immediately by comparing the energy of the minimizer to that of the pair $(\mu, \mathbf{A}^e)$:

$$G_\varepsilon(u^\varepsilon, \mathbf{A}^\varepsilon) \le G_\varepsilon(\mu, \mathbf{A}^e) = \frac{\mu^2}{\varepsilon^2} \int_{\Omega_\varepsilon} |\mathbf{A}^e|^2 \, dx \le \frac{\mu^2 \mathrm{vol}(\Omega_\varepsilon)}{\varepsilon^2} \|\mathbf{A}^e\|^2_{L^\infty(\Omega_\varepsilon)} \le C_1$$

since $\mathrm{vol}(\Omega_\varepsilon) = \mathcal{O}(\varepsilon^2)$.

We next establish the convergence (2.15) and the bound (2.14) using (2.13) by decomposing G_ε as

$$\begin{aligned} \varepsilon^2 G_\varepsilon(u^\varepsilon, \mathbf{A}^\varepsilon) &= \int_{\Omega_\varepsilon} |\nabla u^\varepsilon|^2 \, dx + i \int_{\Omega_\varepsilon} (\overline{u^\varepsilon}\nabla u^\varepsilon - u^\varepsilon \nabla \overline{u^\varepsilon}) \cdot \mathbf{A}^\varepsilon \, dx \\ &+ \int_{\Omega_\varepsilon} |u^\varepsilon|^2 |\mathbf{A}^\varepsilon|^2 + \frac{\nu^2}{2}(|u^\varepsilon|^2 - \mu^2)^2 \, dx + \int_{\mathbb{R}^3} |\nabla \times (\mathbf{A}^\varepsilon - \mathbf{A}^e)|^2 \, dx. \end{aligned} \tag{2.17}$$

Applying (2.12), we find that

$$\left| i \int_{\Omega_\varepsilon} (\overline{u^\varepsilon} \nabla u^\varepsilon - u^\varepsilon \nabla \overline{u^\varepsilon}) \cdot \mathbf{A}^\varepsilon \, dx \right| \leq 2\mu \int_{\Omega_\varepsilon} |\nabla u^\varepsilon| \, |\mathbf{A}^\varepsilon| \, dx$$
$$\leq \frac{1}{2} \int_{\Omega_\varepsilon} |\nabla u^\varepsilon|^2 \, dx + C(\mu) \int_{\Omega_\varepsilon} |\mathbf{A}^\varepsilon|^2 \, dx.$$

Hence, (2.13) and (2.17) yield the bound

$$\int_{\Omega_\varepsilon} |\nabla u^\varepsilon|^2 \, dx \leq 2C_1 \varepsilon^2 + C(\mu) \int_{\Omega\varepsilon} |\mathbf{A}^\varepsilon|^2 \, dx$$
$$\leq 2C_1 \varepsilon^2 + C(\mu) \left(\|\mathbf{A}^\varepsilon - \mathbf{A}^e\|^2_{L^\infty(\Omega_\varepsilon;\mathbb{R}^3)} + \|\mathbf{A}^e\|^2_{L^\infty(\Omega_\varepsilon;\mathbb{R}^3)} \right) \mathrm{vol}\,(\Omega_\varepsilon). \tag{2.18}$$

Now we turn to (2.10) satisfied by the difference $\mathbf{A}^\varepsilon - \mathbf{A}^e$ and utilize the fact that this difference lies in $\mathcal{H}_0$. This decay at infinity allows us to express it via the fundamental solution to the Laplacian in $\mathbb{R}^3$:

$$\mathbf{A}^\varepsilon - \mathbf{A}^e = \int_{\Omega_\varepsilon} \Gamma(x - z) f^\varepsilon(z) \, dz, \tag{2.19}$$

where $\Gamma(x) \equiv \frac{1}{4\pi|x|}$ and $f^\varepsilon(z) \equiv \frac{i}{2}(\overline{u^\varepsilon} \nabla u^\varepsilon - u^\varepsilon \nabla \overline{u^\varepsilon}) - |u^\varepsilon|^2 \mathbf{A}^\varepsilon$.

Given any $R > 0$, one readily checks that

$$\int_{\Omega_\varepsilon} |\Gamma(x - z)|^2 \, dz \leq C(R),$$

provided $|x| \leq R$, so that by Hölder's inequality, we obtain

$$\|\mathbf{A}^\varepsilon - \mathbf{A}^e\|_{L^\infty(B_R(0);\mathbb{R}^3)} \leq \sqrt{C(R)} \, \|f^\varepsilon\|_{L^2(\Omega_\varepsilon)}. \tag{2.20}$$

Then writing

$$f^\varepsilon = \frac{i}{2}(\overline{u^\varepsilon} \nabla u^\varepsilon - u^\varepsilon \nabla \overline{u^\varepsilon}) - |u^\varepsilon|^2 (\mathbf{A}^\varepsilon - \mathbf{A}^e) - |u^\varepsilon|^2 \mathbf{A}^e,$$

we can combine inequalities (2.12), (2.18), and (2.20) to conclude that

$$\|\mathbf{A}^\varepsilon - \mathbf{A}^e\|_{L^\infty(B_R(0);\mathbb{R}^3)}$$
$$\leq C \left(\|\nabla u^\varepsilon\|_{L^2(\Omega_\varepsilon;\mathbb{R}^3)} + \|\mathbf{A}^\varepsilon - \mathbf{A}^e\|_{L^2(\Omega_\varepsilon;\mathbb{R}^3)} + \|\mathbf{A}^e\|_{L^2(\Omega_\varepsilon;\mathbb{R}^3)} \right) \tag{2.21}$$
$$\leq C \left(\varepsilon + \|\mathbf{A}^\varepsilon - \mathbf{A}^e\|_{L^\infty(\Omega_\varepsilon;\mathbb{R}^3)} \sqrt{\mathrm{vol}\,\Omega_\varepsilon} + \|\mathbf{A}^e\|_{L^\infty(\Omega_\varepsilon;\mathbb{R}^3)} \sqrt{\mathrm{vol}\,\Omega_\varepsilon} \right),$$

where again C depends on R. Hence we obtain (2.15) in that

$$\|\mathbf{A}^\varepsilon - \mathbf{A}^e\|_{L^\infty(B_R(0);\mathbb{R}^3)} \leq C\varepsilon \tag{2.22}$$

for some constant C independent of ε but depending on R, μ, C_1, and $\mathbf{A}^e$. Then (2.14) follows from (2.15) and (2.18). ∎

Our characterization of the asymptotic behavior of the sequence of minimizers $\{u^\varepsilon\}$ is most conveniently carried out using the variables (y_1, y_2, y_3) defined in (2.4). Thus, we introduce the notation

$$U^\varepsilon(y_1, y_2, y_3) := u^\varepsilon(T_\varepsilon(y_1, y_2, y_3)). \tag{2.23}$$

PROPOSITION 2.3. *There exists a subsequence* $\{\varepsilon_j\} \to 0$ *and a function* $U^0 \in BV(\mathcal{C};\mathbb{C})$ *such that* $U^{\varepsilon_j} \to U^0$ *in* $L^1(\mathcal{C};\mathbb{C})$. *Furthermore,* U^0 *is a function of* y_1 *only.*

Here $BV(\mathcal{C};\mathbb{C})$ denotes the space of complex-valued functions of bounded variation. The compactness assertion is based on the fact that sequences uniformly bounded in $W^{1,1}(\mathcal{C};\mathbb{C})$ (hence in BV) contain L^1-convergent subsequences; cf., e.g., [22, section 5.3].

Proof. To reiterate, the goal is a uniform $W^{1,1}(\mathcal{C};\mathbb{C})$-bound on the sequence $\{U^\varepsilon\}$. Such a bound will come from (2.14) once we express $\int_{\Omega_\varepsilon} |\nabla u^\varepsilon|^2 \, dx$ in terms of U^ε.

To this end, one carries out a lengthy but routine calculation to obtain

$$\int_{\Omega_\varepsilon} |\nabla u^\varepsilon|^2 \, dx = \frac{1}{2} \int_{\mathcal{C}} a_{ik} \big(U^\varepsilon_{y_i} \overline{U^\varepsilon}_{y_k} + \overline{U^\varepsilon}_{y_i} U^\varepsilon_{y_k} \big) \, dy, \tag{2.24}$$

where the 3×3 matrix A with entries $a_{ik} = a_{ik}(y_1, y_2, y_3)$ can be written in the form

$$A = \frac{1}{(1 + g_\varepsilon(y_1) y_2)} \big(B + D \big) \tag{2.25}$$

with

$$B = \begin{pmatrix} g_\varepsilon(y_1)^2 & -g_\varepsilon(y_1) g'_\varepsilon(y_1) y_2 & -g_\varepsilon(y_1) g'_\varepsilon(y_1) y_3 \\ -g_\varepsilon(y_1) g'_\varepsilon(y_1) y_2 & 1 + g'_\varepsilon(y_1)^2 y_2^2 & g'_\varepsilon(y_1)^2 y_2 y_3 \\ -g_\varepsilon(y_1) g'_\varepsilon(y_1) y_3 & g'_\varepsilon(y_1)^2 y_2 y_3 & 1 + g'_\varepsilon(y_1)^2 y_3^2 \end{pmatrix} \tag{2.26}$$

and

$$D = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 g_\varepsilon(y_1) y_2 + g_\varepsilon(y_1)^2 y_2^2 \end{pmatrix}. \tag{2.27}$$

At this point, we appeal to [6], where the eigenvalues of the matrix B are explicitly calculated and found to be given by the formulas

$$\begin{aligned} \lambda_1 &= \frac{1 + g_\varepsilon(y_1)^2 + g'_\varepsilon(y_1)^2 (y_2^2 + y_3^2)}{2} \\ &\quad - \frac{\sqrt{[1 + g_\varepsilon(y_1)^2 + g'_\varepsilon(y_1)^2 (y_2^2 + y_3^2)]^2 - 4 g_\varepsilon(y_1)^2}}{2}, \end{aligned} \tag{2.28}$$

$$\begin{aligned} \lambda_2 &= \frac{1 + g_\varepsilon(y_1)^2 + g'_\varepsilon(y_1)^2 (y_2^2 + y_3^2)}{2} \\ &\quad + \frac{\sqrt{[1 + g_\varepsilon(y_1)^2 + g'_\varepsilon(y_1)^2 (y_2^2 + y_3^2)]^2 - 4 g_\varepsilon(y_1)^2}}{2}, \end{aligned} \tag{2.29}$$

$$\lambda_3 = 1.$$

Note, in particular, that $\lambda_2, \lambda_3 \geq 1$, while expansion of (2.28) reveals that

$$\frac{g_\varepsilon^2}{1 + g_\varepsilon^2 + (g'_\varepsilon)^2} \leq \lambda_1 \leq g_\varepsilon^2 \quad \text{for small } \varepsilon. \tag{2.30}$$

Let us now denote the eigenvalues of A by μ_1, μ_2, and μ_3. Since A is a regular perturbation of B, it follows easily that

$$\lim_{\varepsilon \to 0} |\mu_2 - \lambda_2| = 0 \quad \text{and} \quad \lim_{\varepsilon \to 0} |\mu_3 - 1| = 0. \tag{2.31}$$

Also, expanding $\det A$ along its last row, one readily checks that

$$(1 + g_\varepsilon(y_1)y_2)^3 \det A = \det B + \mathcal{O}(g_\varepsilon^3) \quad \text{as } \varepsilon \to 0. \tag{2.32}$$

Then phrasing (2.32) in terms of λ_i and μ_i and using (2.30) and (2.31), it is not hard to verify that

$$|\mu_1 - \lambda_1| = o(g_\varepsilon^2). \tag{2.33}$$

Hence, up to terms of order $o(g_\varepsilon^2)$,

$$\frac{g_\varepsilon^2}{1 + g_\varepsilon^2 + (g_\varepsilon')^2} \le \mu_1 \le g_\varepsilon^2. \tag{2.34}$$

If we then introduce the quantity

$$a_\varepsilon(y_1) := \frac{g_\varepsilon(y_1)^2}{\varepsilon^2\big(1 + g_\varepsilon(y_1)^2 + (g_\varepsilon'(y_1))^2\big)}, \tag{2.35}$$

we can combine (2.14), (2.24), (2.31), and (2.34) to conclude that

$$\begin{aligned} \int_{\mathcal{C}} a_\varepsilon \left|U^\varepsilon_{y_1}\right|^2 dy &\le \frac{1}{2\varepsilon^2} \int_{\mathcal{C}} a_{ik}\big(U^\varepsilon_{y_i}\overline{U^\varepsilon}_{y_k} + \overline{U^\varepsilon}_{y_i} U^\varepsilon_{y_k}\big)\, dy \\ &= \frac{1}{\varepsilon^2} \int_{\Omega_\varepsilon} |\nabla u^\varepsilon|^2 \, dx \le C_2. \end{aligned} \tag{2.36}$$

Hence,

$$\int_{\mathcal{C}} \left\{ g_\varepsilon(y_1)^2 \left|U^\varepsilon_{y_1}\right|^2 + \left|U^\varepsilon_{y_2}\right|^2 + \left|U^\varepsilon_{y_3}\right|^2 \right\} dy \le C\varepsilon^2 \tag{2.37}$$

for some constant C independent of ε. In particular, it follows that

$$\int_{\mathcal{C}} \left|U^\varepsilon_{y_2}\right|^2 + \left|U^\varepsilon_{y_3}\right|^2 dy \to 0 \quad \text{as } \varepsilon \to 0. \tag{2.38}$$

Arguing as in [6, Theorem 6.1], this leads to control of $\{\|\nabla U^\varepsilon\|_{L^1(\mathcal{C})}\}$ via (2.36), as follows:

$$\begin{aligned} \int_{\mathcal{C}} |\nabla U^\varepsilon| \, dy &= \int_{\mathcal{C}} \frac{1}{\sqrt{a_\varepsilon}} \sqrt{a_\varepsilon}\, |\nabla U^\varepsilon| \, dy \\ &\le \left(\int_{\mathcal{C}} \frac{1}{a_\varepsilon} \, dy \right)^{1/2} \left(\int_{\mathcal{C}} a_\varepsilon \, |\nabla U^\varepsilon|^2 \, dy \right)^{1/2} \\ &\le C_2^{1/2} \left(\int_{\mathcal{C}} \frac{1}{a_\varepsilon} \, dy \right)^{1/2}. \end{aligned} \tag{2.39}$$

Referring back to assumptions (2.1)–(2.2), we note that

$$\int_{-\pi}^{\pi} \frac{1}{a_\varepsilon(y_1)} \, dy_1 \to 2\pi + 1 \quad \text{as } \varepsilon \to 0. \tag{2.40}$$

In light of the bound (2.12), we see that

$$\|U^\varepsilon\|_{W^{1,1}(\mathcal{C})} < C,$$

and the L^1-convergence of a subsequence $\{U^{\varepsilon_j}\}$ to a $BV(\mathcal{C};\mathbb{C})$ function U^0 follows. In view of (2.38), one sees that U^0 is independent of y_2 and y_3. ∎

Next we wish to identify a limiting energy for G_ε. To this end, we introduce notation for the tangential component of the applied potential restricted to the unit circle:

$$A_1^e(y_1) \equiv \mathbf{A}^e(\cos y_1, \sin y_1, 0) \cdot (-\sin y_1, \cos y_1, 0). \tag{2.41}$$

We also introduce the function λ_ε via the formula

$$\frac{1}{a_\varepsilon(y_1)} = 1 + \lambda_\varepsilon(y_1). \tag{2.42}$$

One can easily check that the $\lambda_\varepsilon \, dy_1 \overset{*}{\rightharpoonup} \delta_0$ weakly as measures so that

$$\frac{1}{a_\varepsilon} \, dy_1 \overset{*}{\rightharpoonup} 1 \, dy_1 + \delta_0. \tag{2.43}$$

This allows us to establish a generalization of Theorem 6.2 of [6]. To state the result, we define the functional G_0 acting on functions in $L^1((-\pi, \pi); \mathbb{C})$ by the formula

$$G_0(U) = \begin{cases} \int_{(-\pi,\pi)\setminus\{0\}} \left(\left| (i\frac{d}{dy_1} + A_1^e)U \right|^2 + \frac{\nu^2}{2}(|U|^2 - \mu^2)^2 \right) dy_1 + |U^+ - U^-|^2 \\ \qquad \text{if } U \in W^{1,2}((-\pi,\pi) \setminus \{0\}; \mathbb{C}), \quad U(-\pi) = U(\pi), \\ \qquad +\infty \quad \text{otherwise,} \end{cases} \tag{2.44}$$

where

$$U^+ = \lim_{y_1 \to 0^+} U(y_1) \quad \text{and} \quad U^- = \lim_{y_1 \to 0^-} U(y_1).$$

We point out that the condition $U \in W^{1,2}((-\pi, \pi) \setminus \{0\}; \mathbb{C})$ in particular implies that U can be continuously defined on $[-\pi, 0]$ and on $[0, \pi]$; see, e.g., [22].

We then will prove the following.

THEOREM 2.4. *The function U^0 provided by Proposition* 2.3 *lies in the space $W^{1,2}((-\pi, \pi) \setminus \{0\}; \mathbb{C})$ and minimizes G_0.*

Proof. The identification of U^0 as a minimizer of G_0 will be achieved through two claims. First we will show that

$$\liminf_{\varepsilon_j \to 0} G_{\varepsilon_j}(u^{\varepsilon_j}, \mathbf{A}^{\varepsilon_j}) \geq \pi G_0(U^0). \tag{2.45}$$

Then we will show that for any $V \in L^1((-\pi, \pi); \mathbb{C})$, there exists a sequence $\{v^\varepsilon\} \subset W^{1,2}(\Omega_\varepsilon; \mathbb{C})$ such that

$$\lim_{\varepsilon \to 0} G_\varepsilon(v^\varepsilon, \mathbf{A}^e) = \pi G_0(V). \tag{2.46}$$

Using the minimizing property of $\{(u^\varepsilon, \mathbf{A}^\varepsilon)\}$, we can then combine (2.45) and (2.46) to obtain $G_0(U^0) \leq G_0(V)$ as asserted. Of course, it will then also follow that $U^0 \in W^{1,2}((-\pi, \pi) \setminus \{0\}; \mathbb{C})$.

Proof of claim (2.45). For this argument it will be convenient to work in a different gauge. Specifically, we take an applied magnetic potential to still satisfy condition (2.7) but now also to satisfy

$$\mathbf{A}^e(x_1, x_2) \cdot (x_1, x_2, 0) = 0 \quad \text{for } x_1^2 + x_2^2 = 1. \tag{2.47}$$

We can achieve this if we drop the divergence-free requirement and insist only that $\operatorname{div} \mathbf{A}^e = 0$ in $\{x : x_1^2 + x_2^2 < 1\}$ by replacing $\mathbf{A}^e$ with $\mathbf{A}^e - \nabla\phi$, where ϕ is any smooth extension to $\mathbb{R}^3$ of the solution to

$$\begin{aligned} \Delta\phi &= 0 \quad \text{in } x_1^2 + x_2^2 < 1, \\ \nabla\phi \cdot (x_1, x_2, 0) &= \mathbf{A}^e \cdot (x_1, x_2, 0) \quad \text{on} \quad x_1^2 + x_2^2 = 1. \end{aligned}$$

Note that a solution ϕ exists in light of the divergence-free condition on $\mathbf{A}^e$ inside the disc, and the solution is independent of x_3 since the original $\mathbf{A}^e$ was as well. Consequently, $\mathbf{A}^e - \nabla\phi$ will, in particular, still satisfy (2.7).

We observe that as a consequence of (2.7) and (2.47), we have that

$$|\mathbf{A}^e(\cos y_1, \sin y_1, 0)| = |A_1^e(y_1)| \quad \text{for } -\pi \le y_1 \le \pi \tag{2.48}$$

(cf. (2.41)).

For the remainder of the proof, we then replace the original $\mathbf{A}^e$ by $\mathbf{A}^e - \nabla\phi$, $\mathbf{A}^\varepsilon$ by $\mathbf{A}^\varepsilon - \nabla\phi$, and u^ε by $u^\varepsilon e^{-i\phi}$. Of course, through gauge-invariance, we have that $G_\varepsilon(u^\varepsilon, \mathbf{A}^\varepsilon) = G_\varepsilon(u^\varepsilon e^{-i\phi}, \mathbf{A}^\varepsilon - \nabla\phi)$. We will not introduce new notation, but through an abuse of notation we still denote these three quantities using their original designations. We should also remark that this change does not affect the estimates (2.15) and (2.16) measuring the L^∞-norm of $\mathbf{A}^\varepsilon - \mathbf{A}^e$ since this difference is unchanged by the gauge transformation.

Discarding the nonnegative term $\int_{\mathbb{R}^3} |\nabla \times \mathbf{A}^{\varepsilon_j} - \mathbf{H}^e|^2 \, dx$, we begin with the decomposition of G_{ε_j} as

$$\begin{aligned} G_{\varepsilon_j}(u^{\varepsilon_j}, \mathbf{A}^{\varepsilon_j}) \ge{}& \frac{1}{\varepsilon_j^2} \int_{\Omega_{\varepsilon_j}} |\nabla u^{\varepsilon_j}|^2 \, dx + \frac{1}{\varepsilon_j^2} \int_{\Omega_{\varepsilon_j}} i(\overline{u^{\varepsilon_j}} \nabla u^{\varepsilon_j} - u^{\varepsilon_j} \nabla \overline{u^{\varepsilon_j}}) \cdot \mathbf{A}^{\varepsilon_j} \, dx \\ &+ \frac{1}{\varepsilon_j^2} \int_{\Omega_{\varepsilon_j}} |u^{\varepsilon_j}|^2 |\mathbf{A}^\varepsilon|^2 \, dx + \frac{1}{\varepsilon_j^2} \int_{\Omega_{\varepsilon_j}} \frac{\nu^2}{2} (|u^{\varepsilon_j}|^2 - \mu^2)^2 \, dx. \end{aligned} \tag{2.49}$$

We will analyze the limit of each of the four terms above separately. Most crucial is the first term, where in light of (2.36) we have

$$\begin{aligned} \liminf_{\varepsilon_j \to 0} \frac{1}{\varepsilon_j^2} \int_{\Omega_{\varepsilon_j}} |\nabla u^{\varepsilon_j}|^2 \, dx &\ge \liminf_{\varepsilon_j \to 0} \int_{\mathcal{C}} a_{\varepsilon_j}(y_1) \left| U_{y_1}^{\varepsilon_j} \right|^2 dy \\ &= \liminf_{\varepsilon_j \to 0} \int_{\mathcal{C}} \left| a_\varepsilon(y_1) U_{y_1}^{\varepsilon_j} \right|^2 \frac{1}{a_{\varepsilon_j}(y_1)} \, dy. \end{aligned}$$

Then conditions (2.40) and (2.43) allow for an appeal to [6, Theorem 6.2], to conclude that

$$\liminf_{\varepsilon_j \to 0} \frac{1}{\varepsilon_j^2} \int_{\Omega_{\varepsilon_j}} |\nabla u^{\varepsilon_j}|^2 \, dx \ge \pi \int_{(-\pi,\pi)\setminus\{0\}} \left| \frac{d\, U^0}{d\, y_1} \right|^2 dy_1 + \pi \left| (U^0)^+ - (U^0)^- \right|^2. \tag{2.50}$$

(See also [3] and [4].)

It remains to determine the limits of the last three integrals on the right-hand side of (2.49). To this end, we fix a positive number δ, and denote by $\mathcal{C}_\delta$ the set $\{y \in \mathcal{C} : |y_1| > \delta\}$. It then follows from (2.37) that $\{U^\varepsilon\}$ is bounded uniformly in $W^{1,2}(\mathcal{C}_\delta; \mathbb{C})$. Hence, we conclude from (2.38) and the Sobolev embedding theorem (see, e.g., [22]) that

$$U^{\varepsilon_{j_k}} \rightharpoonup V^0 \quad \text{in} \quad W^{1,2}(\mathcal{C}_\delta; \mathbb{C}) \quad \text{and} \tag{2.51}$$

$$U^{\varepsilon_{j_k}} \to V^0 \quad \text{in} \quad L^q(\mathcal{C}_\delta; \mathbb{C}) \quad \text{for all} \quad q < 6 \tag{2.52}$$

for some $W^{1,2}(\mathcal{C}_\delta;\mathbb{C})$ function V^0 that is independent of y_2 and y_3. In light of Proposition 2.3, we may then identify $V^0 = U^0$ and observe that the convergences above must hold along the full sequence $\{\varepsilon_j\}$. We caution, however, that the $W^{1,2}(\mathcal{C}_\delta;\mathbb{C})$-norm of U^0 depends on δ, so we should not in general expect that $U^0 \in W^{1,2}(\mathcal{C};\mathbb{C})$.

Since the Jacobian of the mapping T_ε given by (2.4) is found to be $g_\varepsilon(y_1)^2(1 + g_\varepsilon(y_1)y_2)$, one uses (2.1)–(2.2), (2.12), and (2.52) to calculate the limit of the last term of (2.49), as follows:

$$\begin{aligned}
&\lim_{\varepsilon_j\to 0}\int_{\Omega_{\varepsilon_j}} \frac{\nu^2}{2\varepsilon_j^2}(|u^{\varepsilon_j}|^2-\mu^2)^2\,dx\\
&=\lim_{\varepsilon_j\to 0}\int_{\mathcal{C}} \frac{\nu^2}{2\varepsilon_j^2}(|U^{\varepsilon_j}|^2-\mu^2)^2 g_{\varepsilon_j}(y_1)^2(1+g_\varepsilon(y_1)y_2)\,dy\\
&=\lim_{\varepsilon_j\to 0}\int_{\mathcal{C}_\delta} \frac{\nu^2}{2}(|U^{\varepsilon_j}|^2-\mu^2)^2\,dy+\mathcal{O}(\delta)\\
&=\int_{\mathcal{C}_\delta} \frac{\nu^2}{2}(|U^0|^2-\mu^2)^2\,dy+\mathcal{O}(\delta)\\
&=\pi\int_{\{\delta<|y_1|<\pi\}} \frac{\nu^2}{2}(|U^0|^2-\mu^2)^2\,dy_1+\mathcal{O}(\delta). \qquad (2.53)
\end{aligned}$$

Similarly, in light of (2.48), (2.16), and (2.52), we have for the second to last integral in (2.49) that

$$\begin{aligned}
&\lim_{\varepsilon_j\to 0}\frac{1}{\varepsilon_j^2}\int_{\Omega_{\varepsilon_j}} |u^{\varepsilon_j}|^2\,|\mathbf{A}^{\varepsilon_j}|^2\,dx\\
&=\lim_{\varepsilon_j\to 0}\frac{1}{\varepsilon_j^2}\int_{\mathcal{C}} |U^{\varepsilon_j}(y)|^2\,|\mathbf{A}^{\varepsilon_j}(y)|^2\,g_{\varepsilon_j}(y_1)^2(1+g_{\varepsilon_j}(y_1)y_2)\,dy\\
&=\lim_{\varepsilon_j\to 0}\frac{1}{\varepsilon_j^2}\int_{\mathcal{C}} |U^{\varepsilon_j}(y)|^2\,|\mathbf{A}^{e}(\cos y_1,\sin y_1,0)|^2\,g_{\varepsilon_j}(y_1)^2(1+g_{\varepsilon_j}(y_1)y_2)\,dy\\
&=\lim_{\varepsilon_j\to 0}\frac{1}{\varepsilon_j^2}\int_{\mathcal{C}} |U^{\varepsilon_j}(y)|^2\,|A_1^{e}(y_1)|^2\,g_{\varepsilon_j}(y_1)^2(1+g_{\varepsilon_j}(y_1)y_2)\,dy\\
&=\lim_{\varepsilon_j\to 0}\int_{\mathcal{C}_\delta} |U^{\varepsilon_j}|^2\,|A_1^{e}|^2\,dy+\mathcal{O}(\delta)\\
&=\int_{\mathcal{C}_\delta} |U^0|^2\,|A_1^{e}|^2\,dy_1+\mathcal{O}(\delta)\\
&=\pi\int_{\{\delta<|y_1|<\pi\}} |U^0|^2\,|A_1^{e}|^2\,dy_1+\mathcal{O}(\delta). \qquad (2.54)
\end{aligned}$$

Finally, we turn to the limit of the remaining integral in (2.49), namely,

$$\liminf_{\varepsilon_j\to 0}\frac{1}{\varepsilon_j^2}\int_{\Omega_{\varepsilon_j}} i(\overline{u^{\varepsilon_j}}\nabla u^{\varepsilon_j}-u^{\varepsilon_j}\nabla\overline{u^{\varepsilon_j}})\cdot\mathbf{A}^{\varepsilon_j}\,dx. \qquad (2.55)$$

Another straightforward but laborious calculation based on the change of variables

$x = T_\varepsilon(y)$ leads to the fact that

$$\begin{aligned}
&\int_{\Omega_{\varepsilon_j}} i(\overline{u^{\varepsilon_j}}\nabla u^{\varepsilon_j} - u^{\varepsilon_j}\nabla\overline{u^{\varepsilon_j}})\cdot \mathbf{A}^{\varepsilon_j}\,dx \\
&= \int_{\mathcal{C}} i\left(\overline{U^{\varepsilon_j}}U^{\varepsilon_j}_{y_1} - U^{\varepsilon_j}\overline{U^{\varepsilon_j}}_{y_1}\right)\left[(-\sin y_1, \cos y_1, 0)\cdot \mathbf{A}^{\varepsilon_j}(T_{\varepsilon_j})\right]\frac{g_{\varepsilon_j}(y_1)^2}{\varepsilon_j^2}\,dy \\
&\quad + \int_{\mathcal{C}} i\left(\overline{U^{\varepsilon_j}}U^{\varepsilon_j}_{y_2} - U^{\varepsilon_j}\overline{U^{\varepsilon_j}}_{y_2}\right)\left[(\cos y_1, \sin y_1, 0)\cdot \mathbf{A}^{\varepsilon_j}(T_{\varepsilon_j})\right]\frac{g_{\varepsilon_j}(y_1)}{\varepsilon_j^2}\,dy \\
&\quad + \int_{\mathcal{C}} i\left(\overline{U^{\varepsilon_j}}U^{\varepsilon_j}_{y_2} - U^{\varepsilon_j}\overline{U^{\varepsilon_j}}_{y_2}\right)\left[(\sin y_1, -\cos y_1, 0)\cdot \mathbf{A}^{\varepsilon_j}(T_{\varepsilon_j})\right]\frac{g_{\varepsilon_j}(y_1)g'_{\varepsilon_j}(y_1)y_2}{\varepsilon_j^2}\,dy \\
&\quad + \int_{\mathcal{C}} i\left(\overline{U^{\varepsilon_j}}U^{\varepsilon_j}_{y_3} - U^{\varepsilon_j}\overline{U^{\varepsilon_j}}_{y_3}\right)\left[(0,0,1)\cdot \mathbf{A}^{\varepsilon_j}(T_{\varepsilon_j})\right]\frac{g_{\varepsilon_j}(y_1)(1+g_{\varepsilon_j}(y_1)y_2)}{\varepsilon_j^2}\,dy \\
&\quad + \int_{\mathcal{C}} i\left(\overline{U^{\varepsilon_j}}U^{\varepsilon_j}_{y_3} - U^{\varepsilon_j}\overline{U^{\varepsilon_j}}_{y_3}\right)\left[(\sin y_1, -\cos y_1, 0)\cdot \mathbf{A}^{\varepsilon_j}(T_{\varepsilon_j})\right]\frac{g_{\varepsilon_j}(y_1)g'_{\varepsilon_j}(y_1)y_3}{\varepsilon_j^2}\,dy \\
&= I + II + III + IV + V.
\end{aligned} \tag{2.56}$$

In light of (2.47) and (2.16), we see that

$$\begin{aligned}
\left|(\cos y_1, \sin y_1, 0)\cdot \mathbf{A}^{\varepsilon_j}(T_{\varepsilon_j}(y))\right| &\le |(\cos y_1, \sin y_1, 0)\cdot \mathbf{A}^{e}(\cos y_1, \sin y_1, 0)| \\
&\quad + \left|(\cos y_1, \sin y_1, 0)\cdot (\mathbf{A}^{\varepsilon_j}(T_{\varepsilon_j}(y)) - \mathbf{A}^{e}(\cos y_1, \sin y_1, 0)\right| \\
&\le \left\|\mathbf{A}^{\varepsilon_j}(T_{\varepsilon_j}(y)) - \mathbf{A}^{e}(\cos y_1, \sin y_1, 0)\right\|_{L^\infty(\mathcal{C})} \to 0 \quad \text{as } \varepsilon \to 0
\end{aligned}$$

and that $|(0,0,1)\cdot \mathbf{A}^{\varepsilon_j}| \to 0$ as well, using (2.7). Hence we can apply Hölder's inequality, (2.12), and (2.37) to see that

$$|II| + |IV| \le \mu\left(\frac{\left\|g_{\varepsilon_j}\right\|_{L^\infty(0,2\pi)}}{\varepsilon_j^2}\right)\cdot o(1)\cdot\left(\int_{\mathcal{C}} \left|U^{\varepsilon_j}_{y_2}\right|^2 + \left|U^{\varepsilon_j}_{y_3}\right|^2\,dy\right)^{1/2} \to 0 \tag{2.57}$$

as $\varepsilon \to 0$.

Next, observe that since $|(\sin y_1, -\cos y_1, 0)\cdot \mathbf{A}^{\varepsilon_j}|$ is uniformly bounded in $\mathcal{C}$, and since $|g'_{\varepsilon_j}| = \mathcal{O}(\varepsilon^{1-p})$ (cf. (2.1)), through (2.37) one calculates that

$$|III + V| \le \mu\left(\frac{\left\|g_{\varepsilon_j}g'_{\varepsilon_j}\right\|_{L^\infty((0,2\pi))}}{\varepsilon_j^2}\right)\left(\int_{\mathcal{C}} \left|U^{\varepsilon_j}_{y_2}\right|^2 + \left|U^{\varepsilon_j}_{y_3}\right|^2\,dy\right)^{1/2} \to 0 \tag{2.58}$$

as well.

Consequently, we conclude from (2.56), (2.57), and (2.58) that

$$\lim_{\varepsilon_j\to 0}\int_{\Omega_{\varepsilon_j}} i(\overline{u^{\varepsilon_j}}\nabla u^{\varepsilon_j} - u^{\varepsilon_j}\nabla\overline{u^{\varepsilon_j}})\cdot \mathbf{A}^{\varepsilon_j}\,dx = \lim_{\varepsilon_j\to 0} I. \tag{2.59}$$

We now split the integral I into two integrals over the regions $\mathcal{C}_\delta$ and $\mathcal{C}\setminus\mathcal{C}_\delta$, which we label as I_1 and I_2, respectively.

We first treat the limit of I_2. Through (2.12) and (2.16), we find that

$$|I_2| \le C \int_{-\delta}^{\delta} \int_{\{y_2^2+y_3^2<1\}} \left|U_{y_1}^{\varepsilon_j}\right| \frac{g_{\varepsilon_j}(y_1)^2}{\varepsilon_j^2}\, dy_2\, dy_3\, dy_1.$$

As a consequence of (2.1) and (2.37), we obtain that

$$\begin{aligned}
&\int_{-\delta}^{\delta} \int_{\{y_2^2+y_3^2<1\}} \left|U_{y_1}^{\varepsilon_j}\right| \frac{g_{\varepsilon_j}(y_1)^2}{\varepsilon_j^2}\, dy_2\, dy_3\, dy_1 \\
&\le \left(\int_{-\delta}^{\delta} \int_{\{y_2^2+y_3^2<1\}} \frac{g_{\varepsilon_j}(y_1)^2}{\varepsilon_j^2} \left|U_{y_1}^{\varepsilon_j}\right|^2 dy_2\, dy_3\, dy_1 \right)^{1/2} \left(\int_{-\delta}^{\delta} \int_{\{y_2^2+y_3^2<1\}} \frac{g_{\varepsilon_j}(y_1)^2}{\varepsilon_j^2}\, dy_2\, dy_3\, dy_1 \right)^{1/2} \\
&\le C \left(\int_{-\delta}^{\delta} \int_{\{y_2^2+y_3^2<1\}} 1\, dy_2\, dy_3\, dy_1 \right)^{1/2} \le C\delta^{1/2}.
\end{aligned}$$

Hence, we conclude that

$$\lim_{\varepsilon_j \to 0} |I_2| \le C\delta^{1/2}. \tag{2.60}$$

Turning to I_1, through an appeal to (2.2), (2.16), (2.48), (2.51), and (2.52), we may compute

$$\begin{aligned}
&\lim_{\varepsilon_j \to 0} I_1 \\
&= \lim_{\varepsilon_j \to 0} \int_{\{\delta<|y_1|<\pi\}} \int_{\{y_2^2+y_3^2<1\}} i \left(\overline{U^{\varepsilon_j}} U_{y_1}^{\varepsilon_j} - U^{\varepsilon_j} \overline{U^{\varepsilon_j}}_{y_1} \right) [(-\sin y_1, \cos y_1, 0) \cdot \mathbf{A}^{\varepsilon_j}] \frac{g_{\varepsilon_j}(y_1)^2}{\varepsilon_j^2}\, dy \\
&= \lim_{\varepsilon_j \to 0} \int_{\{\delta<|y_1|<\pi\}} \int_{\{y_2^2+y_3^2<1\}} i \left(\overline{U^{\varepsilon_j}} U_{y_1}^{\varepsilon_j} - U^{\varepsilon_j} \overline{U^{\varepsilon_j}}_{y_1} \right) A_1^e\, dy \\
&= \pi \int_{\{\delta<|y_1|<\pi\}} \int_{\{y_2^2+y_3^2<1\}} i \left(\overline{U^0} U_{y_1}^0 - U^0 \overline{U^0}_{y_1} \right) A_1^e\, dy_1.
\end{aligned}$$

Applying (2.57), (2.58), (2.60), and (2.61) to (2.56), we finally obtain

$$\begin{aligned}
&\lim_{\varepsilon_j \to 0} \int_{\Omega_{\varepsilon_j}} i(\overline{u^{\varepsilon_j}} \nabla u^{\varepsilon_j} - u^{\varepsilon_j} \nabla \overline{u^{\varepsilon_j}}) \cdot \mathbf{A}^{\varepsilon_j}\, dx \\
&\quad = \pi \int_{\{\delta<|y_1|<\pi\}} i \left(\overline{U^0} U_{y_1}^0 - U^0 \overline{U^0}_{y_1} \right) A_1^e\, dy_1 + \mathcal{O}(\delta^{1/2}).
\end{aligned} \tag{2.61}$$

Combining (2.50), (2.53), (2.54), and (2.61) and letting $\delta \to 0$, we establish (2.45).

Proof of claim (2.46). We may assume $V \in W^{1,2}((-\pi,\pi) \setminus \{0\}; \mathbb{C})$ and $V(-\pi) = V(\pi)$, since otherwise the construction of a sequence satisfying (2.46) is trivial. In order to highlight the fact that, in general, V' will be singular at $y_1 = 0$, we denote by $h \in L^2((-\pi,\pi); \mathbb{C})$ the regular part of the derivative of V, that is, the part which is absolutely continuous with respect to Lebesgue measure. Hence,

$$\int_{-\pi}^{0} h(y_1)\, dy_1 = V^- - V(-\pi) \quad \text{and} \quad \int_{0}^{\pi} h(y_1)\, dy_1 = V(\pi) - V^+, \tag{2.62}$$

where V^- and V^+ denote the left- and right-hand limits of V at $y_1 = 0$, respectively.

We proceed to define a sequence $\{V^\varepsilon\} \subset W^{1,2}((-\pi,\pi);\mathbb{C})$ satisfying $V^\varepsilon(\pi) = V^\varepsilon(-\pi)$ and from this we will define the sequence $\{v^\varepsilon\} \subset W^{1,2}(\Omega_\varepsilon;\mathbb{C})$ verifying (2.46) by viewing the argument of V^ε as the polar angle in a cylindrical coordinate system on $\mathbb{R}^3$. Note that the polar angle is precisely the variable y_1 as defined in (2.4).

To this end, recall the definition of the sequence $\{\lambda_\varepsilon\}$ given in (2.42) and denote by β_ε the quantity

$$\beta_\varepsilon := \int_{-\pi}^{\pi} \lambda_\varepsilon(y_1)\, dy_1.$$

A routine calculation shows that

$$\int_{-\varepsilon^p}^{\varepsilon^p} \lambda_\varepsilon(y_1)\, dy_1 = 1 + \mathcal{O}(\varepsilon^p), \quad \text{while} \quad \int_{\{|y_1|>\varepsilon^p\}} \lambda_\varepsilon(y_1)\, dy_1 = \mathcal{O}(\varepsilon^2), \tag{2.63}$$

so, in particular, we have $\beta_\varepsilon = 1 + \mathcal{O}(\varepsilon^p)$.

Now we are prepared to define the sequence $\{V^\varepsilon\} \subset W^{1,2}((-\pi,\pi);\mathbb{C})$ via the formula

$$V^\varepsilon(y_1) = \int_{-\pi}^{y_1} \left\{ h(s) + \frac{1}{\beta_\varepsilon}(V^+ - V^-)\lambda_\varepsilon(s) \right\} ds + V(-\pi). \tag{2.64}$$

This construction follows that found in [4].

With the aid of (2.62) and the periodicity of V, one sees that $V^\varepsilon(-\pi) = V^\varepsilon(\pi)$, and with the aid of (2.63), one readily checks that

$$|V^\varepsilon - V| + |(V^\varepsilon)' - V'| \le C\varepsilon^2 \quad \text{a.e. on} \quad \{|y_1| > \varepsilon^p\}, \tag{2.65}$$

while

$$|V^\varepsilon - V| \le \left|V^+ - V^-\right| \quad \text{a.e. on} \quad \{|y_1| < \varepsilon^p\}. \tag{2.66}$$

We also observe that $V^\varepsilon \to V$ in $L^1((-\pi,\pi);\mathbb{C})$ and that $(V^\varepsilon)' \overset{*}{\rightharpoonup} V'$ as measures on $(-\pi,\pi)$.

In light of the periodicity of V^ε we can now define the sequence $\{v^\varepsilon\} \subset W^{1,2}(\Omega_\varepsilon;\mathbb{C})$ through the relation $v^\varepsilon(x) = V^\varepsilon(\tan^{-1}(x_2/x_1)) = V^\varepsilon(y_1)$. We proceed to verify (2.46) by decomposing the energy $G_\varepsilon(v^\varepsilon, \mathbf{A}^e)$ and studying the limit of each term in the same manner as was done in the proof of claim (2.45).

We write

$$\begin{aligned} G_\varepsilon(v^\varepsilon, \mathbf{A}^e) &= \frac{1}{\varepsilon^2}\int_{\Omega_\varepsilon} |\nabla v^\varepsilon|^2\, dx + \frac{1}{\varepsilon^2}\int_{\Omega_\varepsilon} i(\overline{v^\varepsilon}\nabla v^\varepsilon - v^\varepsilon \nabla \overline{v^\varepsilon}) \cdot \mathbf{A}^e\, dx \\ &+ \frac{1}{\varepsilon^2}\int_{\Omega_\varepsilon} |v^\varepsilon|^2 |\mathbf{A}^e|^2\, dx + \frac{1}{\varepsilon^2}\int_{\Omega_\varepsilon} \frac{\nu^2}{2}(|v^\varepsilon|^2 - \mu^2)^2\, dx. \end{aligned} \tag{2.67}$$

First note that through (2.24) one has

$$\begin{aligned} \frac{1}{\varepsilon^2}\int_{\Omega_\varepsilon} |\nabla v^\varepsilon|^2\, dx &= \frac{1}{2\varepsilon^2}\int_{\mathcal{C}} a_{ik}\big(V^\varepsilon_{y_i}\overline{V^\varepsilon}_{y_k} + \overline{V^\varepsilon}_{y_i} V^\varepsilon_{y_k}\big)\, dy \\ &= \pi \int_{-\pi}^{\pi} \frac{g_\varepsilon(y_1)^2}{\varepsilon^2(1+g_\varepsilon(y_1)y_2)} \left|\frac{d\,V^\varepsilon}{d\,y_1}\right|^2 dy_1 \\ &= \pi \int_{-\pi}^{\pi} a_\varepsilon(y_1) \left|\frac{d\,V^\varepsilon}{d\,y_1}\right|^2 dy_1 + o(1) \quad \text{as} \quad \varepsilon \to 0. \end{aligned}$$

Thus, from (2.42) and (2.64) we see that

$$\begin{aligned}
&\frac{1}{\pi}\lim_{\varepsilon\to 0}\frac{1}{\varepsilon^2}\int_{\Omega_\varepsilon}\left|\nabla v^\varepsilon\right|^2\,dx\\
&=\lim_{\varepsilon\to 0}\int_{-\pi}^{\pi}a_\varepsilon(y_1)\left|h(y_1)+\frac{1}{\beta_\varepsilon}(V^+-V^-)\lambda_\varepsilon(y_1)\right|^2\,dy_1\\
&=\lim_{\varepsilon\to 0}\int_{-\pi}^{\pi}\left|a_\varepsilon(y_1)h(y_1)+\frac{1}{\beta_\varepsilon}(V^+-V^-)a_\varepsilon(y_1)\lambda_\varepsilon(y_1)\right|^2\frac{1}{a_\varepsilon(y_1)}\,dy_1\\
&=\lim_{\varepsilon\to 0}\int_{-\pi}^{\pi}\left|a_\varepsilon(y_1)h(y_1)+\frac{1}{\beta_\varepsilon}(V^+-V^-)(1-a_\varepsilon(y_1))\right|^2(1+\lambda_\varepsilon(y_1))\,dy_1\\
&=\lim_{\varepsilon\to 0}\left\{\int_{-\pi}^{\pi}a_\varepsilon(y_1)^2\left|h(y_1)\right|^2\,dy_1+\frac{1}{\beta_\varepsilon^2}\left|V^+-V^-\right|^2+\int_{-\pi}^{\pi}(1-a_\varepsilon(y_1))f_\varepsilon(y_1)\,dy_1\right\},
\end{aligned}$$

where in the last integral we have introduced the real-valued function f_ε so as to include all the remaining terms coming from expanding the square in the previous line. One readily checks that $\|f_\varepsilon\|_{L^\infty((-\pi,\pi))}\le C$ for some C independent of ε. Then, since $a_\varepsilon\to 1$ in $L^1((-\pi,\pi))$ and $\beta_\varepsilon\to 1$, we conclude that

$$\frac{1}{\pi}\lim_{\varepsilon\to 0}\frac{1}{\varepsilon^2}\int_{\Omega_\varepsilon}\left|\nabla v^\varepsilon\right|^2\,dx=\int_{-\pi}^{\pi}\left|V'\right|^2\,dy_1+\left|V^+-V^-\right|^2. \tag{2.68}$$

Turning to the next integral on the right-hand side of (2.67), we see from (2.56), with u^ε replaced by v^ε and $\mathbf{A}^\varepsilon$ replaced by $\mathbf{A}^e$, that

$$\begin{aligned}
&\int_{\Omega_\varepsilon}i(\overline{v^\varepsilon}\nabla v^\varepsilon-v^\varepsilon\nabla\overline{v^\varepsilon})\cdot\mathbf{A}^e\,dx\\
&=\int_{\mathcal{C}}i\left(\overline{V^\varepsilon}V^\varepsilon_{y_1}-V^\varepsilon\overline{V^\varepsilon}_{y_1}\right)\mathbf{A}^e(T_\varepsilon)\cdot(-\sin y_1,\cos y_1,0)\frac{g_\varepsilon(y_1)^2}{\varepsilon^2}\,dy\\
&=\pi\int_{-\pi}^{\pi}i\left(\overline{V^\varepsilon}V^\varepsilon_{y_1}-V^\varepsilon\overline{V^\varepsilon}_{y_1}\right)A_1^e\frac{g_\varepsilon(y_1)^2}{\varepsilon^2}\,dy_1+\mathcal{O}(\varepsilon).
\end{aligned}$$

Here we have used the fact that $|\mathbf{A}^e(T_\varepsilon)\cdot(-\sin y_1,\cos y_1,0)|=A_1^e+\mathcal{O}(\varepsilon)$.

Hence, from (2.65) we have

$$\begin{aligned}
&\frac{1}{\pi}\int_{\Omega_\varepsilon}i(\overline{v^\varepsilon}\nabla v^\varepsilon-v^\varepsilon\nabla\overline{v^\varepsilon})\cdot\mathbf{A}^e\,dx\\
&=\int_{\{|y_1|>\varepsilon^p\}}i\left(\overline{V^\varepsilon}V^\varepsilon_{y_1}-V^\varepsilon\overline{V^\varepsilon}_{y_1}\right)A_1^e\frac{g_\varepsilon(y_1)^2}{\varepsilon^2}\,dy_1\\
&\quad+\int_{\{|y_1|<\varepsilon^p\}}\bullet\quad dy_1\quad+\quad\mathcal{O}(\varepsilon)\\
&=\int_{\{|y_1|>\varepsilon^p\}}i\left(\overline{V}V_{y_1}-V\overline{V}_{y_1}\right)A_1^e\,dy_1\\
&\quad+\int_{\{|y_1|<\varepsilon^p\}}i\left(\overline{V^\varepsilon}V^\varepsilon_{y_1}-V^\varepsilon\overline{V^\varepsilon}_{y_1}\right)A_1^e\frac{g_\varepsilon(y_1)^2}{\varepsilon^2}\,dy_1+\mathcal{O}(\varepsilon).
\end{aligned}\tag{2.69}$$

Now as a consequence of (2.35), (2.42), and (2.64), we can estimate this last term as follows:

$$\begin{aligned}
&\left| \int_{\{|y_1|<\varepsilon^p\}} i\left(\overline{V^\varepsilon} V^\varepsilon_{y_1} - V^\varepsilon \overline{V^\varepsilon}_{y_1}\right) A_1^e \frac{g_\varepsilon(y_1)^2}{\varepsilon^2}\, dy_1 \right| \\
&\le C \int_{\{|y_1|<\varepsilon^p\}} a_\varepsilon(y_1) \left|V^\varepsilon_{y_1}\right| \, dy_1 \\
&\le C \int_{\{|y_1|<\varepsilon^p\}} a_\varepsilon(y_1) \left|h(y_1)\right| \, dy_1 + C\left|V^+ - V^-\right| \int_{\{|y_1|<\varepsilon^p\}} a_\varepsilon(y_1) \frac{1}{\beta_\varepsilon} \left|\lambda_\varepsilon(y_1)\right| \, dy_1 \\
&= C \int_{\{|y_1|<\varepsilon^p\}} a_\varepsilon(y_1) \left|h(y_1)\right| \, dy_1 + C\left|V^+ - V^-\right| \int_{\{|y_1|<\varepsilon^p\}} \frac{1}{\beta_\varepsilon} \left|1 - a_\varepsilon(y_1)\right| \, dy_1 = \mathcal{O}(\varepsilon^p).
\end{aligned} \tag{2.70}$$

Combining (2.69) and (2.70) and passing to the limit as $\varepsilon \to 0$ we conclude that

$$\lim_{\varepsilon\to 0} \int_{\Omega_\varepsilon} i(\overline{v^\varepsilon}\nabla v^\varepsilon - v^\varepsilon \nabla \overline{v^\varepsilon}) \cdot \mathbf{A}^e \, dx = \pi \int_{-\pi}^{\pi} i\left(\overline{V} V_{y_1} - V \overline{V}_{y_1}\right) A_1^e \, dy_1. \tag{2.71}$$

Finally, one finds that the limits of the last two integrals in (2.67) are given by

$$\begin{aligned}
&\lim_{\varepsilon\to 0} \frac{1}{\varepsilon^2} \int_{\Omega_\varepsilon} \left\{ \left|v^\varepsilon\right|^2 \left|\mathbf{A}^e\right|^2 + \frac{\nu^2}{2}(\left|v^\varepsilon\right|^2 - \mu^2)^2 \right\} dx \\
&= \pi \int_{-\pi}^{\pi} \left\{ \left|V\right|^2 \left|A_1^e\right|^2 + \frac{\nu^2}{2}(\left|V\right|^2 - \mu^2)^2 \right\} dy_1
\end{aligned} \tag{2.72}$$

as an easy consequence of (2.65) and (2.66) using a calculation similar to that carried out in (2.53) and (2.54).

Claim (2.46) follows by combining (2.68), (2.71), and (2.72). ∎

3. Discussion. In this section we discuss some physical implications of the one-dimensional model (2.44). We start by observing that adding a suitable parameter to the geometric characterization of the constriction enables us to control the relative magnitude of the different terms in the limit functional G_0 (2.44). For example, if we replace (2.1)–(2.2) by

$$g_\varepsilon(y_1) = (\varepsilon^{1-p} - 2\sqrt{b}\varepsilon)\left|y_1\right| + 2\sqrt{b}\varepsilon^{1+p} \quad \text{for} \quad 0 \le |y_1| \le \varepsilon^p, \tag{3.1}$$

$$g_\varepsilon(y_1) = \varepsilon \quad \text{for} \quad \varepsilon^p \le |y_1| \le \pi, \tag{3.2}$$

where b is a fixed positive parameter, then we obtain the modified limit functional

$$G_0(U) = \begin{cases} \int_{(-\pi,\pi)\setminus\{0\}} \left(\left| (i\frac{d}{dy_1} + A_1^e)U \right|^2 + \frac{\nu^2}{2}(|U|^2 - \mu^2)^2 \right) dy_1 + b\left|U^+ - U^-\right|^2 \\ \qquad \text{if} \quad U \in W^{1,2}((-\pi,\pi)\setminus\{0\};\mathbb{C}), \quad U(-\pi) = U(\pi), \\ \qquad +\infty \quad \text{otherwise.} \end{cases} \tag{3.3}$$

To simplify the presentation in this section, we replace A_1^e by A and use θ to denote the variable y_1. Equating the first variation of G_0 to zero, we obtain the following jump condition at the weak point ($\theta = 0$):

$$\left(\frac{d}{d\theta} - iA\right) U^+ = \left(\frac{d}{d\theta} - iA\right) U^- = b\left(U^+ - U^-\right). \tag{3.4}$$

Multiplying both sides of (3.4) by $\overline{U^+}$ and taking the imaginary part of the obtained complex-valued expression, we find

$$J^+ := \operatorname{Im}\left(\left(\frac{dU^+}{d\theta} - iAU^+\right)\overline{U^+}\right) = -b\operatorname{Im}\left(U^-\overline{U^+}\right). \tag{3.5}$$

The object J^+ is the supercurrent immediately after ($\theta = 0^+$), the weak link. Similarly we multiply both sides of (3.4) by $\overline{U^-}$ and obtain

$$J^- := \operatorname{Im}\left(\left(\frac{dU^-}{d\theta} - iAU^-\right)\overline{U^-}\right) = b\operatorname{Im}\left(U^+\overline{U^-}\right). \tag{3.6}$$

We therefore deduce that $J^+ = J^-$; i.e., the current is conserved across the junction.

In contrast to the continuity of the supercurrent, the order parameter U and its derivative are *not* continuous across the junction. To write the jump in the amplitude's derivative, we express U in a polar form, $U = \rho e^{i\phi}$. Now, taking the real part of (3.4) multiplied by U^+, we obtain after a quick calculation

$$\frac{d\rho^+}{d\theta} = b\left(\rho^+ - \rho^-\cos(\phi^+ - \phi^-)\right), \tag{3.7}$$

where as before, the superscripts $\cdot^+$ and $\cdot^-$ denote evaluation on either side of $\theta = 0$. Similarly we find

$$\frac{d\rho^-}{d\theta} = b\left(\rho^+\cos(\phi^+ - \phi^-) - \rho^-\right). \tag{3.8}$$

Applying the polar form of U to the current formula (3.5), we get

$$J^+ = J^- = b\rho^+\rho^-\sin(\phi^+ - \phi^-). \tag{3.9}$$

This is the celebrated Josephson formula (cf. (1.1)). Moreover we derived an explicit expression for J_M. It is important to observe that in contrast to most models of Josephson junctions (see, e.g., [2] or [21]), in the model being presented here, the order parameter is *not* continuous at the junction. We point out that our equations form a special case of an ad hoc model due to de Gennes (cf. (7.66) of [7]). In this model, de Gennes postulates that the pair $\left(U^+, (\frac{d}{d\theta} - iA)U^+\right)$ is linearly related to the pair $\left(U^-, (\frac{d}{d\theta} - iA)U^-\right)$ through multiplication by a 2×2 matrix he denotes by M. To make the comparison precise, we recast our connection formula across the junction in the form

$$U^+ = U^- + \frac{1}{b}\left(\frac{d}{d\theta} - iA\right)U^-, \tag{3.10}$$

$$\left(\frac{d}{d\theta} - iA\right)U^+ = \left(\frac{d}{d\theta} - iA\right)U^-. \tag{3.11}$$

Using the notation of [7], we can then identify $M_{11} = M_{22} = 1,\ M_{12} = \frac{1}{b},\ M_{21} = 0$.

The collapse of the three-dimensional domain Ω_ε onto a one-dimensional wire is analogous to similar limits computed in, e.g., [5] and [18]. The new feature here is the strong effect of inhomogeneities in the wire. To demonstrate the effect of the term proportional to b in (3.3), we consider the particular problem of phase transition

in thin wires with constrictions. It is known that the critical temperature in one-dimensional rings depends upon the magnetic flux threading the hole bounded by the ring. This was discovered experimentally by Little and Parks about 40 years ago and has been well established theoretically since then [17]. We proceed to compute this dependency using the model given by G_0. The critical temperature is associated with μ, as we explained below (2.5), while μ, in turn, is determined by the eigenvalue problem that is obtained through linearizing the Euler–Lagrange equation associated with G_0 about the normal state $U \equiv 0$. We refer to [11] for a justification of this statement. The linearized equation is

$$\left(\frac{d}{d\theta} - iA\right)^2 U + (\nu\mu)^2 U = 0, \tag{3.12}$$

together with the jump conditions (3.10)–(3.11) enforced at $\theta = 0$ and periodic boundary conditions at the endpoints $\theta = \pm\pi$.

Proceeding as in [11], we make a gauge transformation $U(\theta) \to U(\theta)e^{i\int_0^\theta A}$. This simplifies the eigenvalue problem (3.12) into

$$\frac{d^2}{d\theta^2}U + (\nu\mu)^2 U = 0 \quad \text{for } -\pi < \theta < 0 \text{ and } 0 < \theta < \pi, \tag{3.13}$$

coupled with the boundary conditions

$$\begin{aligned} &U(-\pi) = U(\pi)e^{i\Phi}, \quad U'(-\pi) = U'(\pi)e^{i\Phi}, \\ &U^+ = U^- + \frac{1}{b}(U')^- \quad \text{at } \theta = 0, \\ &(U')^+ = (U')^- \quad \text{at } \theta = 0, \end{aligned}$$

where $\Phi := \int_{-\pi}^{\pi} A\, d\theta$ is the magnetic flux through the hole bounded by the ring. After writing U in terms of sines and cosines, we obtain through a simple calculation the following transcendental equation for μ:

$$\cos(2\pi\nu\mu) - \frac{\nu\mu}{2b}\sin(2\pi\nu\mu) = \cos\Phi. \tag{3.14}$$

Notice that in the limit $b \to \infty$, the order parameter is continuous, and (3.14) reduces to the clean ring limit [11].

We wish to compare (3.14) to the analogous one obtained in [11]. In [11], a junction is modeled not through a constriction but rather through a modification of the one-dimensional GL energy in a thin region corresponding to angular values $0 \le \theta \le d$, where $d \ll 1$. Specifically, one replaces the potential term $\frac{\nu^2}{2}(|u|^2 - \mu^2)^2$ in (2.5) for these θ-values with the term $\frac{\alpha}{d}|u|^2$, where $\alpha > 0$ is a parameter related to the strength of the junction. Then carrying out an asymptotic analysis of the normal/superconducting phase transition in the small d limit, one finds that the corresponding eigenvalue μ solves the transcendental equation

$$\cos 2\pi\mu + \frac{\alpha}{2\mu}\sin 2\pi\mu = \cos\Phi. \tag{3.15}$$

A comparison of the two phase transition curves corresponding to (3.14) and (3.15) with $\nu = b = \alpha = 1$ can be found in Figure 2. We note that qualitatively the two transition curves coming from the two different models are quite similar, though the

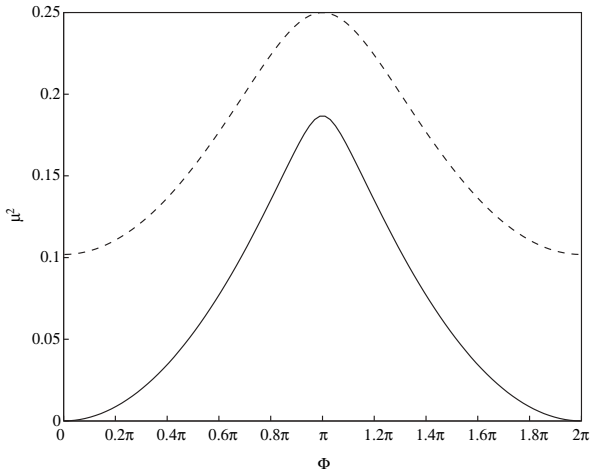


FIG. 2. *Comparison of phase transition curves μ^2 versus Φ for the constricted model (solid) and the modified GL model (dashed).*

transition temperature for the model from [11] is lower (higher μ). Another distinction is that in the model from [11], the curve does not pass through the origin. That is, even for zero magnetic flux through the ring, the transition temperature in that model is lower than the critical temperature associated with the normal/superconducting phase transition in the absence of any applied magnetic field for a ring without normal inclusions. This is not the case in the constricted model.

Returning to the general question of Josephson junctions, we comment that de Gennes did not distinguish between different kinds of junctions. Together with earlier investigations (see [11] and [20]) we are able now to identify two distinguished kinds of junctions. The first kind is a classical SNS junction, where a thin normal layer separates two bulk superconducting samples. Under appropriate scaling it can be shown then that the current is proportional to the sin of the magnetic flux threading through the hole bounded by the wire, but the amplitude of the order parameter is continuous [11]. The second class consists of a geometric weak link (our constriction). Here the current is a periodic function of the phase jump, but the topological constraint implied by the ring is not as strong as in the first class in that the phase is no longer required to jump by a multiple of 2π along the ring. Finally, we point out the heuristics behind our construction. For sufficiently narrow constrictions, it is energetically preferable for the minimizer to have a rapid transition across the link with less drastic variations in the bulk. In this light, the specific geometry of the constriction given by the graph of g_ε is not crucial.

REFERENCES

[1] L. G. ASLAMAZOV AND A. I. LARKIN, *Josephson effect in superconducting point contacts*, JETP Lett., 9 (1969), pp. 87–91.

[2] A. BARONE AND G. PATERNO, *Physics and Applications of the Josephson Effect*, Wiley, New York, 1982.

[3] G. BOUCHITTÉ, *Représentation int'egrale de fonctionnelles convexes sur un espace de mesures. II. Cas de l'épi-convergence*, Ann. Univ. Ferrara. Sez. VII (N.S.), 33 (1987), pp. 113–156.

[4] G. BUTTAZZO AND L. FREDDI, *Functionals defined on measures and applications to non-equi-uniformly elliptic problems*, Ann. Mat. Pura Appl. (4), 159 (1991), pp. 133–149.

[5] S. J. Chapman, Q. Du, and M. D. Gunzburger, *A Ginzburg Landau model of superconducting/normal junctions including Josephson junctions*, European J. Appl. Math., 6 (1996), pp. 97–114.
[6] E. Cabib, L. Freddi, A. Morassi, and D. Percivale, *Thin notched beams*, J. Elasticity, 64 (2002), pp. 157–178.
[7] P. G. de Gennes, *Superconductivity of Metals and Alloys*, Addison-Wesley, Redwood City, CA, 1989.
[8] Q. Du, M. D. Gunzburger, and J. S. Peterson, *Analysis and approximation of the Ginzburg–Landau model of superconductivity,* SIAM Rev., 34 (1992), pp. 54–81.
[9] Q. Du and J. Remski, *Simplified models for superconducting-normal-superconducting junctions and their numerical approximations,* European J. Appl. Math., 10 (1999), pp. 1–25.
[10] Q. Du and J. Remski, *Limiting models for Josephson junctions and superconducting weak links*, J. Math. Anal. Appl., 266 (2002), pp. 357–382.
[11] E. Hill, J. Rubinstein, and P. Sternberg, *A modified Ginzburg Landau model for Josephson junctions in rings*, Quart. Appl. Math., 60 (2002), pp. 485–503.
[12] K. H. Hoffman, L. Jiang, and W. Yu, *Models of superconducting-normal-superconducting junctions*, Math. Methods Appl. Sci., 21 (1998), pp. 59–91.
[13] S. Jimbo and P. Sternberg, *Nonexistence of permanent currents in planar convex samples*, SIAM J. Math. Anal., 33 (2002), pp. 1379–1392.
[14] B. D. Josephson, *Possible new effects in superconducting tunneling*, Phys. Lett., 1 (1962), pp. 251–253.
[15] R. V. Kohn and V. Slastikov, *Geometrically Constrained Walls*, preprint, 2003.
[16] K. Likharev, *Superconducting weak links*, Rev. Modern Phys., 51 (1979), pp. 101–159.
[17] W. A. Little and R. D. Parks, *Observation of quantum periodicity in the transition temperature of a superconducting cylinder,* Phys. Rev. Lett., 9 (1962), pp. 9–12.
[18] J. Rubinstein and M. Schatzman, *Variational problems in multiply connected thin strips* II: *The asymptotic limit of the Ginzburg Landau functional*, Arch. Ration. Mech. Anal., 160 (2001), pp. 309–324.
[19] J. Rubinstein and M. Schatzman, *On mesoscopic superconducting samples*, in Proceedings of the Third International Palestinian Conference on Mathematics and Mathematics Education, S. Elyadi et al., eds., World Scientific, Singapore, 2002, pp. 253–263.
[20] J. Rubinstein and P. Sternberg, *Limits of a Ginzburg-Landau model with codimension-one defects*, J. Math. Phys., 44 (2003), pp. 1240–1251.
[21] M. Tinkham, *Introduction to Superconductivity*, McGraw–Hill, New York, 1996.
[22] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.